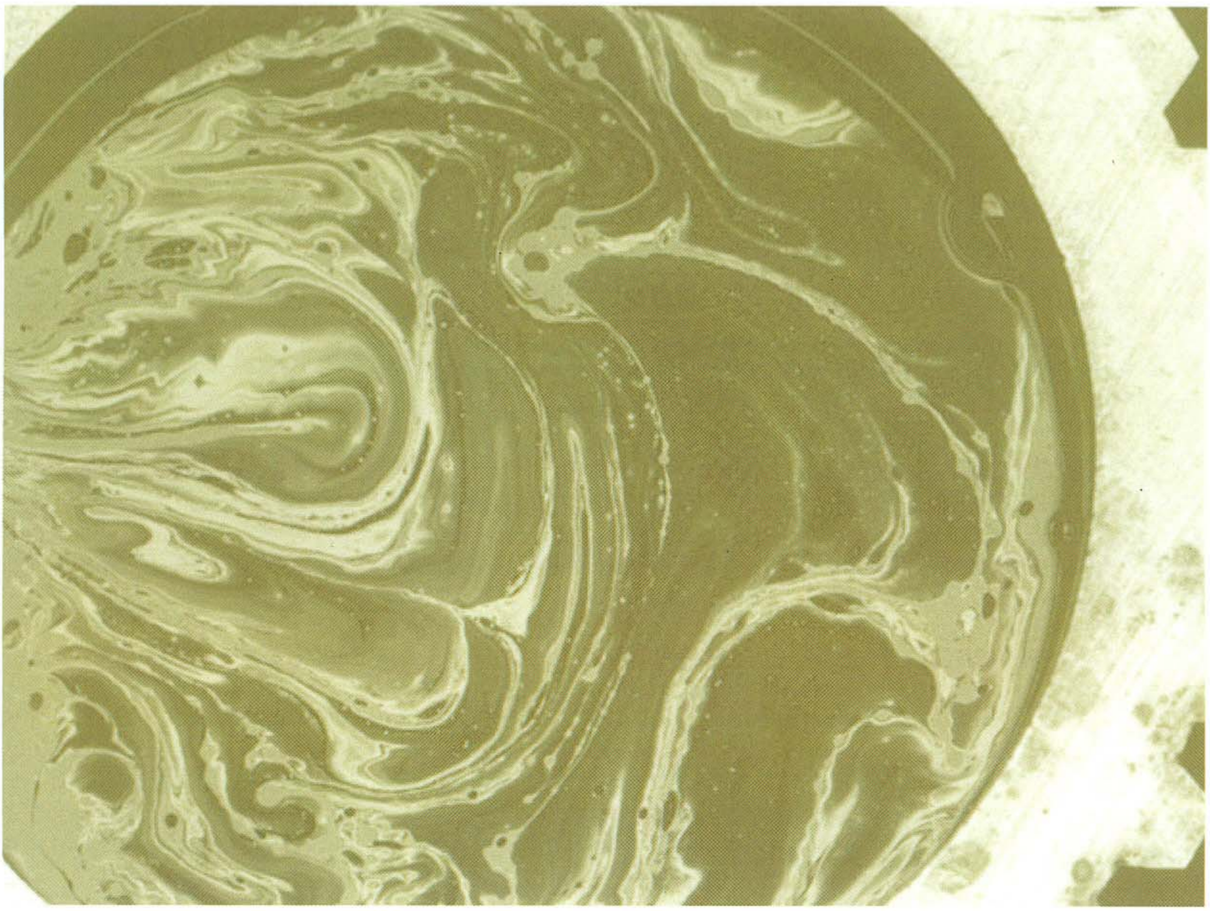
Vol. 1, No. 1
January - March 2005

An official publication of
the Information Resources
Management Association

INTERNATIONAL JOURNAL ON

SEMANTIC WEB AND INFORMATION SYSTEMS



IDEA GROUP PUBLISHING

Publisher of IT books, journals and cases since 1988
www.idea-group.com

IJSWIS is listed in the following indexes: GetCited; MediaFinder; Standard Periodical Directory; Ulrich's International Periodicals Directory

The *International Journal on Semantic Web & Information Systems* (ISSN 1552-6283; eISSN 1552-6291) is published quarterly by Idea Group Publishing, 701 E. Chocolate Avenue, Suite 200, Hershey, PA 17033-1240, USA. Subscription fee to this journal is US$85 individual and US$195 institutional. Sample copies may be requested (based on availability). **Advertising information:** Idea Group Publishing, 701 E. Chocolate Avenue, Suite 200, Hershey, PA 17033-1240, USA, 717/533-8845, <www.idea-group.com>. **Postmaster:** Send all address changes to above address.

Idea Group Publishing

Publisher in information science, technology and management

701 E. Chocolate Avenue, Suite 200, Hershey, PA 17033-1240
Tel: (717) 533-8845 • Fax: (717) 533-8661 • E-mail: cust@idea-group.com
www.idea-group.com

Special Issue Proposal

IJSWIS publishes special issues on the recent trends in Semantic Web and related issues. A short proposal containing the title of proposed special issue, tentative list of invited authors, suggested date of publication and a call-for-papers should be sent to *mdl@aueb.gr* for review.

Idea Group Publishing
http://www.idea-group.com

Hershey • London • Melbourne • Singapore
Tel: 717/533-8845 • Fax: 717/533-8661 • E-mail: cust@idea-group.com

International Journal on Semantic Web & Information Systems

Jan-March 2005, Vol 1, No. 1

Table of Contents

International Journal on Semantic Web & Information Systems

Guidelines for Manuscript Submissions

Mission:
The *International Journal on Semantic Web & Information Systems* is an open forum aiming to cultivate the Semantic Web vision within the Infromation Systems research community. In the common practice of anticipating Semantic Web as a technology driven phenomenon, a scientific insight is provided, which reveals the practical implications and the research challenges of SW in the context of Information Systems. It goes beyond the traditional research agenda of Information Systems and critical themes are analyzed through a Semantic Web perspective in horizontal and vertical pillars. The main idea is to communicate high quality research findings in the leading edge aspects of Semantic Web and Information Systems convergence. This statement distinguishes our journal and differentiates our publishing strategy from other publications: Traditionally Semantic Web is treated as a technological phenomenon with the main emphasis on technologies, languages and tools without similar attention given to theoretical constructions or linkages to multidisciplinary references: Our focus is on the Information Systems Discipline and we are working towards the delivery of the main implications that the Semantic Web brings to Information Systems and the nformation/Knowledge Society.

Coverage:
The main themes covered in the journal include: Semantic Web Issues, challenges and implications in each of the IS research streams; Real world applications toward the development of the knowledge society; New Semantic Web enabled tools for the citizen/ learner/organization/business; New Semantic Web enabled business models; New Semantic Web enabled information systems; Integration with other disciplines; Intelligent systems standards; Semantic enabled business intelligence; Enterprise application integration metadata-driven (bottom-up) vs. ontology-driven (top-down) SW development from e-Government to e-Democracy.

Originality:
Prospective authors should note that only original and previously unpublished manuscripts will be considered. Furthermore, simultaneous submissions are not acceptable. Submission of a manuscript is interpreted as a statement of certification that no part of the manuscript is copyrighted by any other publication nor is under review by any other formal publication. It is the primary responsibility of the author to obtain proper permission for the use of any copyrighted materials in the manuscript, prior to the submission of the manuscript.

Style: Submitted manuscripts must be written in the APA (American Psychological Association) editorial style. References should relate only to material cited within the manuscript and be listed in alphabetical order, including the author's name, complete title of the cited work, title of the source, volume, issue, year of publication, and pages cited. Please do not include any abbreviations. See the following examples:

Example 1: Single author periodical publication.
Smith, A.J. (2002). Information and organizations. *Management Ideology Review*, 16(2), 1-15.

Example 2: Multiple authors periodical publication.
Smith, A. (1988). Organizations and information processing. *Management Source*, 10(4), 77-88.

Example 3: Books.
Smith, A.J. (2002). *Information Booklet*. New York: J.J. Press.

State author's name and year of publication where you use the source in the text. See the following:

Example 1: In most organizations, information resources are considered a major resource (Brown, 1988; Smith, 2002).

Example 2: Brown (2002) states that the value of information is recognized by most organizations.

Direct quotations of another author's work should be followed by the author's name, date of publication, and the page(s) on which the quotation appears in the original text.

Example 1: Brown (1989) states that "the value of information is realized by most organizations" (p. 45).

Example 2: In most organizations, "information resources are considered to be a major organization asset" (Smith, 2002, pp. 35-36) and must be carefully monitored by the senior management.

For more information please consult the APA manual.

Review process:
To ensure the high quality of published material, **IJSWIS** utilizes a group of experts to review submitted manuscripts. Upon receipt of the manuscript, two reviewers are selected from the Editorial Review Board of the journal. The selection is based upon the particular area of expertise of the reviewers, matched to the subject matter of the submission. An additional ad-hoc reviewer is also selected to review the manuscript. Therefore, each submission is accordingly blind reviewed by at least three reviewers. Revised manuscripts will be reviewed again by the original review panel with the addition of one new reviewer. Return of a manuscript to the author(s) for revision does not guarantee acceptance of the manuscript for publication. The final decision will be based upon the comments of the reviewers, upon their second review of the revised manuscript.

Copyright:
Authors are asked to sign a warranty and copyright agreement upon acceptance of their manuscript, before the manuscript can be published. All copyrights, including translation of the published material into other languages are reserved by the publisher, Idea Group Inc. Upon transfer of the copyright to the publisher, no part of the manuscript may be reproduced in any form without written permission of the publisher.

Submission:
Authors are asked to submit their manuscripts for possible publication by e-mail as a file attachment in Microsoft Word or RTF (Rich Text Format) to *mdl@aueb.gr.* The main body of the e-mail message should contain the title of the paper and the names and addresses of all authors. Manuscripts must be in English. The author's name should not be included anywhere in the manuscript, except on the cover page. Manuscripts must also be accompanied by an abstract of 100-150 words, precisely summarizing the mission and object of the manuscript.

Length: The length of the submitted manuscript is not specifically limited, however, the length should be reasonable in light of the chosen topic. Discussion and analysis should be complete but not unnecessarily long or repetitive.

Correspondence:
The acknowledgment letter regarding the receipt of the manuscript will be promptly sent. The review process will take approximately 8-16 weeks, and the author will be notified concerning the possibility of publication of the manuscript as soon as the review process is completed. All correspondence will be directed to the first author of multi-authored manuscripts. It is the responsibility of the first author to communicate with the other author(s). Authors of accepted manuscript will be asked to provide a final copy of their manuscript in either Word or RTF text format stored on a 3 1/2" disk, zip disk or CD-ROM, accompanied by a hardcopy of the manuscript and the original signed copy of the Warranty and Copyright Agreement. The accepted manuscript will be edited by the Journal copyeditor for format and style.

Book Reviews:
IJSWIS invites prospective book reviewers to submit their review of either textbooks or professional books for possible inclusion in the journal. Reviewers should focus on the following guidelines when developing the book review:

- Book reviews must not exceed 1,500 words
- Reviews should summarize the book and indicate the highlights, strengths, and weaknesses of the book.
- Reviews should evaluate the organizational and managerial applications of the material discussed in the book relevant to information resources and technology management.
- Reviews should critique and constructively evaluate the author's work and not merely list the chapters' contents.
- The writing style, accuracy, relevance, and the need for such a work in the discipline should be analyzed.
- The review must include the title of the book, author, publishing company, publication date, number of pages, cost (if listed), and ISBN number.
- Each submission must be accompanied by a short biography of the reviewer.

Case Studies:
The **IJSWIS** encourages submissions of case studies based on actual scenarios related to practice-based issues and practical applications of Semantic Web. Case studies must not exceed 2000 words and must provide adequate information regarding the educational environment upon which the study is based, presentation of the issues involved, coverage of any experiments or techniques involved, and elaborations of the lessons learned or conclusions drawn from the study.

All submissions and questions should be forwarded to:
Dr. Miltiadis D. Lytras
Athens University of Economics and Business, Dept. of Management Science & Technology
ELTRUN - The Research Center
47A Evelpidon Str., 113 62 Athens, Greece
P: 30 210 8203663 * F: 30 210 8203664
mdl@aueb.gr; mdl@eltrun.gr
www.eltrun.gr

Semantic Web and Information Systems:

An Agenda Based on Discourse with Community Leaders

Miltiadis D. Lytras, Athens University of Economics and Business, Greece

ABSTRACT

When last January, AIS approved our proposal for a new SIG on Semantic Web and Information Systems (www.sigsemis.org), I never thought that we could gain the support of so many renowned academics and practitioners. Moreover, I couldn't imagine that all these people would be so excited concerning knowledge sharing and community building around the Semantic Web and its catalytic influence on our traditional perceptions of expressing and exploiting meaning through tools, services, and applications. Having already interviewed four key people for the evolution of the Semantic Web, we decided instead to provide a simple editorial to sketch the open Semantic Web and information systems research agenda.

INTRODUCTION

Semantic Web = Semantics + Web — a simple equation, yet so many ways to interpret it. I feel blessed that I had the opportunity and the honor to find guidance, vision, and different views from all the leaders of the Semantic Web — namely (and in order of time) Amit Sheth, James Hendler, Chris Bussler, and Eric Miller. Every time I finished an interview, I felt that these people contributed to all of us by stipulating our energy to be part of a great and silent revolution, the one of the Semantic Web (SW). I will try in the next few pages to put together their thoughts and ideas with my questioning. My ultimate objective in the end is to interpret the initial equation. Semantics is the first, the most important part of the Semantic Web.

According to Sheth:

"Semantics has long been recognized to be very important in IS, Databases, AI, Linguistics and many other fields. From the IS/DB perspective, I remember talking about 'So Far (Schematically) yet So Near (Semantically)' in 1992, but lots of smarter people have talked about semantics for some time. More recently however, two things have happened—one positive, one potentially not so positive. The positive thing is that we have now been able to engineer semantic technology that supports large-scale semantic applications, and use large populated ontologies to provide semantic underpinning. At the same time a questionable development is a rather overwhelming importance attached to 'formal semantics.'" (Note: Sheth does not argue against the importance on formal semantics, he merely questions sole or overwhelming reliance on it since, for semantics, one needs to bridge a gap between humans, real world domain knowledge and the machines, and the formal representation works adequately well only for the machines.)

In this inaugural issue, Amit Sheth, Cartic Ramakrinshan, and Christopher Thomas, all from the University of Georgia, provide an excellent discussion on "Semantics for the Semantic Web: The Implicit, the Formal, and the Powerful." Considering the role of semantics in a number of research areas in computer science, they organize semantics in three forms — implicit, formal, and powerful — and explore their roles in enabling some of the key capabilities related to the Semantic Web. The central message of this article is that building the Semantic Web purely on description logics will artificially limit its potential, and that we will need to both exploit well-known techniques that support implicit semantics, and develop more powerful semantic techniques. This article is surely an excellent starting point for everyone interested in SW. Please note that since this first issue consists of invited papers rather than refereed, I have taken the liberty of asking the EIC to contribute.

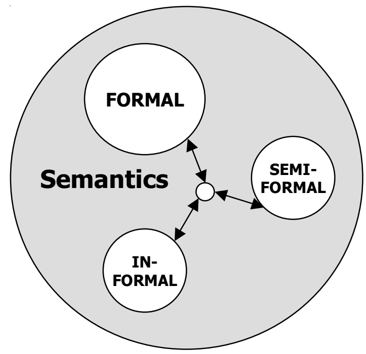
If you ask a newcomer in SW the first thing he would like to know, likely his response will address the impact of a Semantic Web on services and applications. It is the key issue for promoting the visibility and the worthiness of SW. The starting point of such questioning is always the same: "What is the difference in comparison to the WWW that we all know?"

Eric Miller set an interesting "framework" for understanding the evolution of the Semantic Web in relation to the WWW evolution:

"If we think back to the phases associated with Web deployment: 1. The Web was born at CERN, 2. Was first picked up by high-energy physicists, 3. Then by academia at large, 4. Then by small businesses and start-ups, 5. Big business came only later! I'd suggest the Semantic Web is now at #4, and very quickly moving to #5."

I believe this is only the half story. And of course Miller just pointed out the critical aspect: the Semantic Web is not an initiative hermetically tight to close academic clubs.

Figure 1. Semantics types



"Semantic Web is here to stay," Sheth says with emphasis.

So from this perspective I totally agree with Sheth:

"More exiting and important goals of the next generation Web research are improving the human experience and enriching the living, and I can now see a possibility of a major shift from focus on computing to improving human experience — not only with better ability to use heterogeneous content and apply knowledge, but also to incorporate perception and pervasive computing."

This ultimate objective reveals the key argumentation for technology adoption. We need technologies as means for improving our lives and expanding our frontiers towards the common wealth. But his statement hides a lot of engineering.

Chris Bussler provided a very interesting insight to this issue:

"When you look behind the scenes, and study what information systems infrastructure has to be put in place and maintained in order to provide that level of services, be it for customers or businesses, the state of affairs can be improved quite a bit from a technological side, especially the semantics side of it. Too many glitches

happen due to missing semantic underpinning. Once semantics-based technologies are available, the situation for customers and businesses can be advanced a lot beyond the current state, too, in addition to overcoming today's problems."

The emphasis on today's problems and on the vision for tomorrow incorporates the two major pillars of the so-called knowledge society. Sheth provided one more excellent insight with a philosophic flavor:

"…It is fairly certain that nothing we are seeing is a utopia. We all have [the] tendency to get unduly excited with every new trend and fad, and after a field matures, we find out that instead of them being [a] major life-changing technology or science, they are a step towards a continuing evolution."

This is the Semantic Web for me also. We do not have to underestimate the social context, not even the societal inquiries for new services.

Danny Ayer contributed in *AIS SIGSEMIS Bulletin* a very interesting article. I found his positions quite informing. His main point is that the current Web has many inefficiencies and characteristics that limit its value. Ayers' excellent description in Figure 2 sets an interesting context for revealing the required elements of the Semantic Web. In the next section we try to outline the key research issues of the Semantic Web research agenda.

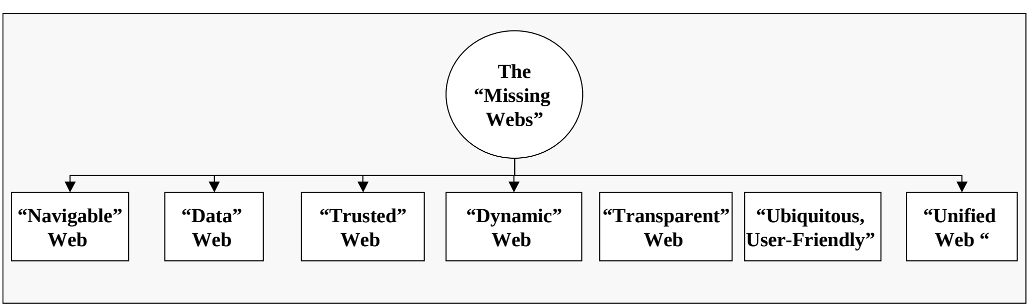
Navigable Web

There is a huge amount of information on the Web, but it is of limited use without it being possible to access that information with ease. Compared to traditional systems the current Web is closer to a file system than a relational database. We have a means of storing and labeling the documents, what we don't have is any built-in technique for indexing and searching them. Catalogue-styled portals do help, and search engines like Google are extremely good at finding a needle in a haystack. The hierarchies of catalogue portals and Google point to ways in which information can be more efficiently retrieved. Many portals are built from taxonomic hierarchies, in effect metadata-based navigation.

Data Web

The current Web is primarily a very large number of hyperlinked documents. Whether they're written in loose HTML or more controlled XHTML format, these documents are designed for human reading. The intended path of use goes directly from the organized bits of data through a renderer to the end user. The Web is currently closer to a microfiche repository with an optical viewer than

Figure 2. The missing Webs (adopted by Ayer, 2004)



a knowledge representation system. But much of the information on the world's computers isn't in this form; it exists as chunks of information relating to real-world or abstract notions and the relationships between these chunks. As a generalization it could be called relational data, and in fact much of it is stored in quasi-relational SQL databases. But on the current Web, data like this is only usually available through very narrow, human-oriented interfaces. A lot of data held by companies and other organizations will be commercially or politically sensitive, and would need to be kept private. But a considerable proportion of it could be made available more widely to general benefit. Given a framework that supports differing levels of access control, data could be published anywhere in between the private/public extremes.

Trusted Web

If information relating to the source of information can be reliably managed, then this opens up potential in several directions. Being sure of aspects like 'who asserted' and 'when' related to facts enables any conclusions inferred from statements based on those facts to carry some of that assurance.

Dynamic Web

A visitor from another planet might be forgiven for thinking that computers are solely communication devices. Apart from infrastructure wiring, the Web barely acknowledges that computers are good for computing. *To take the computing model beyond the isolated mainframe or desktop PC requires integration of software across organization and even application boundaries.*

Transparent Web

The Web Service approach of passing messages between systems offers a partial solution to making the Web more dynamic. For example, material contained in relational databases can be exposed, so their information becomes as available as that of published documents. But as already noted, the interface tends to be narrow. A database of a hundred tables, a thousand columns, and a million rows may appear on the Web as a single node through which queries have to be tunneled. For efficient interaction between end users and services and between services, a level of transparency is needed in which parcels don't have to be opened to discover their contents.

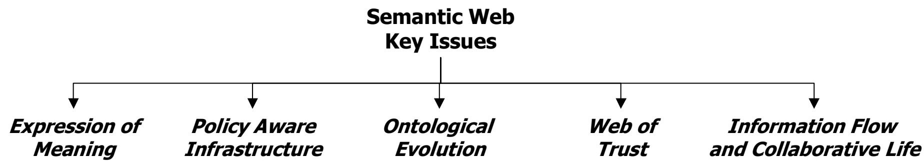
Ubiquitous, User-Friendly Web

Currently most access to the Web takes place through PCs or laptops. There has been some extension into smaller mobile devices, as well as TV-based systems. Wireless has also helped to break some physical restrictions. But still, these are still relatively specialized interfaces; access is far from being on hand everywhere to everyone. In terms of user-friendliness, the Web is generally accessed through a[n] HTML-oriented browser. This usually means read-only access, in very limited single mode of interaction. It lags far behind what is expected of "fat" desktop PC applications. Ubiquity and user-friendliness are key to humanity getting the maximum benefit of the Web, for people to have their abilities augmented at individual and societal levels.

Unified Web

The connectivity of the Web occurs at the level of hyperlinks, in effect the only

Figure 3. Semantic Web key issues



shared languages are fairly low-level protocols. For the Web to be really useful, more sophisticated connectivity is needed. This requires language to describe the entities involved and the relationships between them. Given the scale and diversity of information sources, whatever language is used must be applicable in a very generic way. The only languages that are likely to fit the bill are mathematical, and the prime contenders are understandable in terms of first-order logic.

SEMANTIC WEB KEY ISSUES

The Semantic Web Activity of the W3C is the key driver for promoting the Semantic Web vision. In our interview with Eric Miller, he outlined four interesting areas:

- **Creating a Policy Aware Infrastructure:** The development of a Policy Aware Infrastructure for the Web is required. The Semantic Web will only achieve its potential as an information space for the free flow of scientific and cultural information if its infrastructure supports a full range of fine-grained policy controls over the information contained in the Semantic Web. If we are going to entrust more of our knowledge to the Semantic Web, we must be assured that the Web will respect many more of the social agreements that we enforce in the physical world. For the Semantic Web includes not only freely available information, but also personal information and information available to a person or agent only as a result of its membership in groups. A policy-aware infrastructure — one that gives information creators and users the types of control over information we have all become accustomed to in the physical world such as the ability to assert and exercise privacy and intellectual property rights — will make the Semantic Web into a vibrant and humane environment for sharing knowledge and collaborating on [a] wide range of intellectual enterprises.
- **Ontological Evolution:** An important goal of the Semantic Web is to address the problem that, in the course of scientific (or any) endeavor, one changes the vocabularies one uses to organize, discover, and communicate. A given vocabulary may be refined, resulting in a need for migration from old to new. Communication between distinct groups using different vocabularies creates the need to create common vocabularies, which optimally suit all involved. Semantic Web techniques should make this difficult process of creating new common vocabularies as easy as possible. The Semantic Web already removes confusion by giv-

ing each term a globally unique URI. OWL ontologies and rules languages allow relationships between old and new terms to be expressed. There is, however, little experience with the serious management of such evolution. The Semantic Web needs to incorporate versioning and provenance within its foundation. Human understanding changes and statements that we once thought were accurate are later described to be inaccurate. However, the original statement should not be deleted from our corpus of human knowledge. The Semantic Web should not be required to forget that a statement was once believed to be a true statement. Versioning is such a common approach to representing discrete states of understanding that it warrants explicit treatment in the Semantic Web.

- **Web of Trust:** Trust in the human social context is based on constantly evolving and adapting information. Two parties may trust each other based on a history of mutual interaction, based on formal contracts that in turn rely on other established systems (e.g., legal and legislative), and based on risk analysis of a failure of any party to perform as agreed. A trust language for the Semantic Web that is capable of representing these complex and evolving relationships will be crucial to our future ability to build software that behaves more in the manner of an intelligent assistant than a rote rules processor.
- **Information Flow and Collaborative Life**: Many tools used with collaborating groups today instrument the flow of data, information, and knowledge. One of the challenges we will meet is to strike a balance between requiring authors to do more at the outset to make information machine processable, insisting that everything the machine could use to answer a question be recognized and identified by the (human) questioner, and leaving large quantities of information inaccessible to the machine.

In this list of semantic Web key issues, information systems researchers can find many interesting research topics to contribute. In the interviews, we tried to get feedback on the open research agenda and the "hot topics" that require a multidisciplinary approach. Hendler and Sheth shared with us their thoughts on this issue. More specifically, Hendler's short list includes five significant areas of research:

1. On the Semantic Web the **ontologies are linked together** and can use terms from other ontologies and change them. The system is open and distributed, and there is no way to guarantee consistency. How do we do reasoning in this kind of distributed and inconsistent system?
2. **Social networks** are becoming very popular on the Web, and it is clear that Semantic Web technologies help support large, distributed networks of people who know other people (like the Friend of a Friend, FOAF, work). What new and exciting things can explicit semantics add to these?
3. How are **traditional** technologies (information retrieval, artificial intelligence, etc.) changed by Web semantics and Semantic Web languages?
4. One of the promises of the Semantic Web is that it will let us **bring databases and structured information sources (like spreadsheets) to the Web**. How will query and search engines for this kind of information work?

5. How will semantics function in the emerging world of mobile and ubiquitous computing and other emerging IT trends?

I find Sheth's perspective refreshing because he is among the very few lucky guys who had an opportunity to simultaneously work with the entire span of research, prototyping, technology transfer, commercialization, and real-world application deployment:

"At LSDIS, I can work with colleagues and a large group of PhD students on long-term and conceptual research which allows me to collaborate with industry and provide inputs to standards activities. We have twice licensed technology resulting from our research, leading to start ups, including Semagix (earlier Taalee). At Semagix, I get to work with smart engineers — some of whom are LSDIS alumni — to develop a leading product in SW and architect customer specific solutions. On the same day I can work on research papers and prototypes, as well as deal with challenges of a deployment at a Fortune 500 customer. It has been incredibly exciting."

This is not the only reason why I respect his opinion, but this mix of activities is enough to trust Sheth as a leader. So I asked him to sketch for me his hot topics in Next-Generation Web Research:

"It's hard to pick a few, but here I have a few favorite ones.

In [the] research arena, these include:

1. increasing automatic extraction/annotation of newer forms of digital media, including streaming media, broadcast TV, and sensor-generated data streams;
2. complementing semantic or thematic metadata (and corresponding domain ontologies) with spatial and temporal metadata and ontologies, and providing comprehensive spatio-temporal thematic reasoning; and
3. extending semantics description of static aspects (such as data input and output) of resources or Web services to descriptions related to functional and execution behavior and quality of service, along with increasing semantic support for dynamic nature of Web processes.

In [the] commercial and application arena, some of the favorites include:

4. automated literature search and mining for pharmaceutical R&D;
5. business intelligence applications of opinion and brand management for marketing; and
6. increasing use of semantics in Web search especially as more major players compete with Google.

THE WAY AHEAD

A lot of questioning is related to the time required for the realization of the Semantic Web. Sheth put his vision for the Semantic Web in an interesting triangle:

"My view is an amalgamation of what I have seen on 'experiential computing' by Ramesh Jain, 'computing with words' by Lotfi Zadeh, and 'humanist computing' by Jonathan Rossiter. For those focused on semantics and IS, we still need to address the difficult and fundamental problem of identifying entities (from unstructured text), semantic disambiguation, and discovering (potentially fuzzy, inexact, or probabilistic) relationships. And while formal representation and techniques certainly have a role, we need to find [a] much better way for involving humans — much more than in human interfaces and visualization issues — in any approach supporting semantics and knowledge management."

Hendler contributed to this debate with an to-the-point comment:

"I think we are going to see continually evolving capabilities based on the Semantic Web infrastructure. I think the first application area where we see it deployed is in enterprise application. That will let us see the creation of 'islands' of Semantic Web functionality. We will also see the Semantic Web allowing the creation of easier-to-build and -run Web portals. These will also give us areas of content to link together, and as all these things do link together, the Web of metadata will grow, and we will see the Semantic Web really emerge."

Certainly, the Semantic Web cannot be considered as a general milestone or an illusion. A lot of things have to be done, and this road ahead seems to require a well-defined step-by-step approach. Bussler put things in perspective:

"The Semantic Web is a long-term effort working towards a clear goal, not at all changing every year. Solid progress based on real impact creates a successful area, solid as well as healthy growth, and never a bubble. All involved parties — DERI, research groups, industry, standards organizations, and customers — are interested in making the Semantic Web a reality, not a fashion or a bubble at all. Milestones going forward will be industry-wide pick-up of Semantic Web and Semantic Web Services technology, broad application in all industrial domains, and an ongoing establishment of Semantic Web and Semantic Web Services research groups in universities and research institutes worldwide."

So the obvious question is, where are we now? Miller commented on the success of the WWW2004 Web Conference:

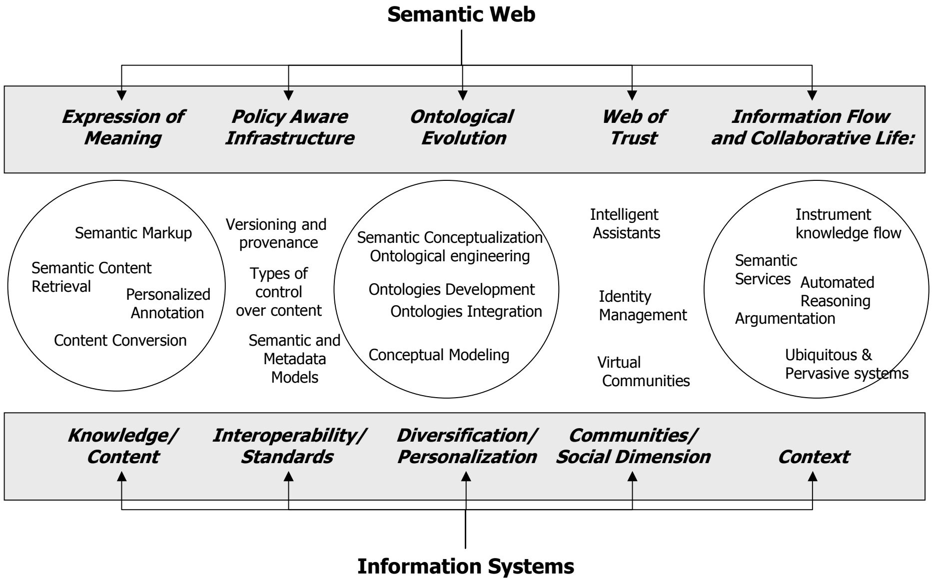
"The WWW2004 Web Conference had a huge Semantic Web focus that permeated almost all aspects of the conference. The energy at the meeting, the collaboration occurring in the corners and throughout the night, reminded me of the second Web conference in Chicago. In Chicago, it seems to me this was a turning point, as everyone who attended realized the Web was not a fad, but rather something that was going to revolutionize how we communicate. The WWW2004 Conference had a similar impact on me with regards to the Semantic Web. The technologies and toolkits are maturing. Semantic Web applications are becoming far more prevalent. Novel ideas for how these technologies may be used are happening on a daily basis. It was quite a week!"

THE SEMANTIC WEB EQUATION IN IS

This article is not a research paper; rather, it is a synthesis of opinions, ideas, and thoughts. I decided to draw a line and to summarize my understanding of the Semantic Web and its role in information systems from a naïve's perspective. In Figure 4, key Semantic Web issues are combined with some ontological perceptions for information systems. This rich picture of Semantic Web and information systems research agendas can be a useful guide for putting ourselves within the new context of Semantic Web-enabled information systems. In the three cyclical areas, we can see the basic research streams and topics that currently gain the main interest of researchers. In fact five pairs describe converging actions of semantic Web and information systems research:

- Expression of meaning/managing of knowledge content
- Ontological evolution/diversification-personalization
- Information flow and collaborative life/context
- Policy aware infrastructure/interoperability-standards
- Web of trust/communities—social dimension

Figure 4. Semantic Web/information systems landscape



ALSO IN THIS ISSUE

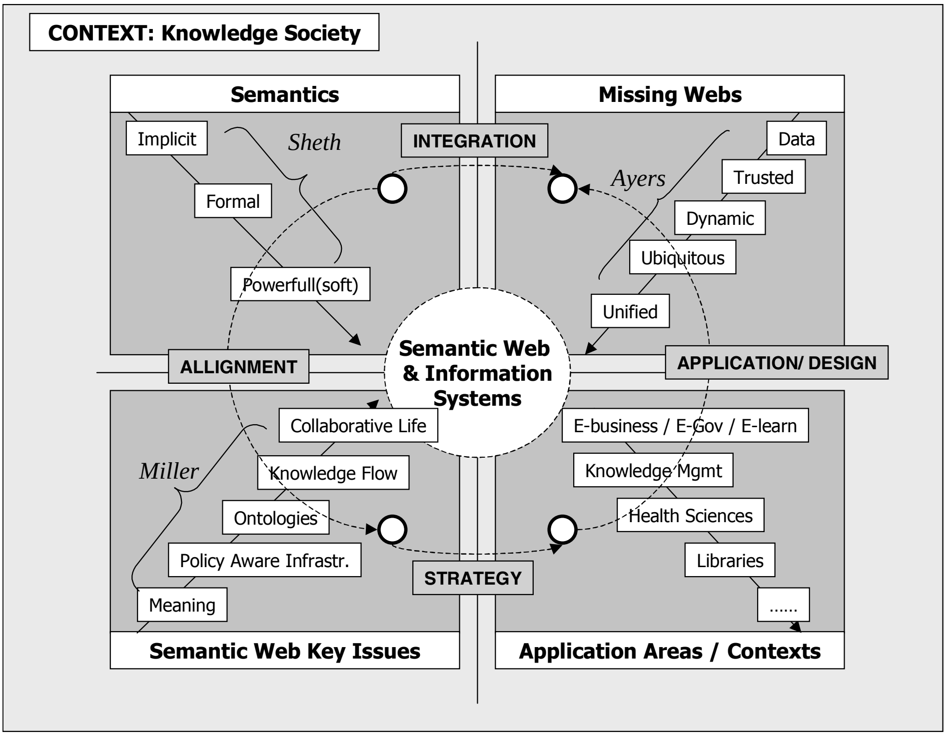
In this first issue we tried to have a balanced mix of articles. The second article of the issue, by Rahul Singh, Lakshmi Iyer, and A.F. Salam, co-organizers of the Semantic E-Business Track in AMCIS and guest editors of a special issue on the same topic in *Communications of the ACM*, unfold the research agenda of "Semantic E-Business." They present a holistic view of semantic e-business that integrates emergent and well-grounded Semantic Web technologies to improve the current state of the art in the transparency of e-business processes.

Given the great interest on ontologies development and tools that facilitate the relevant process, we decided to include in the inaugural issue two papers that explain some ontological engineering considerations. Of course we must from the beginning distinguish the nature of our journal. We are not seeking to provide to the research community a solid technical journal. Instead, we want to pay more attention to the business issues and the drivers of applied technologies, and from this perspective we emphasize the discussion of research problems, the business justification, and the new facts of the Semantic Web towards real-world problems.

In the third article of this inaugural issue, Aditya Kalyanpur, Bijan Parsia, and James Hendler, from the University of Maryland, describe "A Tool for Working with Web Ontologies." Beginners in the SW will find this article very interesting since they will see how ontologies and the Semantic Web affect the way we structure and exploit knowledge. Without emphasis on research findings, this article is a good starting point for people interested in working with Web ontologies.

The last article of this issue has a similar orientation. Artem Chebotko, Yu Deng, Shiyong Lu, Farshad Fotouhi, and Anthony Aristar present "An Ontology-Based Multimedia Annotator for the Semantic Web of

Figure 5. Semantic Web definition for information systems



Language Engineering" for multimedia linguistic data. This type of research, I do believe, in the next year will be a must for SW evolution. The management of multimedia content lacks from several perspectives and ontological insights, and SW technologies provide new insights to the problem.

I conclude this editorial with a figure of the knowledge society I would develop in order to communicate my understanding of the Semantic Web. In a way this is for me the translation of the initial equation: "What is the difference in comparison to the WWW that we all know?"

The Semantic Web in the context of information systems research is "*the integration of semantics in the context of 'missing Webs' towards the alignment and proposition of new strategies that capitalize on semantic Web key issues and provide value in specific information systems contexts.*"

This general definition is the mission statement for our journal. In the forthcoming issues we plan very interesting things. We decided to follow the difficult journey of planning and organizing a new journal from scratch in a very short time. In this journey we want your help; we are looking forward for your comments, we need your participation in AIS SIGSEMIS and SIGODIS activities, and moreover we are open for ideas on collaboration. See you in the next issue.

ACKNOWLEDGMENTS

I would like personally to thank the Council of AIS for their great support and especially Ritu Agarwal. Without their approval our efforts could not find such a creative place to foster growth.

Moreover I would like to thank Gottfried Vossen and Gerd Wagner for their warm participation in the initial proposal of SIGSEMIS.

My deepest appreciation to all the renewed academics and practitioners who joined our SIGSEMIS board and also provide the Editorial Board of *IJWSIS*.

Special thanks and great respect to Professor Amit Sheth, who is kindly providing inspiration for our current strategy and the future of our activities. Without his help, this journal could not be a reality.

Many thanks also to AIS SIGSEMIS members for their collaboration towards our objective for cultivating the Semantic Web vision in the information systems research community. I invite you to join our SIG and be part of an exciting community. (www.sigsemis.org/sig/membership/document_view).

I am grateful to Mehdi Khosrow-Pour, Jan Travers, and all the personnel at Idea Group Inc. who do their best for serving our journal.

My last GREAT THANKS to my wife Theodora, and my little children Dimitris and Hara, who see me every night after midnight authoring and editing important papers.

REFERENCES

Ayers, D. (2004). The missing Webs. *AIS SIGSEMIS Bulletin*, *1*(2), 18-23. Available at *www.sigsemis.org*

Bussler, C. (2004). An interview with Christoph Bussler. *AIS SIGSEMIS Bulletin*, *1*(2), 74-80. Available at *www.sigsemis.org*

Miller, E. (2004). An interview with Eric Miller. *AIS SIGSEMIS Bulletin*, *1*(2), 3-10. Available at *www.sigsemis.org*

Hendler, J. (2004). An interview with James Hendler. *AIS SIGSEMIS Bulletin*, *1*(1), 2-4. Available at *www.sigsemis.org*

Sheth, A. (2004). An interview with Amit Sheth. *AIS SIGSEMIS Bulletin*, *1*(1), 14-19. Available at *www.sigsemis.org*

Semantics for the Semantic Web:
The Implicit, the Formal and the Powerful

Amit Sheth, University of Georgia, USA
Cartic Ramakrishnan, University of Georgia, USA
Christopher Thomas, University of Georgia, USA

ABSTRACT

Enabling applications that exploit heterogeneous data in the Semantic Web will require us to harness a broad variety of semantics. Considering the role of semantics in a number of research areas in computer science, we organize semantics in three forms — implicit, formal, and powerful — and explore their roles in enabling some of the key capabilities related to the Semantic Web. The central message of this article is that building the Semantic Web purely on description logics will artificially limit its potential, and that we will need to both exploit well-known techniques that support implicit semantics, and develop more powerful semantic techniques.

Keywords: analytical processing; data exploration; data extraction; document management and retrieval; fuzzy logic; informal semantics; knowledge discovery; metadata; relationship discovery; semantic analytics; semantic integration; semantic matching; semantic search; soft computing; text management

INTRODUCTION

Semantics has been a part of several scientific disciplines, both in the realm of Computer Science and outside of it. Research areas such as Information Retrieval (IR), Information Extraction (IE), Computational Linguistics (CL), Knowledge Representation (KR) Artificial Intelligence (AI), and Data(base) Management (DB) have all addressed issues pertaining to semantics in their own ways. Most of these areas have very different views of what "meaning" is, and these views are all built on some meta-theoretical and epistemological assumptions. These different views imply very different views of cognition, of concepts, and of meaning (Hjorland, 1998). In this article, we organize these different views to three forms of semantics: implicit, formal, and powerful (a.k.a. soft). We use these forms to explore the role of semantics that go beyond the narrower interpretation of the Semantic Web (that involve adherence to contemporary Semantic Web standards) and encompass those required for a broad variety of semantic applications. We advocate that for the Semantic Web

(SW) to be realized, we must harness the power of a broad variety of semantics encompassing all three forms.

IR, IE, and CL techniques primarily draw upon analysis of unstructured texts in addition to document repositories that have a loosely defined and less formal structure. In these sorts of data sources, the semantics are *implicit*.

In the fields of KR, AI, and DB, however, the data representation takes a more formal and/or rigid form. Well-defined syntactic structures are used to represent information or knowledge where these structures have definite semantic interpretations associated with them. There are also definite rules of syntax that govern the ways in which syntactic structures can be combined to represent the meaning of complex syntactic structures. In other words, techniques used in these fields rely on *formal semantics*.

Usually, efforts related to formal semantics have involved limiting expressiveness to allow for acceptable computational characteristics. Since most KR mechanisms and the Relational Data Model are based on set theory, the ability to represent and utilize knowledge that is imprecise, uncertain, partially true, and approximate is lacking, at least in the base/standard models. However, there have been several efforts to extend the base models (e.g., Barbara, Garcia-Molina, & Porter, 1992). Representing and utilizing these types of more powerful knowledge is, in our opinion, critical to the success of the Semantic Web. Soft computing has explored these types of powerful semantics. We deem these *powerful (soft)* semantics as distinguished, albeit not distinct from or orthogonal to *formal* and *implicit semantics*.

More recently, semantics has been driving the next generation of the Web as the Semantic Web, where the focus is on the role of semantics for automated approaches to exploiting Web resources. This involves two well–recognized, critical enabling capabilities — ontology generation (Maedche & Staab, 2001; Omelayenko, 2001) and automated resource annotation (Hammond, Sheth, & Kochut, 2002; Dill et al., 2003; Handschuh, Staab, & Ciravegna, 2002; Patil, Oundhakar, Sheth, & Verma, 2004), which should be complemented by an appropriate computational approach such as reasoning or query processing. We use a couple of such enabling capabilities to explore the role and importance of all three forms of semantics.

A majority of the attention in the Semantic Web has been centered on a logic-based approach, more specifically that of description logic. However, looking at past applications of semantics, it is very likely that more will be expected from the Semantic Web than what the careful compromise of expressiveness and computability represented by description logic and the W3C adopted ontology representation language OWL (even its three flavors) can support. Supporting expressiveness that meet requirements of practical applications and the techniques that support their development is crucial. It is not desirable to limit the Semantic Web to one type of representation where expressiveness has been compromised at the expense of computational property such as decidability.

This article is not the first to make this above observation. We specifically identify a few. Uschold (2003) has discussed a semantic continuum involving informal to formal and implicit to explicit, and Gruber (2003) has talked about informal, semi-formal, and formal ontologies. The way we use the term *implicit semantics*, however, is different compared to Uschold (2003) insofar as we see implicit semantics in all kinds of data sets, not only in language. We assume that machines can analyze implicit semantics with several, mostly statistical, techniques. Woods has written extensively regarding the limitations of first-order logics (FOLs) — and hence description logics, or DLs — in the context of natural language understanding, although limitations emanating from rigidness and limitation of expressive power, as well as limited value reasoning supported in DLs, can also be identified:

"Over time, many people have responded to the need for increased rigor in knowledge representation by turning to first-order logic as a semantic criterion. This is distressing, since it is already clear that first-order logic is insufficient to deal with many semantic problems inherent in understanding natural language as well as the semantic requirements of a reasoning system for an intelligent agent using knowledge to interact with the world." (Woods, 2004)

We also recall Zadeh's long-standing work (such as Zadeh, 2002), in which he extensively discussed the need for what constitutes a key part of the "powerful semantics" here. In essence, we hope to provide an integrated and complementary view on the range of options. One may ask what the uses of each of these types of semantics are in the context of the Semantic Web. Here is a quick take.

- *Implicit semantics* is either largely present in most resources on the Web or can easily (quickly) be extracted. Hence mining and learning algorithms applied to these resources can be utilized to extract structured knowledge or enrich existing structured formal representations. Since formal semantics intrinsically does not exist, implicit semantics is useful in processing data sets or corpus to obtain or bootstrap semantics that can be then represented in formal languages, potentially with human involvement.
- *Formal semantics* in the form of ontologies is relatively scarce, but representation mechanisms with such semantics have definite semantic interpretations that make them more machine-processable. Representation mechanisms with formal semantics therefore afford applications the luxury of automated reasoning, making the applications more intelligent.
- *Powerful (soft) semantics* in the form of fuzzy or probabilistic KR mechanisms attempt to overcome the shortcomings of the rigid set-based interpretations associated with currently prevalent representation mechanisms by allowing for representation of degree of membership and degree of certainty. Some of the domain knowledge human experts possess is intrinsically complex, and may require these more expressive representations and associated computational techniques.

These uses are further exemplified later on using Semantic Web applications as driving examples. In the next section we define and describe *implicit, formal* and *powerful (soft) semantics*.

TYPES OF SEMANTICS

In this section we give an overview of the three types of semantics mentioned. It is rather informal in nature, as we only give a broad overview without getting in depth about the various formalisms or methods used. We assume that the reader is somewhat familiar with statistical methods on the one hand and Description Logics/ OWL on the other. We present a view of these methods in order to lead towards the necessity of powerful (soft) semantics.

Implicit Semantics

This type of semantics refers to the kind that is implicit from the patterns in data and that is not represented explicitly in any strict machine processable syntax. Examples of this sort of semantics are the kind implied in the following scenarios:

- Co-occurrence of documents or terms in the same cluster after a clustering process based on some similarity measure is completed.
- A document linked to another document via a hyperlink, potentially associating semantic metadata describing the concepts that relate the two documents.
- The sort of semantics implied by two documents belonging to categories that are siblings of each other in a concept hierarchy.
- Automatic classification of a document to broadly indicate what a document is about with respect to a chosen taxonomy. Further, use the implied semantics to disambiguate (does the word "palm" in a document refer to a palm tree, the palm of your hand, or a palm-top computer?).
- Bioinformatics applications that exploit patterns like sequence alignment, secondary and tertiary protein structure analysis, and so forth

One may argue that although there is no strict syntactic and explicit representation, the knowledge about patterns in data may yet be machine processable. For instance, it is possible to get a numeric similarity judgment between documents in a corpus. Although this is possible, this is the only sort of processing possible. It is not possible to look at documents and automatically infer the presence of a named relationship between concepts in the documents.

Even though the exploitation of implicit semantics draws upon well-known statistical techniques, the wording is not a mere euphemism, but meant to give a different perception of the problem.

Many tools and applications for implicit semantics have been developed for decades and are readily available. Basically all machine learning exploits implicit semantics, namely clustering, concept and rule learning, Hidden Markov Models, Artificial Neural Networks, and others. These techniques supporting implicit semantics are found in early steps towards the Semantic Web, such as clustering in the Vivisimo search engine, as well as in early Semantic Web products, such as metadata extraction on Web Fountain technology (Dill et al., 2003), automatic classification, and automatic metadata extraction in Semagix Freedom (Sheth et al., 2002).

Formal Semantics

Humans communicate mostly through language. Natural language, however, is inherently ambiguous—semantically, but also syntactically. Computers lack the ability to disambiguate and understand complex natural language. For these reasons, it is infeasible to use natural language as a means for machines to communicate with other machines. As a first step, statements or facts need to be expressed in a way that computers can process them. Semantics that are represented in some well-formed syntactic form (governed by syntax rules) is referred to as *formal semantics*. There are some necessary and sufficient features that make a language formal and by association their semantics formal. These features include:

- *The Notions of Model and Model Theoretic Semantics:* Expressions in a formal language are *interpreted* in *models*. The structure common to all models in which a given language is interpreted (the *model structure* for the model-theoretic interpretation of the given language) reflects certain basic presuppositions about the "structure of the world" that are implicit in the language.
- *The Principle of Compositionality:* The meaning of an expression is a function of the meanings of its parts and of the way they are syntactically combined. In other words, the semantics of an expression is computed using the semantics of its parts, obtained using an interpretation function.

From a less technical perspective, formal semantics means machine-processable semantics where the formal language representing the semantics has the above-mentioned features. Basically, the semantics of a statement are unambiguously expressed in the syntax of the statement in the formal language. A very limited subset of natural language is thus made available for computer processing. Examples of such semantics are:

- The semantics of subsumption in Description Logics, reflecting the human tendency of categorizing by means of broader or narrower descriptions.
- The semantics of Partonomy, accounting for what is part of an object, not which category the object belongs to.

Description Logics

Recently, description logics have been the dominant formalisms for knowledge representation. Although DLs have gained substantial popularity, there are some fundamental properties of DLs that can be seen as drawbacks when viewed in the context of the Semantic Web and its future. The *formal semantics* of DLs is based on set theory. A concept in description logics is interpreted as a set of things that share one required common feature. Relationships between concepts or roles are interpreted as a subset of the cross-product of the domain of interpretation. This leaves no scope for the representation of degrees of concept membership or uncertainty associated with concept membership.

DL-based representation and reasoning for both schema and instance data is being applied in Network Inference's Cerebra product for such problems as data integration. This product uses a highly optimized tableaux algorithm to speed up ABox reasoning, which was the bane of description logics. Although a favorable trade-off between computational complexity and expressive power has been achieved, there is still the fundamental issue of the inability of DLs to allow for representation of fuzzy and probabilistic knowledge.

Powerful (Soft) Semantics

The statistical analysis of data allows the exploration of relationships that are not explicitly stated. Statistical techniques give us great insight into a corpus of documents or a large collection of data in general, when a program exists that can actually "pose the right questions to the data," that is, analyze the data according to our needs. All derived relationships are statistical in nature, and we only have an idea or a likelihood of their validity.

The above-mentioned formal knowledge representation techniques give us certainty that the derived knowledge is correct, provided the explicitly stated knowledge was correct in the first place. Deduction is truth preserving. Another positive aspect of a formal representation is its universal usability. Every system that adheres to a certain representation of knowledge will understand, and a well-founded formal semantics guarantees that the expressed statements are interpreted the same way on every system. The restriction of expressiveness to a subset of FOL also allows the system to verify the consistency of its knowledge.

But here also lies the crux of this approach. Even though it is desirable to have a consistent knowledge base, it becomes impractical as the size of the knowledge base increases or as knowledge from many sources is added. It is rare that human experts in most scientific domains have a full and complete agreement. In these cases it becomes more desirable that the system can deal with inconsistencies.

Sometimes it is useful to look at a knowledge base as a map. This map can be partitioned according to different criteria, for example, the source of the facts or their domain. While on such a map the knowledge is usually locally consistent, it is almost impossible and practically infeasible to maintain a global consistency. Experience in developing the Cyc ontology demonstrated this challenge. Hence, a system must be able to identify sources of inconsistency and deal with contradicting statements in such a way that it can still produce derivations that are reliable.

In the traditional bivalent-logic-based formalisms, we — that is, the users or the systems — have to make a decision. Once two contradictory statements are identified, one has to be chosen as the right one. While this is possible in domains that are axiomatized, fully explored, or in which statements are true by definition, it is not possible for most scientific domains. In the life sciences, for instance, hypotheses have to be evaluated, contradicting statements have promoting data, and so forth. Decisions have to be deferred until enough data is available that either verifies or falsifies the hypothesis. Nevertheless, it is desirable to express these hypotheses formally to have means to computationally evaluate them on the one hand and to exchange them between different systems on the other.

In order to allow the sort of reasoning that would allow this, the expressiveness of the formalism needs to be increased. It is known that increasing the expressive power of a KR language causes problems relating to computability. This has been the main reason for limiting the expressive power of KR languages. The real power behind human reasoning, however, is the ability to do so in the face of imprecision, uncertainty, inconsistencies, partial truth, and approximation. There have been attempts made in the past at building KR languages that allow such expressive power.

Major approaches to reasoning with imprecision are: (1) probabilistic reasoning, (2) possibilistic reasoning (Dubois, Lang, & Prade, 1994), and (3) fuzzy reasoning. Zadeh (2002) proposed a formalism that combines fuzzy logic with probabilistic reasoning to exploit the merits of both approaches. Other formalisms have focused on resolving local inconsistencies in knowledge bases, for instance the works of Blair, Kifer, Lukasiewicz, Subrahmanian, and others in annotated logic and paraconsistent logic (see Kifer & Subrahmanian, 1992; Blair & Subrahmanian, 1989). Lukasiewicz

(2004) proposes a weak probabilistic logic and addresses the problem of inheritance. Cao (2000) proposed an annotated fuzzy logic approach that is able to handle inconsistencies and imprecision; Straccia (e.g., 1998, 2004) has done extensive work on fuzzy description logics. With P-CLASSIC, Koller, Levi, and Peffer (1997) presented an early approach to probabilistic description logics implemented in Bayesian Networks. Other probabilistic description logics have been proposed by Heinsohn (1994) and Jaeger (1994). Early research on Bayesian-style inference on OWL was done by Ding and Peng (2004). In her formalism, OWL is augmented to represent prior probabilities. However, the problem of inconsistencies arising through inheritance of probability values (see Lukasiewicz, 2004) is not taken into account.

The combination of probabilistic and fuzzy knowledge under one representation mechanism proposed in Zadeh (2002) appears to be a very promising approach. Zadeh argues that fuzzy logics and probability theory are "complementary rather than competitive." Under the assumption that humans tend to linguistically categorize a continuous world into discrete classes, but in fact still perceive it as continuous, fuzzy set theory classifies objects into sets with fuzzy boundaries and gives objects degrees of set membership in different sets. Hence it is a way of dealing with a multitude of sets in a computationally tractable way that also follows the human perception of the world. Fuzzy logic allows us to blur artificially imposed boundaries between different sets. The other powerful tool in soft computing is probabilistic reasoning. Definitely in the absence of complete knowledge of a domain and probably even in its presence, there is a degree of uncertainty or randomness in the ways we see real-world entities interact. OWL as a description language is meant to explicitly represent knowledge and to deductively derive implicit knowledge. In order to use a similar formalism as a basis for tools that help in the derivation of new knowledge, we need to give this formalism the ability to be used in abductive or inductive reasoning. Bayesian-type reasoning is a way to do abduction in a logically feasible way by virtue of applying probabilities. In order to use these mechanisms, the chosen formalism needs to express probabilities in a meaningful way, that is, a reasoner must be able to meaningfully interpret the probabilistic relationships between classes and between instances. The same holds for the representation of fuzziness. The formalism must give a way of defining classes by their membership functions.

A major drawback of logics dealing with uncertainties is the required assignment of prior probabilities and/or fuzzy membership functions. Obviously, there are two ways of doing that — manual assignment by domain experts and automatic assignment using techniques such as machine learning. Manual assignments require the domain expert to assign these values to every class and every relationship. This assignment will be arbitrary, even if the expert has profound knowledge of the domain. Automatic assignments of prior values require a large and representative dataset of annotated instances, and finding or agreeing on what is a representative set is difficult or at times impossible. Annotating instances instead of categorizing them in a top-down approach is tedious and time consuming. Often, however, the probability values for relationships can be obtained from the dataset using statistical methods, thus we categorize these relationships as implicit semantics.

Another major problem here is that machine learning usually deals with flat categories rather than with hierarchical categorizations. Algorithms that take these hierarchies into account need to be developed. Such an algorithm needs to change the prior values of the superclasses according to the changes in the subclasses, when necessary. Most likely, the best way will be a combination of both, when the domain expert assigns prior values that have to be validated and refined using a testing set from the available data.

In the end, powerful semantics will combine the benefits of both worlds: hierarchical composition of knowledge and statistical analysis; reasoning on available information, but with the advantage over statistical methods that it can be formalized in a common language and that general purpose reasoners can utilize it, and with the advantage over traditional formal DL representation that it allows abduction as well as induction in addition to deduction.

It might be argued that more powerful formalisms are already under development, such as SWRL (Straccia, 1998), which works on top of OWL. These languages extend OWL by a function-free subset of first-order logics, allowing the definition of new rules in the form of Horn clauses. The paradigm is still that of bivalent FOLs, and the lack of function symbols makes it impossible to define functions that can compute probability values. Furthermore, SWRL is undecidable. We believe that abilities to express probabilities and fuzzy membership functions, as well as to cope with inconsistencies, are important. It is desirable (and some would say necessary) that the inference mechanism is sound and complete with respect to the semantics of the formalism and the language is decidable. Straccia (1998) proves this for a restricted fuzzy DL; Giugno and Lukasiewicz (2002) prove soundness and completeness for the probabilistic description logic formalism P-SHOQ(D).

So far, this powerful semantic and soft computing research has not been utilized in the context of developing the Semantic Web. In our opinion, for this vision to become a reality, it will be necessary to go beyond RDFS and OWL, and work towards standardized formalisms that support powerful semantics.

CORRELATING SEMANTIC CAPABILITIES WITH TYPES OF SEMANTICS

Building practical Semantic Web applications (e.g., see TopQuadrant, 2004; Sheth & Ramakrishnan, 2003; Kashyap & Shklar, 2002) require certain core capabilities. A quick look at these core capabilities reveals a sequence of steps towards building such an application. We group this sequence into two categories as shown in Table 1 and identify the type of semantics utilized by each.

APPLICATIONS AND TYPES OF SEMANTICS THEY EXPLOIT

In this section we describe some research fields and some specific applications in each field. This list is by no means a comprehensive list, but rather samples of some research areas that attempt solve problems that are crucial to realizing the Semantic Web vision. We cover *information integration*, *information extraction/retrieval*, *data mining*, and *analytical applications*. We also discuss *entity identification/disambiguation* in some detail. We associate with each of the techniques in these research areas one or more of the types of semantics we identified earlier.

Information Integration

There is, now more than ever, a growing need for several information systems to interoperate in a seamless manner. This sort of interoperation requires that the syntactic, structural, and semantic heterogeneities (Hammer & McLeod, 1993; Kashyap & Sheth, 1996) between such information systems be resolved. Resolving such heterogeneities has been the focus of a lot of work in schema integration in the past. With the recent interest in the Semantic Web, there has been a renewed interest in resolving such heterogeneities. A survey of schema matching techniques (Rahm & Bernstein, 2001) identifies a wide variety of techniques that are deployed to solve this problem.

Schema Integration

A look at the leaf nodes and the level immediately above it, in the classification tree of schema matching techniques in Rahm and Bernstein (2001), reveals the combination of the technique used and the type of information about the schema used for matching schemas. Depending on whether the schema or the instances are used to determine the match, the type of information harnessed varies. Our aim is to associate one or more types of semantics (from our classification) with each of the bulleted entries at the leaf nodes of the tree shown. Table 1 does just that.

Table 1. Some key semantic capabilities and the type of semantics exploited

	Capabilities	*Implicit Semantics*	*Formal Semantics*	*Possible Use of Powerful (Soft) Semantics*
Bootstrapping Phase (building phase)	Building ontologies either automatically or semi-automatically	Analyzing word co-occurrence patterns in text to learn taxonomies/ontologies (Kashyap et al., 2003)		Using fuzzy or probabilistic clustering to learn taxonomic structures or ontologies
	Annotation of unstructured content *wrt.* these ontologies (resulting in semantic metadata)	Analyzing word occurrence patterns or hyperlink structures to associate concept names from and ontology with both resources and links between them (Naing, Lim, & Goh, 2002)		Using fuzzy or probabilistic clustering to learn taxonomic structures or ontologies OR Using fuzzy ontologies
	Entity disambiguation	Using clustering techniques or support vector machines (SVMs) for entity disambiguation (Han, Giles, Zha, Li, & Tsioutsiouliklis, 2004)	Using an ontology for entity disambiguation	Using fuzzy KR mechanisms to represent ontologies that may be used for disambiguation
	Semantic integration of different schemas and ontologies	Analyzing the extension of the ontologies to integrate them (Wang, Wen, Lochovsky, & Ma, 2004)	Schema-based integration techniques (Castano, Antonellis, & Vimercati, 2001)	
	Semantic metadata enrichment (further enriching the existing metadata)	Analyzing annotated resources in conjunction with an ontology to enhance semantic metadata (Hammond et al., 2002)		This enrichment could possibly mean annotating with fuzzy ontologies

Table 1. Some key semantic capabilities and the type of semantics exploited (cont.)

	Capabilities	*Implicit Semantics*	*Formal Semantics*	*Possible Use of Powerful (Soft) Semantics*
Utilization Phase	Complex query processing		Hypothesis validation queries (Sheth, Thacker, & Patel, 2003) or path queries (Anyanwu & Sheth, 2002)	
	Question answering (QA) systems[1]	Word frequency and other CL techniques to analyze both the question and answer sources (Ramakrishnan, Chakrabarti, Paranjpe, & Bhattacharya, 2004)	Using *formal* ontologies for QA (Atzeni et al., 2004)	Providing confidence levels in answers based on fuzzy concepts or probabilistic
	Concept-based search[1]	Analyzing occurrence of words that are associated with a concept, in resources	Using hypernymy, partonomy, and hyponymy to improve search (Townley, 2000)	
	Connection and pattern explorer[1]	Analyzing semi-structured data stores to extract patters (technique in Kuramochi & Karypis, 2004, applied to RDF graphs)	Using ontologies to extract patterns that are meaningful (Aleman-Meza, Halaschek, & Sahoo, 2003)	
	Context-aware retriever[1]	Word frequency and other CL techniques to analyze resources that match the search phrase	Using formal ontologies to enhance retrieval	Using fuzzy KR mechanisms to represent context
Utilization Phase	Dynamic user interfaces[1]		Using ontologies to dynamically reconfigure user interfaces (Quan & Karger, 2004)	
	Interest-based content delivery[1]	Analyzing content to identify concept of content so as to match with interest profile	User profile will have ontology associated with it which contains concepts of interest	
	Navigational and research (Guha, McCool, & Miller, 2003) search	Navigational searches will need to analyze unstructured content	Discovery style queries (Anyanwu & Sheth, 2002) on semi-structured data which is a combination of implicit and formal semantics	Fuzzy matches for research search results

Entity Identification/Disambiguation (EI/D)

A much harder, yet fundamental (and related) problem is that of *entity identification/disambiguation*. This is the problem of identifying that two entities are in fact either the same but treated as being different or that they are in fact two different entities that are being treated as one entity. Techniques used for *identification/disambiguation* vary widely depending on the nature of the data being used in the process. If the application uses unstructured text as a data source, then the techniques used for EI/D will rely on *implicit*

Table 2. Techniques used for schema integration and the type of semantics they exploit

	Type of Information Used	What Does it Mean?	Types of Semantics Exploited
Linguistic Techniques	Name Similarity	Using canonical name representations, synonymy, hypernymy, string edit distance, pronunciation, and N-gram-like techniques to match schemas' attribute and relation names	*Implicit Semantics* are exploited by string edit distance, pronunciation, and N-gram-like techniques. *Formal Semantics* are exploited by synonymy, etc.
	Description Similarity	Processing natural language descriptions associated with attributes and relations	*Implicit Semantics* are exploited by the NLP techniques deployed.
	Word Frequencies of Key Terms	Using relative frequencies of keywords and word combinations at the instance level	*Implicit Semantics*
Constraint Based Techniques	Type Similarity	Using information about data types of attributes as an indicator of a match between schemas	*Formal Semantics*
	Key Properties	Using foreign keys, part-of relationships, and other constraints	*Formal Semantics*
	Graph Matching	Treating the structure of schemas as graph algorithms to determine match degree; between graphs are used to match schemas.	Combination of *Implicit* and *Formal Semantics*
	Value Patterns and Ranges	Using ranges of attributes and patterns in the value of attributes as an indicator of similarity between the corresponding schemas	*Implicit Semantics*

semantics. On the other hand, if EI/D is being attempted on semi-structured data, the application can, for instance, disambiguate entities based on the properties they have. This implies harnessing the power of *formal* or *semi-formal semantics*. As listed in Table 1, the constraint-based techniques are ideally suited for use in EI/D when semi-structured data is being used. Dealing with unstructured data will require the use of the techniques listed under linguistic techniques.

Information Retrieval and Information Extraction

Let us consider information retrieval applications and the types of data they exploit. Given a request for information by the user, information retrieval applications have the task of processing unstructured (text corpus) or loosely connected documents (hyperlinked Web pages) to answer the "query." There are various flavors of such applications.

Search Engines

Search engines exploit both the content of Web documents and the structure *implicit* from the hyperlinks connecting one document to the other. Kleinberg (1998) defines the notions of hubs and authorities in a hyperlinked environment. These notions are crucial to the structural analysis and the eventual indexing of the Web. A modification of this approach aimed at achieving scalability is used by Google (Brin

& Page, 1998). Google has fairly good precision and recall statistics. However, the demands that the Semantic Web places on search engine technology will mean that future search engines will have to deal with information requests that are far more demanding. Guha et al. (2003) identify two kinds of searches:

- *Navigational Searches:* In this class of searches, the user provides the search engine with a phrase or combination of words which s/he expects to find in the documents. There is no straightforward, reasonable interpretation of these words as denoting a concept. In such cases, the user is using the search engine as a navigation tool to navigate to a particular intended document. Using the domain knowledge as specified in relevant domain ontology can enable an improved semantic search (Townley, 2000).
- *Research Searches:* In many other cases, the user provides the search engine with a phrase that is intended to denote an object about which the user is trying to gather/research information. There is no particular document that the user knows about that s/he is trying to get to. Rather, the user is trying to locate a number of documents, which together will give her/him the information s/he is trying to find.

We believe that research searches will require a combination of *implicit semantics, formal semantics,* and what we refer to as *powerful semantics.*

Question Answering Systems

Question answering systems can be viewed as more advanced and more "intelligent" search engines. Current question-answering systems (Brin & Page, 1998; Etzioni et al., 2004; Ramakrishnan et al., 2004) use Natural Language Processing (NLP) and pattern matching techniques to analyze both the question asked of the system and the potential sources of the answers. By comparing the results of these analyses, such systems attempt to match portions of the sources of the answer (for instance, Web pages) with the question, thereby answering them. Such systems therefore still use data like unstructured text and attempt to extract information from it. In other words the semantics are *implicit* in the text and are extracted from this text. To facilitate question answering, Zadeh (2003) proposes the use of an epistemic lexicon of world knowledge, which would be represented by a weighted graph of objects with uncertain attributes; in our terminology this is the equivalent of an ontology using *powerful semantics*.

Data Mining

The goal of data mining applications is to find non-trivial patterns in unstructured and structured data.

Clustering

Clustering is defined as the process of grouping *similar* entities or objects together in groups based on some notion of similarity. Clustering is considered a form of *unsupervised learning*. The applications of clustering use a given similarity metric and, as a result of the grouping of data points into clusters, attempt to use this information *(implicit semantics)* to learn something about the interactions between the clustered entities. The sort of information sought from the clustered data points may range from simple similarity judgments as in Query-By-Example (QBE) document retrieval systems or systems aimed at extracting *more formal* semantics from the underlying data, as is the aim of Semi-Automatic Taxonomy Generation.

Semi-Automatic Taxonomy Generation (ATG)

As described in Kashyap et al. (2003), the aim of Automated Taxonomy Generation is to hierarchically cluster a document corpus and extract from the resulting hierarchy of clusters a *sequence* of clusters that best captures all the levels of specificity/generality in the corpus, where this sequence is ordered by the value of the specificity/generality measure. This is then followed by a node label extraction phase, where each cluster in the sequence is analyzed to extract from it a set of labels that best captures the topic its documents represent. These sets of labels are then pruned to reduce the number of potential labels for nodes in the final output hierarchy.

Association Rule Mining

An example of an association rule is given in Agrawal, Imielinski, and Swami (1993) and Agrawal and Srikant (1994) as follows: 90% of the transactions in a transaction database that involve the purchase of bread and butter together also have the purchase of milk involved. This is an example of an application where occurrence patterns of attribute values in a relational database *(implicit semantics)* are converted in association rules *(formal semantics)*.

Analytical Applications

These come under the purview of applications that support complex query processing. It would be reasonable to hypothesize that search engines of the future will be required to answer analytical or discovery style queries (Guha et al., 2003; Anyanwu & Sheth, 2002). This is in sharp contrast to the kinds of information requests today's search engines have to deal with, where the focus is on retrieving resources from the Web that may contain information about the desired keyword. In this current scenario the user is left to sift through vast collections of documents and further analyze the returned results. In addition to querying data from the Web, future search engines will also have to query vast metadata repositories. We discuss one such technique in the following section.

Complex Relationship Discovery

As described in Anyanwu and Sheth (2002):

"Semantic Associations capture complex relationships between entities involving sequences of predicates, and sets of predicate sequences that interact in complex ways. Since the predicates are semantic metadata extracted from heterogeneous multi-source documents, this is an attempt to discover complex relationships between objects described or mentioned in those documents. Detecting such associations is at the heart of many research and analytical activities that are crucial to applications in national security and business intelligence."

The datasets that Semantic Associations operate over are RDF/RDFS graphs. The semantics of an edge connecting two nodes in an RDF/RDFS graph are *implicit*, in the sense that there is no explicit interpretation of the semantics of the edge other than the fact that it is a predicate in a statement (except for *rdfs:subPropertyOf* or edges that represent data type properties

— for which there is model-theoretic (formal) semantics). Hence the RDF/RDFS graph is composed of a combination of *implicit* and *formal* semantics. The objective of Semantic Associations is therefore to find all contextually relevant edge sequences that relate two entities. This is in effect an attempt to combine the *implicit* and *formal* semantics of the edges in the RDF/RDFS graph in conjunction with the context of the query to determine the *multifaceted (multivalent) semantics* of a set of "connections" between entities. We view this *multivalent semantics* as a form of *powerful semantics*. In the context of search, Semantic Associations can be thought of as a class of research searches or discovery-style searches.

CONCLUSION

We have identified three types of semantics and in the process assorted key capabilities required to build a practical semantic application involving Web resources. We have also qualified each of the listed capabilities with one or more types of semantics, as in Table 1. This table reveals some very basic problems that need to be solved for an application to be termed "semantic." It is clear from this table that *entity disambiguation, question answering capability, context-based retrieval,* and *navigational and research (discovery) style query capability* require the use of all three types of semantics. Therefore by focusing research efforts in representation mechanisms for *powerful (soft) semantics* in conjunction with fuzzy/probabilistic computational methods supporting techniques that use implicit and formal semantics, it might be possible to solve some of the difficult but practically important problems. In our opinion the current view taken by the Semantic Web community is heavily biased in favor of *formal semantics*. It is clear, however, that the focus of effort in pursuit of the Semantic Web vision needs to move towards an approach that encompasses all three types of semantics in representation, creation methods, and analysis of knowledge. If the capabilities that we identified do in fact turn out to be fundamental capabilities that make an application semantic, these capabilities could serve as a litmus test or a standard against which other applications may be measured to determine if they are "semantic applications."

REFERENCES

Agrawal, R., Imielinski, T., & Swami, A.N. (1993). Mining association rules between sets of items in large databases. In P. Buneman & S. Jajodia (Eds.), *Proceedings of the 1993*.

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. *Proceedings of the 20th International Conference on Very Large Databases,* Santiago, Chile.

Aleman-Meza, B., Halaschek, C., & Sahoo, S. (2003). *Terrorist-related assessment using knowledge similarity*. LSDIS Lab Technical Report, December.

Anyanwu, K., & Sheth, A.P. (2002). The p operator: Discovering and ranking associations on the Semantic Web. *SIGMOD Record, 31*(4), 42-47.

Atzeni, P., Basili, R., Hansen, D.H., Missier, P., Paggio, P., Pazienza, M.T., & Zanzotto, F.M. (2004, June). Ontology-based question answering in a federation of university sites: The MOSES case study. *Proceedings of the 9th International Conference on Applications* *of Natural Language to Information Systems* (NLDB'04), Manchester, UK.

Barbará, D., Garcia-Molina, H., & Porter, D. (1992). The management of probabilistic data. *IEEE Transactions on Knowledge and Data Engineering, 4*(5), 487-502.

Blair, H.A., & Subrahmanian, V.S. (1989). Paraconsistent logic programming. *Theoretical Computer Science, 68,* 135-154.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Proceedings of the 12th International World Wide Web Conference.*

Cao, T.H. (2000). Annotated fuzzy logic programs. *International Journal on Fuzzy Sets and Systems, 113,* 277-298.

Castano, S., Antonellis, V.D., & Vimercati, S.D.C. (2001). Global viewing of heterogeneous data sources. *IEEE Transactions on Knowledge and Data Engineering, 13*(2), 277-297.

Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J.A., & Zien, J.Y. (2003). Semtag and seeker: Bootstrapping the Semantic Web via automated semantic annotation. *Proceedings of the 12th International Conference on World Wide Web* (pp. 178-186). ACM Press.

Ding, Z., & Peng, Y. (2004, January 5-8). A probabilistic extension to ontology language OWL. *Proceedings of the Hawaii International Conference on System Sciences,* Big Island, Hawaii.

Dubois, D., Lang, J., & Prade, H. (1994). Possibilistic logic. In D.M. Gabbay et al. (Eds.), *Handbook of logic in artificial intelligence and logic programming* (Vol. 3, pp. 439-514). Oxford: Oxford University Press.

Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D.S., & Yates, A. (2004). Web-scale information extraction in knowitall (preliminary results). *WWW,* 100-110.

Giugno, R., & Lukasiewicz, T. (2002). P-SHOQ(D): A probabilistic extension of SHOQ(D) for probabilistic ontologies in the Semantic Web. *Proceedings of the European Conference on Logics in Artificial Intelligence.*

Gruber, T. (2003. March 26-27). It is what it does: The pragmatics of ontology. Invited talk at *Sharing the Knowledge—International CIDOC CRM Symposium,* Washington, DC. Retrieved from *tomgruber.org/writing/cidoc-ontology.htm*

Guha, R., McCool, R., & Miller, E. (2003, May). Semantic search. *Proceedings of the 12th International World Wide Web Conference*, Budapest, Hungary.

Hammer, J., & McLeod, D. (1993). An approach to resolving semantic heterogeneity in a federation of autonomous, heterogeneous database systems. *Journal for Intelligent and Cooperative Information Systems*.

Hammond, B., Sheth, A., & Kochut, K. (2002). Semantic enhancement engine: A modular document enhancement platform for semantic applications over heterogeneous content. In V. Kashyap & L. Shklar (Eds.), *Real-world Semantic Web applications* (pp. 29-49). IOS Press.

Han, H., Giles, C.L., Zha, H., Li, C., & Tsioutsiouliklis, K. (2004). Two supervised learning approaches for name disambiguation in author citations. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries* (JCDL 2004) (pp. 296-305).

Handschuh, S., Staab, S., & Ciravegna, F. (2002, October 1-4). S-CREAM—semi-automatic CREAtion of metadata. *Pro-*

ceedings of the European Conference on Knowledge Acquisition and Management (EKAW-2002), Madrid, Spain. Berlin: Springer-Verlag.

Heinsohn, J. (1994). *Probabilistic description logics* (pp. 311-318). UAI.

Hjorland, B. (1998). Information retrieval, text composition, and semantics. *Knowledge Organization, 25*(1/2), 16-31.

Jaeger, M. (1994). *Probabilistic reasoning in terminological logics* (pp. 305-316). KR.

Kashyap, V., Ramakrishnan, C., Thomas, C., Bassu, D., Rindflesch, T.C., & Sheth, A. (2003). *TaxaMiner: An experimentation framework for automated taxonomy bootstrapping*. Technical Report No. UGA-CS-TR-04-005, Computer Science Department, University of Georgia, USA.

Kashyap, V., & Sheth, A. (1996). Semantic heterogeneity in global information systems: The role of metadata, context and ontologies. *Cooperative Information Systems*.

Kashyap, V., & Shklar, L. (Eds). (2002). *Real-world Semantic Web applications—Volume 92: Frontiers in artificial intelligence and applications*.

Kifer, M., & Subrahmanian, V.S. (1992). Theory of generalized annotated logic programming and its applications. *Journal of Logic Programming, 12,* 335-367.

Kleinberg, J. (1998). Authoritative sources in a hyperlinked environment. *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*.

Koller, D., Levy, A., & Pfeffer, A. (1997). P-CLASSIC: A tractable probabilistic description logic. *Proceedings of the 14th National Conference on Artificial Intelligence* (AAAI-97) (pp. 390-397).

Kuramochi, M., & Karypis, G. (2004). Finding frequent patterns in a large sparse graph. *Proceedings of the SIAM International Conference on Data Mining* (SDM-04).

Lukasiewicz, T. (2004). *Weak nonmonotonic probabilistic logics, principles of knowledge representation and reasoning.* KR.

Maedche, A., & Staab, A. (2001). Ontology learning for the Semantic Web. *IEEE Intelligent Systems, 16*(2), 72-79.

Naing, M.-M., Lim, E.-P., & Goh, D.H.-L. (2002). Ontology-based Web annotation framework for hyperlink structures. *Proceedings of WISE Workshops 2002* (pp. 184-193).

Omelayenko, B. (2001). Learning of ontologies for the Web: The analysis of existent approaches. *Proceedings of the International Workshop on Web Dynamics*.

Patil, A., Oundhakar, S., Sheth, A., & Verma, K. (2004, May). METEOR-S Web service annotation framework. *Proceedings of the World Wide Web Conference* (pp. 553-562). New York.

Quan, D., & Karger, D.R. (2004). *How to make a Semantic Web browser in WWW.*

Rahm, E., & Bernstein, P.A. (2001). A survey of approaches to automatic schema matching. *VLDB Journal, 10,* 4.

Ramakrishnan, G., Chakrabarti, S., Paranjpe, D., & Bhattacharya, P. (2004). Is question answering an acquired skill? *Proceedings of the 13th International Conference on the World Wide Web 2004*.

Sheth, A., Bertram, C., Avant, D., Hammond, B., Kochut, K., & Warke, Y. (2002). Managing semantic content for the Web. *IEEE Internet Computing*, (July/August), 80-87.

Sheth, A.P., Thacker, S., & Patel, S. (2003). Complex relationships and knowledge discovery support in the InfoQuilt system. *VLDB Journal, 12*(1), 2-27.

Sheth, A.P., & Ramakrishnan, C. (2003). Semantic (Web) technology in action: Ontology-driven information systems for search, integration and analysis. *IEEE Data Engineering Bulletin*, 26(4), 40-48.

Straccia, U. (2004). *Uncertainty and description logic programs: A proposal for expressing rules and uncertainty on top of ontologies.* Technical Report 2004-TR-14.

Straccia, U. (1998). A fuzzy description logic. *Proceedings of AAAI-98, 15th Conference of the American Association for Artificial Intelligence.*

Townley, J. (2000). The streaming search engine that reads your mind. Retrieved August 10, 2000: *smw.internet.com/gen/reviews/searchassociation/*

Uschold, M. (2003). Where are the semantics in the Semantic Web? *Artificial Intelligence,* (Fall).

Wang, J., Wen, J.-R., Lochovsky, F.H., & Ma, W.-Y. (2004). Instance-based schema matching for Web databases by domain-specific query probing. *Proceedings of the 2004 Conference on VLDBs.*

Woods, W.A., (2004, June 2-5). Meaning and links: A semantic odyssey. *Principles of Knowledge Representation and Reasoning: Proceedings of the 9th International Conference* (KR2004) (pp. 740-742).

Yen, J. (1991). Generalizing term subsumption languages to fuzzy logic. *IJCAI*, 472-477.

Zadeh, L.A. (2002). Toward a perception-based theory of probabilistic reasoning with imprecise probabilities. *Journal of Statistical Planning and Inference, 105*, 233-264.

Zadeh, L.A. (2003). From search engines to question-answering systems—the need for new tools. *Proceedings of the 1st Atlantic Web Intelligence Conference.*

ADDITIONAL ONLINE RESOURCES

www.networkinference.com/Assets/Products/Cerebra_Server_Datasheet.pdf

www.topquadrant.com/documents/TQ04_Semantic_Technology_Briefing.PDF

Amit Sheth is an educator, researcher and entrepreneur. He joined the University of Georgia and started the LSDIS lab in 1994. Earlier, he served in R&D groups at Bellcore (now Telcordia Technologies), Unisys, and Honeywell. In August 1999, Sheth founded Taalee, Inc., a VC funded enterprise software and Internet infrastructure startup based on the technology developed at the LSDIS lab. He managed Taalee as its CEO until June 2001. Following Taalee's acquisition/merger, he serves as CTO and co-founder of Semagix, Inc. (formerly Voquette, Inc). His research has led to several commercial products and applications. He has published more than 175 papers and articles (in the areas of semantic interoperability, federated databases, workflow management, Semantic Web), given over 130 invited talks and colloquia including 19 keynotes, (co)-organized/chaired 12 conferences/workshops, and served on over 90 program committees. He is co-editor in chief of the International Journal on Semantic Web and Information Systems, *is*

on five journal editorial board, is a member of the W3C Advisory Committee the Semantic Web Services Architecture committee, etc. For more information, visit the following: http://lsdis.cs.uga.edu/~amit and http://www.semagix.com/company_team.html.

Cartic Ramakrishnan is a PhD student in computer science at the University of Georgia. He graduated from the Univeristy of Pune, India, with a BE in computer engineering. He is currently working at the Large Scale Distributed Information Systems (LSDIS) Lab, University of Georgia. He leads the research efforts in ontology learning and semantic association discovery with other members at the LSDIS lab. For more information, please visit http://lsdis.cs.uga.edu/~cartic.

Christopher Thomas is a PhD student in the LSDIS lab, University of Georgia. His academic background is computer science and artificial intelligence. His main research focus is semantics, knowledge representation and reasoning. Recently he has been working in the field of computational biochemistry to develop ontologies and reasoning mechanisms suitable for the complex carbohydrate domain. His research in representation of uncertain, incomplete and partially inconsistent knowledge is geared towards developing formalism that can be applied beyond the life sciences domain in all Semantic Web areas. Please visit http://lsdis.cs.uga.edu/~cthomas for more information.

Semantic eBusiness

Rahul Singh, The University of North Carolina at Greensboro, USA
Lakshmi Iyer, The University of North Carolina at Greensboro, USA
A.F. Salam, The University of North Carolina at Greensboro, USA

ABSTRACT

We define Semantic eBusiness as "an approach to managing knowledge for coordination of eBusiness processes through the systematic application of Semantic Web technologies." Advances in Semantic Web-based technologies offer the means to integrate heterogeneous systems across organizations in a meaningful way by incorporating ontology—a common, standard, and shareable vocabulary used to represent the meaning of system entities; knowledge representation, with structured collections of information and sets of inference rules that can be used to conduct automated reasoning; and intelligent agents that collect content from diverse sources and exchange semantically enriched information. These primary components of the Semantic Web vision form the foundation technology for semantic eBusiness. The challenge for research in information systems and eBusiness is to provide insight into the design of business models and technical architecture that demonstrate the potential of technical advancements in the computer and engineering sciences to be beneficial to business and consumers. Semantic eBusiness seeks to apply fundamental work done in Semantic Web technologies to support the transparent flow of semantically enriched information and knowledge—including content and know-how—to enable, enhance, and coordinate collaborative eBusiness processes within and across organizational boundaries. Semantic eBusiness processes are characterized by the seamless and transparent flow of semantically enriched information and knowledge. We present a holistic view of semantic eBusiness that integrates emergent and well-grounded Semantic Web technologies to improve the current state of the art in the transparency of eBusiness processes.

Keywords: electronic marketplace; intelligent agents; knowledge management; knowledge services; ontology; Semantic eBusiness; Semantic Web

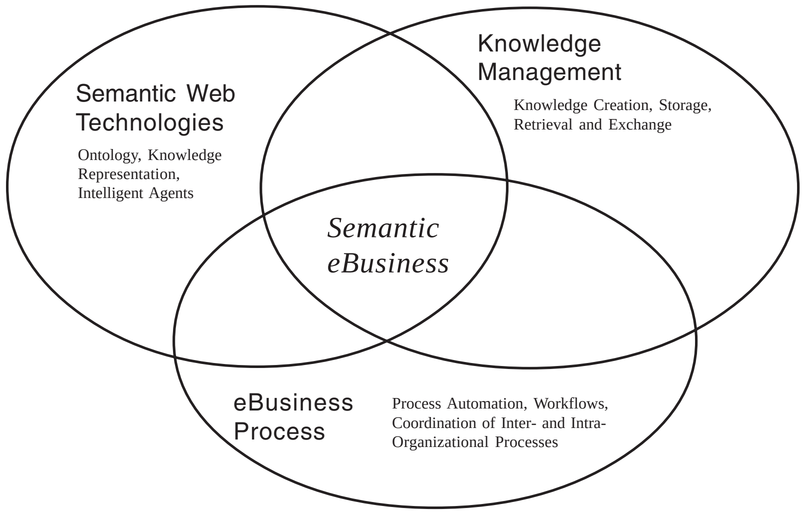
INTRODUCTION

The Semantic Web vision (Berners-Lee, Hendler, & Lassila, 2001) provides the foundation for semantic architecture to support the transparent exchange of information and knowledge among collaborating eBusiness organizations. Recent advances in Semantic Web-based technologies offer means for organizations to exchange knowledge in a *meaningful* way. This requires *ontologies*, to provide a standardized and shareable vocabulary to represent the meaning of system entities; *knowledge representation*, with structured collections of information and sets of inference rules that can be used to conduct automated reasoning; and *intelligent agents* that can exchange semantically enriched information and knowledge, and interpret the knowledge on behalf of the user (Hendler, 2001). It is increasingly clear that semantic technologies have the potential to enhance eBusiness processes. The *challenge* for research in information systems and eBusiness is to provide insight into the design of business models and technical architecture that demonstrate the potential of technical advancements in the computer and engineering sciences to be beneficial to business and consumers.

EBusiness is "*an approach to achieving business goals in which technology for information exchange enables or facilitates execution of activities in and across value chains, as well as supporting decision making that underlies those activities*" (Holsapple & Singh, 2000). Inter-organizational collaborations are effective means for organizations to improve the efficacy of their eBusiness processes and enhance their value propositions. Inter-organizational collaborative business processes require transparent information and knowledge exchange across partner firms. Businesses increasingly operate in a dynamic, knowledge-driven economy and function as knowledge-based organizations. *Knowledge* is defined as the highest order in the continuum of data and information, as having utility and specificity in its context domain. Functionally and in systems, the lines between useful information and knowledge are blurred (Grover & Davenport, 2001). For this research, we define knowledge as "*information, in the context of a specific problem domain, upon which action can be advised or taken*." Knowledge management includes facilities for the creation, exchange, storage, and retrieval of knowledge in an exchangeable and usable format, in addition to the critical facilities to use of knowledge to support business activity (O'Leary, 1998). It is important for eBusiness to explicitly recognize knowledge along with the processes and technologies for knowledge management.

We define Semantic eBusiness as "*an approach to managing knowledge for coordination of eBusiness processes through the systematic application of Semantic Web technologies.*" Semantic eBusiness applies fundamental work done in Semantic Web technologies, including ontologies, knowledge representation, multi-agent systems, and Web-services, to support the *transparent* flow of semantically enriched information and knowledge, including *content* and *know-how,* and enable collaborative eBusiness processes within and across organizational boundaries. In this article, we present an overview of the Semantic eBusiness vision, with emphasis on the conceptual foundations and research directions in Semantic eBusiness. In our view, Semantic eBusiness is founded upon three primary streams of research literature: *Semantic Web technologies*, including ontologies, knowledge Representation and intelligent software agents; *knowledge management,* including the creation, storage and retrieval, and the exchange of machine interpretable and useful information upon which action can be taken or advised; and *eBusiness processes,* including process automation, enterprise systems integration, and the coordination of workflows and activities within and across organizations. We provide a conceptual schematic of this grounding in Figure 1.

Figure 1. Semantic eBusiness vision founded upon existing work in Semantic Web technologies, knowledge management, and in the e-business processes literature



The following sections provide a detailed discussion of these foundations upon which Semantic eBusiness is envisioned. We provide some directions, from our own research initiatives and that of others, leading towards making the Semantic eBusiness vision a reality. Interest in Semantic eBusiness in the information systems community is beginning to gather momentum through the formation of special interest groups in the research and practitioner communities. We provide a description of some of the organizations that are playing an important role in this. This article concludes with a summary and directions for future research in Semantic eBusiness.

FOUNDATIONS

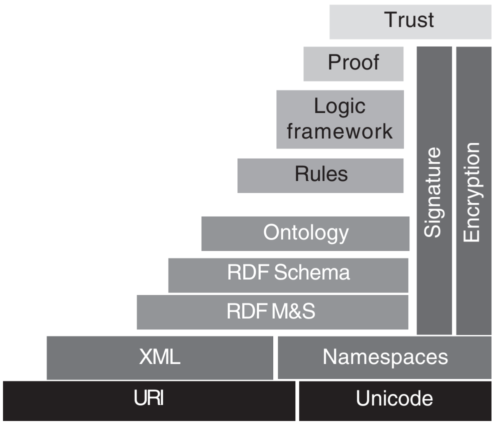
Semantic Web Technologies

The Semantic Web is an extension of the current Web in which information is given "*well-defined meaning*" to allow machines to "*process and understand*" the information presented to them (Berners-Lee et al., 2001).

According to Berners-Lee et al. (2001), the "Semantic Web" comprises and requires the following components in order to function:

- *Knowledge Representation:* Structured collections of information and sets of inference rules that can be used to conduct automated reasoning. Knowledge representations must be linked into a single system.
- *Ontologies:* Systems must have a way to discover common meanings for entity representations. In philosophy, ontology is a theory about the nature of existence; in systems, ontology is a document that formally describes classes of objects and defines the relationship among them. In addition, we need ways to interpret ontology.

Figure 2. Semantic Web architecture (www.w3.org/DesignIssues/diagrams/sw-stack-2002.png; Berners Lee et al., 2001)



Source: http://www.w3.org/DesignIssues/diagrams/sw-stack-2002.png

- *Agents:* Programs that collect content from diverse sources and exchange the result with other programs. Agents exchange "data enriched with semantics."

Intelligent software agents can reach a shared understanding by exchanging ontologies that provide the vocabulary needed for discussion. Agents can even *bootstrap* new reasoning capabilities when they discover new ontologies. Semantics makes it easier to take advantage of a service that only partially matches a request.

"A typical process will involve the creation of a 'value chain' in which subassemblies of information are passed from one agent to another, each one 'adding value,' to construct the final product requested by the end user. Make no mistake: to create complicated value chains automatically on demand, some agents will exploit artificial-intelligence technologies in addition to the Semantic Web." (Berners-Lee et al., 2001)

XML-Based Technologies for Knowledge Representation and Exchange

Technologies for developing meaningful semantic representations of information and knowledge exist through *XML* (eXtensible Markup Language—www.xml.org, www.w3.org/XML/), *RDF* (Resource Description Framework—www.w3.org/RDF/), and *OWL* (Web Ontology language—www.w3.org/TR/owl-features/). XML and its related standards make it feasible to store knowledge in a meaningful way while supporting unambiguous content representation and flexible exchange over heterogeneous platforms (Chiu, 2000). XML allows the creation of customized tags and languages using XML schema, which describe specific elements, the data types in each element, and their relationships. With the appropriate schema, XML documents can be parsed, validated, and processed by application software using XML parsers. Built upon accepted W3C standards, this provides the foundation for semantic technology for the capture, representation, exchange, and storage of knowledge that can be potentially used and shared by software agents. XML provides standardized representations of data structures for processing on heterogeneous systems without case-by-case programming. The use of XML-based technology, including ebXML (www.ebxml.org) and RossettaNet (www.RossettaNet.org), allows for the creation of common vocabularies for eBusiness to help automate business processes, allowing better collaboration and knowledge transfer between partners in semantically integrated systems.

Initiatives to develop technologies for the Semantic Web make the content of the Web unambiguously computer-interpretable to make it amenable to agent interoperability and automated reasoning techniques (McIlraith, Son, & Zeng, 2001). RDF was developed by the W3C as a metadata standard to provide a data model and syntactical conventions to represent data semantics in a standardized interoperable manner (McIlraith et al., 2001). The RDF working group also developed RDF Schema (RDFS), an object-oriented type system that provides an ontology modeling language. Recently, there have been several efforts to build on RDF and RDFS with AI-inspired knowledge representation languages such as SHOE, DAML-ONT, OIL, and DAML+OIL (Fensel, 2000). The Web Ontology Language (OWL) has been standardized by the W3C as a knowledge representation language for the Semantic Web. OWL documents represent domain ontologies and rules, and allow knowledge sharing among agents through the standard *Web services* architecture. Web services technology provides the envelope and transport mechanism for information exchange between software entities. Knowledge exchange architectures use Simple Object Access Protocol (SOAP—www.w3.org/TR/soap/) messages to carry relevant semantic information in the form of OWL documents between agents. The Web services framework consists of the Web Services Definition Language (WSDL—www.wsdl.org), which describes Web services in XML format and provides the basis for tools to create appropriate SOAP messages. These technologies provide the knowledge representation and exchange mechanism to allow collaborating organizations to seamlessly share information and knowledge to coordinate eBusiness processes.

Ontologies

Description logics (DLs) form a basis for developing ontology to further the sharing and use of a common understanding of a specific problem. Description logics model the domain of interest using constructs that describe domain-specific objects and the relationships between them (Baader et al., 2002). Domain-specific objects are represented using the *concept* construct, which is a unary predicate. Relationships between constructs are represented using the *relations* construct, which may be an n-ary predicate. Description logics, at the least, can be used to develop a model of the domain comprising:

- specifications for the creation of complex concept and relation expressions built upon a set of atomic concepts and relations,
- the cumulative set of description logics that forms the basis for a knowledge base containing the properties of domain-dependent concepts and relations specified through a set of assertions on the domain, and
- a set of reasoning procedures that allows suitable inferences from the concepts and the relationships between them.

Ontologies provide a shared and common understanding of specific domains that can be communicated between disparate application systems, and therein provide a means to integrate the knowledge used by online processes employed by eBusiness organizations (Klein et al., 2001). Ontology describes the semantics of the constructs that are common to the online processes, including descriptions of the data semantics that are common descriptors of the domain context. Staab et al. (2001) describe an approach for ontology-based knowledge management through the concept of knowledge metadata, which contains two distinct forms of ontologies that describe the structure of the data itself and issues related to the content of data. We refer the reader to Kishore et al. (2004) for more comprehensive discussion of ontologies and information systems. Ontology documents can be created using FIPA-compliant content languages like BPEL, RDF, OWL, and DAML to generate standardized representations of the process knowledge. The structure of ontology documents will be based on description logics. The recent adoption of the OWL standards by the World Wide Web Consortium (www.w3c.org) includes OWL-DL, which specifies the representation of DL-based models into OWL documents.

In the Semantic eBusiness vision, knowledge exchange and delivery can be facilitated by the availability and exchange of knowledge represented in OWL documents among intelligent software agents. Domain knowledge objects provide an abstraction to create, exchange, and use modular knowledge represented using OWL documents. This allows for a common vocabulary used for exchange of information and knowledge across all system participants. There are many benefits to storing this knowledge in XML format, including standardization of semantics, validation ability and 'well-formedness', ease of use, re-use, and storage. In addition, the ability to exchange complete XML documents in W3C standards affords integration on heterogeneous platforms. All exchanges between agents take place using the standard Web services architecture to allow for platform independence, and facilitate exchange of information and knowledge in OWL documents. Capturing and representing modular knowledge in XML format facilitates their storage in a knowledge repository—a repository that enables storage and retrieval of XML documents of multiple knowledge modules depending upon the problem domain. The benefits of such knowledge repositories are the historical capture of knowledge modules that are available to all agents in the agent community. This ensures that a newly instantiated agent has access to knowledge available to the entire system.

Intelligent Agents

Intelligent agents are action-oriented abstractions in electronic systems, entrusted to carry out various generic and specific goal-oriented actions on behalf of users (Papazoglou, 2001). The agent paradigm can support a range of decision-making activity, including information retrieval, generation of alternatives, preference order ranking of options and alternatives, and supporting analysis of the alternative-goal relationships. An intelligent agent is "*a computer system situated in some environment and that is capable of flexible autonomous action in this environment in order to meet its design objectives*" (Jennings & Wooldridge, 1998). The specific autonomous behavior expected of intelligent agents depends on the concrete application domain and the expected role and impact of intelligent agents on the potential solution for a particular problem for which the agents are designed to provide cognitive support. Criteria for application of agent technology require that the application domain should show *natural distributivity* with autonomous entities that are geographically distributed and work with distributed data; require *flexible interac-*

tion without a priori assignment of tasks to actors; and be embedded in a *dynamic environment* (Muller, 1997).

Intelligent agents are able to organize, store, retrieve, search, and match information and knowledge for effective collaboration among Semantic eBusiness participants. A fundamental implication is that knowledge must be available in formats that allow for processing by software agents. Intelligent agents can be used for knowledge management to support Semantic eBusiness activities. The agent abstraction is created by extending an object with additional features for encapsulation and exchange of knowledge between agents to allow agents to deliver knowledge to users and support decision-making activity (Shoham, 1993). Agents work on a distributed platform and enable the transfer of knowledge by exposing their public methods as Web services using SOAP and XML. In this respect, the interactions among the agents are modeled as collaborative interactions, where the agents in the multi-agent community work together to provide decision support and knowledge-based explanations of the decision problem domain to the user.

Knowledge Management

Emerging business models are causing fundamental changes in organizational and inter-organizational business processes by replacing conflict with cooperation as a means to be economically efficient (Beam, 1998). Operationally, knowledge management (KM) is "*a process that helps organizations find, select, organize, disseminate, and transfer important information and expertise necessary for activities such as problem solving, dynamic learning, strategic planning, and decision making*" (Gupta, Iyer, & Aronson, 2000). From an organizational perspective, it is the management of corporate knowledge that can improve a range of organizational performance characteristics by enabling an enterprise to be more *intelligent acting* (Wiig, 1993). A system managing available knowledge must comprise facilities for the *creation*, *exchange*, *storage,* and *retrieval* of knowledge in an *exchangeable* and usable format, in addition to facilities to use the knowledge in a business activity (O'Leary, 1998). Many organizations are developing KM systems designed specifically to facilitate the exchange and integration of knowledge in business processes for increasing collaboration to gain a competitive advantage.

The Semantic eBusiness vision is built upon transparent information and knowledge exchange across seamlessly integrated systems over globally available Internet technologies to enable *information partnerships* among participants across the entire value chain. Such transparency enhances the utility and extensibility of knowledge management initiatives of an organization by adding the ability to exchange specific and transparent knowledge, utilizing unambiguously interpretable, standards-based representation formats (Singh, Iyer, & Salam, 2003). Implementing and managing such high levels of integration over distributed and heterogeneous information platforms such as the Internet is a challenging task with significant potential benefits for organizations embracing such collaboration. Organizations can gain significant benefits from these initiatives including optimized inventory levels, higher revenues, improved customer satisfaction, increased productivity, and real-time resolution of problems and discrepancies throughout the supply chain. The vision is to achieve dynamic collaboration among business partners and customers throughout a trading community through transparent exchange of semantically enriched information and knowledge.

EBusiness, EBusiness Processes, and E-Marketplaces

Electronic data interchange (EDI) established the preliminary basis for automating business-to-business (B2B) e-commerce (EC) transactions through facilities for organizations to share process information electronically using standardized formats and semantics. Strategies such as supply chain management (SCM) and enterprise resource planning (ERP) go beyond process automation by streamlining and integrating internal and inter-organizational process for improved information availability across value-chain partners. While popular strategies such as SCM and ERP have improved transactional efficiencies, the lack of systems and process integration and the resultant lack of end-to-end value chain visibility continue to hinder collaborative and mutually beneficial partnerships. EBusiness processes require transparent information and knowledge transparency among business partners. The vision is to achieve dynamic collaboration among internal personnel, business partners, and customers throughout a trading community, electronic market, or other form of exchange characterized by the seamless and transparent exchange of *meaningful* information and knowledge. The resultant view is similar to the notions of real-time supply chains and infomediary-based e-marketplaces, where the virtual supply chain is viewed as an inter-organizational information system with seamless and transparent flows of information enabled through highly integrated systems (Rabin, 2003).

The timely sharing of accurate information among collaborating firms and transparency in the supply chain is critical for efficient workflows that support the business processes (Davenport & Brooks, 2004). Information technologies can help streamline business processes across organizations and improve the performance of the value chain by enabling better coordination of inter-firm processes through B2B e-marketplaces (Dai & Kauffman, 2002). The lack of integration of information and knowledge in systems that manage business processes is a stumbling block in enterprise innovation (Badii & Sharif, 2003). The consequent lack of transparencies in information flow across the value chain continue to hinder productive and collaborative partnerships among firms in B2B e-marketplaces. Current e-chains suffer from paucity in information transparency spanning all participant e-marketplaces in the e-supply chain. Integrative systems that support the transparent exchange of information and knowledge can enhance collaboration across organizational value chains by extending support for a range of eBusiness processes and provide aggregate or product-specific cumulative demand or supply conditions in a single e-marketplace and across multiple upstream or downstream links in the e-chain (Singh, Salam, & Iyer, forthcoming). Such systems must provide collaborating value chain partners with intelligent knowledge services capabilities for the seamless and transparent exchange of volatile and dynamic market information, both synchronously and asynchronously.

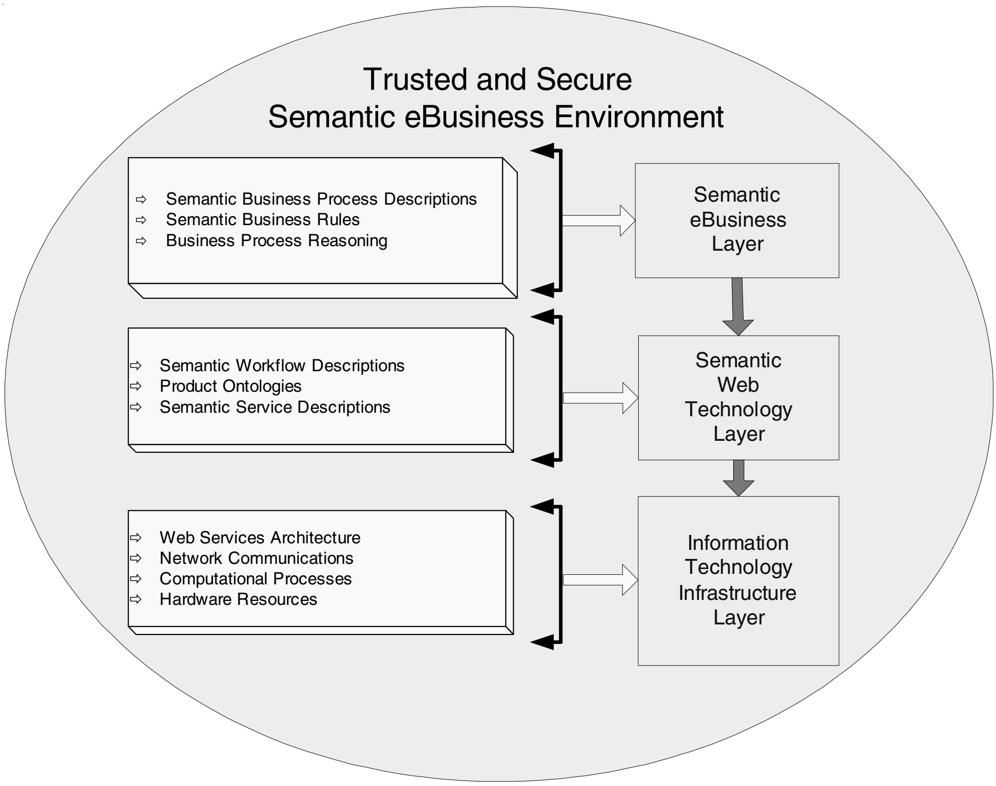
Reductions in transaction coordination costs gained through the effective application of information technologies partly explain the increasing use of markets over hierarchies by organizations to coordinate economic activities (Malone, Yates, & Benjamin, 1987). E-marketplaces offer value-added services by leveraging industry-specific expertise through deciphering complex information and contribute to transaction cost reductions. A survey by Davenport, Brooks, and Cantrell (2001) on B2B e-marketplaces identified lack of trust as a primary barrier for e-marketplace growth. Much of the risk associated with lack of trust can be reduced "*as information becomes more codified, standardized, aggregated, integrated, distributed, and shaped for ready use*" (Davenport et al., 2001). They also state that "*currently achieved e-marketplace integration levels fall far below what is necessary.*" Investments in the IT infrastructure of the e-marketplace can further the effective use of process coordination and communication between participants. While asset-specific technology investments serve to reduce the transaction cost, this leads to significant increases in cost of switching partners. However, when such investments are made by the e-marketplace, the transaction cost reductions can benefit e-marketplace participants, while the increase in switching costs applies to switching from an e-marketplace participant to a non-participant firm.

Integrative technologies that support the transparent exchange of information and knowledge make it easier for the development of inter-organizational relationships through enhanced adaptability and standardization of content representation. This is increasingly prevalent through efforts such as ebXML (www.ebXML.org), Web services, and systems architecture standards, which allow standardization of content representation, with implications for technology adaptation and enterprise applications integration (Davenport & Brooks, 2004). By defining the standards for adaptability and standardization, e-marketplaces can help define the information technology standards that are in use by all participant organizations, allowing for easy interoperability and integration of key systems of participant organizations. In this regard, e-marketplaces are viewed as inter-organizational information systems that allow participant firms to integrate their information technologies in a Semantic eBusiness architecture that facilitates transparent information exchange (Choudhury, 1997).

SEMANTIC EBUSINESS VISION AND APPLICATIONS

Semantic eBusiness applies fundamental work done in semantic Web technologies, knowledge management, intelligent agent systems, and Web services to support the *transparent* flow of knowledge, *content*, and *know-how*, and enable semantically enriched collaborative eBusiness processes. Institutional trust among the collaborative partners engaged in Semantic eBusiness processes, as well as information assurance of all flows between integrated systems in the Semantic eBusiness network, is essential to the adoption of the vision. Semantic eBusiness requires a trusted and secure environment. Organizations develop descriptions of their business processes and business rules using semantic knowledge representation languages, such as OWL, in a format that allows for reasoning by intelligent software agents. Business processes consist of workflow descriptions that describe individual tasks at an atomic transactional level. At this transactional level, the individual services offered by organizations can be described using semantic languages. In

Figure 3. Semantic eBusiness utilizes Semantic Web technologies and existing information technology infrastructure for transparent information and knowledge flows in a secure and trusted environment



addition, product ontologies and meta-ontologies describe the relationships between the various resources utilized, required, or created by an organization in the Semantic eBusiness network. The Semantic eBusiness framework (Figure 3) utilizes (existing) information technology infrastructure, including Web services architecture to provide the transport infrastructure for messages containing semantic content.

The application of Semantic Web technologies to enable Semantic eBusiness provides the organizations the means to design collaborative and integrative, inter- and intra-organizational business processes and systems founded upon the seamless exchange of knowledge. Semantic eBusiness architectures can enable transparent information and knowledge exchange, and intelligent decision support to enhance online eBusiness processes. It can also help organizations fill the chasm that exists in the adaptation of emerging technologies to enable and enhance business processes through the use of distributed heterogeneous knowledge resources. The concept of Semantic eBusiness is potentially applicable to industries with an online presence. Candidates for applications in business include supply chain management and e-marketplaces. In addition, multiple not-for-profit and government processes are also potential application areas, including the health care industry for improving the management of medical records and e-government applications for improving services offered online to citizens. The following scenarios present some areas where we believe Semantic eBusiness can enhance information and knowledge exchange and improve the efficacy of eBusiness processes.

Potential Semantic EBusiness Applications

Supply Chain Management

Supply chain management (SCM) is a common strategy employed by businesses to improve organizational processes to optimize the transfer of goods, information, and services between buyers and suppliers in the value chain (Poirier & Bauer, 2000). A fundamental ongoing endeavor of SCM is to foster *information transparency* (availability of information in an unambiguously interpretable format) that allows organizations to coordinate supply chain interactions efficiently in dynamic market conditions. A standard ontology for all trading partners is necessary for seamless transformation of information and knowledge essential for supply chain collaboration (Singh et al., forthcoming). Increasing complexity in supply chains make the timely sharing of accurate information among collaborating partners a critical element in the efficiency of workflows and eBusiness processes. Information and knowledge exchange facilitated through semantic Web technologies enable the creation of global information partnerships across the entire supply chain. Organizations embracing such paradigms can sustain their competitive advantages by having an effective and efficient e-supply chain and realize benefits such as reduced cycle times, lower product costs, reduced inventory, better quality decision making, and improved customer service.

E-Marketplaces

Infomediaries perform a critical role in bringing together buyers and suppliers in the e-marketplace and facilitating transactions between them. A detailed description of the value-added activities provided by infomediaries in e-marketplaces can be found in Grover and Teng (2001). The infomediary adds value through its role as an enterprise system hub responsible for the critical integration of the information flows across participant firms (Davenport & Brooks, 2004). Infomediaries become vital repositories of knowledge about buyers, suppliers, and the nature of exchanges among them including the past experiences of other buyers' reliability and trustworthiness of the supplier. They provide independent and observed post-transaction assessment of the commitments of the individual buyers and sellers to facilitate the development of coordination structures, leading to collaborative relationships in e-supply chains. The integration of intelligence and knowledge within and across e-marketplaces can enhance the coordination of activities among collaborating firms across e-marketplaces (Singh et al., 2003). Collaborations create information partnerships between organizations to enable the delivery of products and services to the customer in an efficient manner. Such information partnerships are founded upon the transparent exchange of information and knowledge between collaborating organizations in a dynamic manner across participants in the value chain.

Healthcare

Healthcare delivery is very complex and knowledge dependent. Information systems employed for healthcare store information in very disparate and heterogeneous clinical information system data repositories. Pollard (2004) states that knowledge management activities in healthcare center on acquiring and storage of information, and lacks the ability to share and transfer knowledge across systems and organizations to support individual user productivity. In addition the data acquired and stored in islands clinical information systems are in multiple formats. Common vocabulary to represent data and information is needed for efficient knowledge management (Desouza, 2002). The focus has been on building independent applications to make these systems talk to each other. The need is for models to integrate the data and knowledge in these disparate systems for effective knowledge sharing and use (Sittig et al., 2002). To serve the needs, relevant patient-centered knowledge must be accessible to the person supplying care in a timely manner in the workflow. Interoperability standards of emerging Semantic Web technologies can enable health information integration, providing the transparency for healthcare-related processes involving all entities within and between hospitals, as well as stakeholders such as pharmacies, insurance providers, healthcare providers, and clinical laboratories. Further research on using Semantic Web technologies is needed to deliver knowledge services proactively for improved decision making. Such innovations can lead to enhanced caregiver effectiveness, work satisfaction, patient satisfaction, and overall care quality in healthcare (Eysenbach, 2003).

E-Government

E-government refers to the use of Internet technologies for the delivery of government services to citizens and businesses (www.Webster-dictionary.org/definition/EGovernment). The aim of E-government is to streamline processes and improve interactions with business and industry, empower citizens with the right information, and improve the efficiency of government management. Given that e-government services extend across different organizational boundaries and infrastructures, there is a critical need to manage the knowledge and information resources stored in these disparate systems (Teswanich, Anutariya, & Wuwongse, 2002). Emerging Semantic Web technologies have the ability to enable transparent information and knowledge exchange to enhance e-government processes. Klischewski and Jeenicke (2004) examine the use of ontology-driven e-government applications based on Semantic Web technologies to support knowledge management related to e-government services. Further research to investigate requirements, design and develop systems, and examine success factors for systems development employing Semantic Web technologies for effective knowledge management within e-government services is needed.

ORGANIZATIONS AND RESEARCH GROUPS FOSTERING A SEMANTIC EBUSINESS VISION

As research in the foundation technologies for the Semantic Web develops, the application of these technologies to enable Semantic eBusiness is of increasing importance to the professional and academic communities. In this section we would like to inform the readers of several organizations that are involved in furthering research related to Semantic eBusiness.

Association for Information Systems (AIS) (www.aisnet.org)

A professional organization, the Association for Information Systems (AIS) was founded in 1994 to serve as the premier global organization for academics specializing in information systems. This organization has formed several special interest groups (SIGs) to provide substantial benefits to IS students, academics, and practitioners by helping members exchange ideas and keep up to date on common research interests. The following SIGs contribute significantly to advacing Semantic eBusiness research:

- *Special Interest Group on Semantic Web and Information Systems—SIG-SEMIS (www.sigsemis.org):* SIG-SEMIS' goal is to cultivate the Semantic Web vision in IS. The main areas of emphasis in this SIG are: Semantic Web, Knowledge Management, Information Systems, E-Learning, Business Intelligence, Organizational Learning, and Emerging Technologies. The SIG aims to "create knowledge capable of supporting high-quality knowledge and learning experience concerning the integration" of the above main areas. This integration will provide the participants of the SIG an opportunity to create and diffuse knowledge concerning the issues of Semantic Web in the IS research community.
- *Special Interest Group on Agent-Based Information Systems—SIG-ABIS (www.agentbasedis.org):* SIG-ABIS aims to advance knowledge "in the use of agent-based information systems, which includes complex adaptive systems and simulation experiments, to improve organizational performance. SIG-ABIS promises to fill an existing gap in the field, and therefore is more focused on the strategic and business issues with agent technology and less on the artifact itself, such as computational algorithms, which are well investigated by computer science related research groups."
- *Special Interest Group on Ontology Driven Information System—SIG-ODIS (aps.cabit.wpcarey.asu.edu/sigodis/):* The objective of SIG-ODIS is to provide "a unifying international forum for the exchange of ideas about the field of ontology as it relates to design, evaluation, implementation, and study of ontology driven information systems." In helping develop awareness and foster research about the role and impact of computational ontologies on the design, development, and management of business information systems, SIG-ODIS also strives to build bridges between the IS discipline and other related disciplines, such as computer science, information science, philosophy, linguistics, and so forth, that pursue research in the broad area of computational ontologies.
- *Special Interest Group on Process Automation and Management—SIG-PAM (www.sigpam.org):* SIG-PAM's objective is to address the "need of IS researchers and practitioners for information and knowledge sharing in the areas of process design, automation, and management in both organizational and inter-organizational contexts." The SIG

collaborates with other not-for-profit organizations that have related focus on process theories and applications, such as the Workflow Management Coalition (WfMC), the Workflow and Reengineering International Association (WARIA), and the Computer Supported Collaborative Work (CSCW) Conference.

Hewlett-Packard (HP) Labs Semantic Web Research (www.hpl.hp.com/semWeb/)

The HP Labs Semantic Web research group recognizes that Semantic Web technologies can enable new and more flexible approaches to data integration, Web services, and knowledge discovery. The HP Labs' investment in the Semantic Web consists of the development of Semantic Web tools (such as Jena, a Java framework for writing Semantic Web applications) and associated technology, complemented by basic research and application-driven research. HP is also part of several collaborative ventures, including involvement in W3C initiatives (RDF and Web ontologies working groups) and European projects (Semantic Web Advanced Development Europe—SWAD-E and Semantic Web-enabled Web Services—SWWS).

World Wide Web Consortium's Semantic Web Initiative (www.w3.org/2001/sw/)

The main goal of the W3C Semantic Web initiative is to create a universal medium for the exchange of data. "It is envisaged to smoothly interconnect personal information management, enterprise application integration, and the global sharing of commercial, scientific, and cultural data. The W3C Semantic Web activity has been established to serve a leadership role in both the design of specifications and the open, collaborative development of enabling technology."

In addition to these organizations, the formation of this new journal, *International Journal on Semantic Web and Information Systems*, provides an opportunity for the publication and exchange of research discussions of the Semantic Web in the context of information systems.

SUMMARY AND RESEARCH DIRECTIONS

The realization of representing knowledge-rich processes is possible through the broad developments in the Semantic Web initiative of the World Wide Web Consortium. We defined Semantic eBusiness as "*an approach to managing knowledge for coordination of eBusiness processes through the systematic application of Semantic Web technologies.*" Advances in Semantic Web technologies—including ontologies, knowledge representation, multi-agent systems, and the Web services architecture—provide a strong theoretical foundation to develop system architecture that enables semantically enriched collaborative eBusiness process. Semantic eBusiness architecture enables transparent information and knowledge exchange and intelligent decision support to enhance online eBusiness processes.

Developments in the availability of content and business logic on-demand, through technologies such as Web services, offer the potential to allow organizations to create content-based and logic-driven information value chains, enabling the needed information transparencies for Semantic eBusiness processes. Research is needed to understand how conceptualizations that comprise business processes can be captured, represented, shared, and processed by both human and intelligent agent-based information systems to create transparency in eBusiness processes. Further work on these dimensions is critical to the design of knowledge-based and intelligence-driven eBusiness processes in the digital economy.

Research is also needed in the development of business models that can take advantage of emergent technologies to support collaborative, knowledge-rich processes characteristic of Semantic eBusiness. Equally important is the adaptation and assimilation of emergent technologies to enable Semantic eBusiness processes, and the contribution to organizations' value propositions. Topics of research directions include the development of innovative, knowledge-rich business models that enhance collaborations in eBusiness processes, and innovative technical models that enable the vision of Semantic eBusiness.

One of our current research initiatives involves developing models for the representation of knowledge, using ontologies and intelligent agents for semantic processing of cross-enterprise business processes over heterogeneous systems. For the Semantic Web to be a *vibrant and humane environment* for sharing knowledge and collaborating on a wide range of intellectual enterprises, the W3C must include in its Semantic Web initiatives research agenda the creation of policy-aware infrastructure, along with a trust language for the Semantic Web that can represent complex and evolving relationships.

REFERENCES

Baader, F., Calvanese, D., McGuinness, D., Nardi, D., & Patel-Schneider, P.F. (2002). *The description logic handbook: Theory, implementation, and applications*. Cambridge University Press.

Badii, A., & Sharif, A. (2003). Information management and knowledge integration for enterprise innovation. *Logistics Information Management, 16*(2), 145-155.

Beam, H.H. (1998). The infinite resource: Creating and leading the knowledge enterprise. *The Academy of Management Executive, 12*(3).

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

Chiu, C. (2000). Re-engineering information systems with XML. *Information Systems Management*, 17(4).

Choudhury, V. (1997). Strategic choices in the development of interorganizational information systems. *Information Systems Research, 8*(1), 1-24.

Dai, Q., & Kauffman, R.J. (2002). Business models for Internet-based B2B electronic markets. *International Journal of Electronic Commerce, 6*(4), 41-72.

Davenport, T.H., Brooks, J.D., & Cantrell, S. (2001, January). B2B e-market survey: Summary of findings. White Paper, Accenture Institute of Strategic Change.

Davenport, T.H., & Brooks, J.D. (2004). Enterprise systems and the supply chain. *Journal of Enterprise Information Management, 17*(1), 8-19.

Desouza, K.C. (2002). Knowledge management in hospitals: A process oriented view and staged look at managerial issues. *International Journal of Healthcare Technology & Management, 4*(6), 478-497.

Eysenbach, G. (2003). The Semantic Web and healthcare consumers: A new challenge and opportunity on the horizon?

International Journal of Healthcare Technology & Management, 5(3/4/5), 194-212.

Fensel, D. (2000). *IEEE Intelligent Systems*, (November/December), 67.

Grover, V., & Davenport, T.H. (2001). General perspectives on knowledge management: Fostering a research agenda. *Journal of Management Information Systems, 18*(1).

Grover, V., & Teng, J.T.C. (2001, April). E-commerce and the information market. *Communications of the ACM, 44*(4), 79-86.

Gupta, B., Iyer, L., & Aronson, J.E. (2000). Knowledge management: A taxonomy, practices and challenges. *Industrial Management and Data Systems, 100*(1), 17-21.

Hendler, J. (2001). Agents and the Semantic Web. *IEEE Intelligent Systems*, (March/April), 30-37.

Holsapple, C., & Singh, M. (2000). Toward a unified view of electronic commerce, electronic business, and collaborative commerce: A knowledge management approach. *Knowledge and Process Management*, 7(3), 159.

Jennings, N.R., & Wooldridge, M. (1998). *Agent technology: Foundations, applications, and markets*. London: Springer-Verlag.

Kishore, R., Sharman, R., & Ramesh, R. (2004). Computational ontologies and information systems: I. foundations. *Communications of the Association for Information Systems, 14,* 158-183.

Klein, M., Fensel, D., van Harmelen, F., & Horrocks, I. (2001). The relation between ontologies and XML schemas. *Electronic Transactions on Artificial Intelligence (ETAI), Linköping Electronic Articles in Computer and Information Science, 6*(4).

Klischewski, R., & Jeenicke, M. (2004). Semantic Web technologies for information management within e-government services. *Proceedings of the 37th Hawaii International Conference on System Sciences.*

Malone, T.W., Yates, J., & Benjamin, R.I. (1987). Electronic markets and electronic hierarchies. *Communications of the ACM, 30*(6), 484-497.

McIlraith, S., Son, T.C., & Zeng, H. (2001). Semantic Web services. *IEEE Intelligent Systems*, (March/April), 46-53.

Muller, H.J. (1997). Towards agent systems engineering. *Data and Knowledge Engineering, 23,* 217-245.

O'Leary, D. (1998). Knowledge management systems: Converting and connecting. *IEEE Intelligent Systems and Their Applications, 13*(1).

Papazoglou, M.P. (2001). Agent oriented technology in support of eBusiness: Enabling the development of intelligent business agents for adaptive, reusable software. *Communications of the ACM, 44*(4), 71-77.

Poirier, C.C., & Bauer, M. (2000). *E-supply chain: Using the Internet to revolutionize your business*. Berrett-Koehler.

Pollard, D. (2004). Knowledge integration leading to personal knowledge management: Enabling better life science and medical product management and health delivery. *Knowledge Management Blog—The Ferryman,* (June 15). Retrieved September 2004 from barryhardy.blogs.com/theferryman/2004/06/knowledge_integ.html.

Rabin, S. (2003). The real-time enterprise, the real-time supply chain. *Information Systems Management, 20*(2), 58-62.

Singh, R., Salam, A.F., & Iyer, L. Intelligent infomediary-based e-marketplaces:

Agents in e-supply chains. *Communications of the ACM*, (forthcoming).

Singh, R., Iyer, L.S., & Salam A.F. (2003). Web service for knowledge management in e-marketplaces. *eService Journal*, *3*(1).

Shoham, Y. (1993). Agent oriented programming. *Journal of Artificial Intelligence*, *60*(1), 51-92.

Sittig, D.F., Hazlehurst, B.L., Palen, T., Hsu, J., Jimison, H., & Hornbrook, M.C. (2002). A clinical information system research agenda for Kaiser Permanente. *The Permanente Journal*, *6*(3), 41-44.

Staab, S., Studer, R., Schnurr, H.P., & Sure, Y. (2001). Knowledge processes and ontologies. *IEEE Intelligent Systems*, (February).

Teswanich, W., Anutariya, C., & Wuwongse, V. (2002). Unified representation for e-government knowledge management. *Proceedings of the 3rd International Workshop on Knowledge Management in E-Government* (pp. 199-209), University of Linz and University of Roskilde.

Wiig, K.M. (1993). *Knowledge management foundations*. Schema Press.

Rahul Singh is a faculty member in the Information Systems and Operations Management Department at The University of North Carolina at Greensboro. He obtained his PhD in business administration from Virginia Commonwealth University. His research interests are in the area of the design of systems that apply intelligent technologies to business decision systems. Specific interests include intelligent agents, knowledge management systems, data mining and machine learning systems. His research work has been published, or accepted for publication in CACM, eService Journal, Information Resources Management Journal, Journal of Decision Systems, International Journal of Production Engineering Socio-Economic Planning Sciences *and* The Encyclopedia of Information Systems.

Lakshmi Iyer is a faculty member in the Information Systems and Operations Management Department at The University of North Carolina at Greensboro. She obtained her PhD from the University of Georgia, Athens. Her research interests are in the area of electronic commerce, global issues in IS, intelligent agents, decision support systems, knowledge management, and cluster analysis. Her research work has been published or accepted for publication in CACM, eService Journal, Annals of OR, DSS, Journal of Global Information Technology Management, Journal of Scientific & Industrial Research, Encyclopedia of ORMS, Journal of Information Technology Management, Journal of Data Warehousing, *and* Industrial Management and Data Systems.

A.F. Salam is a faculty member in the Information Systems and Operations Management Department at The University of North Carolina at Greensboro. He earned both his MBA and PhD degrees from the State University of New York at Buffalo. His research interests include e-business and emerging technologies, intelligent systems in e-business, and information technology management. His research has been published or accepted for publication in CACM, Information & Management, eService Journal, CAIS, Information Systems Journal, *and* Journal of Information Technology Cases and Applications. *He has served as co-guest editor of a Special Issue of the* JECR *and a Special Section on Internet and marketing in the CACM.*

A Tool for Working with Web Ontologies

Aditya Kalyanpur, University of Maryland, USA
Bijan Parsia, University of Maryland, USA
James Hendler, University of Maryland, USA

ABSTRACT

The task of building an open and scalable ontology browsing and editing tool based on OWL, the first standardized Web-oriented ontology language, requires the rethinking of critical user interface and ontological engineering issues. In this article, we describe Swoop, a browser and editor specifically tailored to OWL ontologies. Taking a "Web view" of things has proven quite instructive, and we discuss some insights into Web ontologies that we gained through our experience with Swoop, including issues related to the display, navigation, editing, and collaborative annotation of OWL ontological data.

Keywords: computer systems; Semantic Web; Web technologies

INTRODUCTION

The Web ontology language, OWL (Dean & Schreiber, 2004), was approved in February 2004 as a World Wide Web Consortium (W3C) Recommendation for the publication of ontologies on the World Wide Web—creating a standard language for the publication and exchange of ontological models on the Web. OWL reflects almost 10 years of research, experimentation, and small-scale deployment of Web ontologies; a number of certain features in its design were made explicitly to help realize the ideal of Web-based ontologies, that is, of integrating knowledge representation with the open, global, and distributed hypermedia system of the Web, compatible with the principles of Web architecture design. In this article we discuss some insights into supporting the use of Web ontologies that we have gained in building Swoop[1], an ontology browser and editor, designed specifically for use with OWL and directly supporting the use of Web-based "cultural metaphors"—that is, based on the way people are used to interacting with documents and data in current Web applications.

A WEB (ONTOLOGY) BROWSER—OWL

OWL is a standard for representing knowledge on the Web, with a focus on both making these documents compatible with Web standards and on being useful for the modeling of knowledge using past research on ontologies and reasoning. OWL comes in three increasingly expressive sublanguages—OWL Lite, DL, and Full. The Lite and DL species of OWL are based on description logics, that is, decidable, class- and property- oriented subsets of first-order logic. OWL Full follows RDF schema in having a higher-order syntax (although first-order semantics)—OWL Full does not enforce a strict separation of classes, properties, individuals, datatypes, or data values. Any entity could be, for example, both a class and an individual. This design was motivated by the Web architecture dictum that "everything is a resource," thus an individual, and from the general modeling consideration that the choice between whether to represent some aspect of a domain as a class or an individual is not always clear. In a world where people are trying to reuse vocabulary and map between concepts, it seems quite natural to be able to express the dual view of certain domain objects as either classes or individuals, and sometimes both.

One characteristic of "Webized" languages, especially Semantic Web languages, is the systematic prevalence of Universal Resource Indicators (URIs)[2] as names for most entities. In OWL, names for classes, properties, individuals, datatypes, and so forth are URIs. URIs have a number of useful properties, including:

1. For a number of URI schemes, notably http URIs, there is a well-developed set of mechanisms for avoiding name collisions, most notably the domain name system (DNS).
2. These mechanisms, especially the DNS, interact with various Internet protocols, notably HTTP, to make it very easy to publish and retrieve information associated with a URI.
3. URIs have various degrees of opacity. For example, HTTP imposes relatively few constraints on the semantics of the scheme specific part 1. A URI is a generalization of the more common URL, roughly composed of a naming scheme or protocol indicator (http, ftp, mailto, etc.), a unique indicator (a domain name space name for http, a mail address for mailto), and a "fragment id," which is a hash mark followed by a set of characters—thus, for example, an OWL class called "person" from an ontology on a university server might be named by the URI: http:/www.thisuniversity.edu/OntologyLib/csontology#person. The hierarchical structure seen in most http URIs can map directly into a file system (which is a very useful default behavior), but it can also map into queries on a relational database, the object structure of a long-running process, or any other Web resource.
4. URIs can work well for end users, who have developed a lot of expertise with using URIs when browsing or authoring. Web browsers are the ubiquitous way that people use URIs, and even in authoring tools, the primary mental model people have of URIs is derived from their use in browsers. In designing Swoop, we took the Web browser as our user interface (UI) paradigm, believing that URIs are central to the understanding and construction of Web ontologies. We contrast this to other ontology editors such as Protégé (Noy et

al., 2001), OilED (Bechhofer, Horrocks, Goble, & Stevens, 2001), and OntoEdit (Sure et al., 2002), which either are or were influenced by traditional KR development tools and applications, and do not reflect this "Webiness" in their UI design. In particular, they do not fully support the use of hypertext to drive the exploration and editing of ontologies.

Hypertextual Navigation

In a Web browser, there are two primary modalities for URIs: manifest and hidden. The address bar is the central mechanism for manifest URIs. URIs must be typed into the address bar and are always visible there. Browser features such as history drop-downs and the use of name completion mean that users need not remember or enter entire URIs, while the address bar requires and abets interaction with raw URIs. The most prominent hidden use of URIs is the hyperlink wherein the URI address is the target of a clickable (in most browsers) region of text (or an image).

There are tight links between hidden and manifest URIs. The URIs hidden "in" hyperlinks appear in the address bar after one has followed a hyperlink or may be revealed by mousing over a hot region, retrieved by pop-up menu commands (i.e., copy hyperlink) or by viewing the actual HTML source.[3]

The ecology of Web pages depends on the ease of access to URIs, both hidden (there is no hypertext without hyperlinks!) and manifest. Much Web browsing starts with URIs discovered in non-Web media, from e-mail to billboards and buses. Writing Web pages requires, even in WYSIWIG HTML editors, familiarity with URIs and the ability to secure the right ones.

Bookmarks are another example of hidden URIs, at least in their most common form. Browsers typically have many ways to review bookmarked URIs. As the natural habitat of Web ontologies is the Web, Swoop allows the interactions with these, using the UI metaphors prevalent on the Web. For loading ontologies, Swoop presents the familiar address bar, and the URI for such an ontology can be secured by whatever means—e-mail, Google, or perhaps one day, a billboard or bus.

Views

It is worth considering the level of detail that needs to be displayed while rendering Web ontological information. While an OWL entity is represented by its URI, it is characterized in a specific context by the axioms dealing with the entity in that context (the document or ontology). Moreover, on the Semantic Web, we expect OWL entities to be characterized by axioms in remote documents. That is, we expect OWL documents and OWL ontologies to use Web links. When rendering the related axioms or definition of an OWL entity, we have taken care that the appropriate information is directly presented in an intelligible manner, and that all the known information is naturally accessible. We consider various levels of detail at which information related to an entity can be displayed:

1. its definition and related axioms (within a single ontology);
2. axioms relating it to imported entities (from an external ontology);
3. inferred information (not explicitly stated in the ontology, but which is inferred from its definition using an OWL reasoner or otherwise);

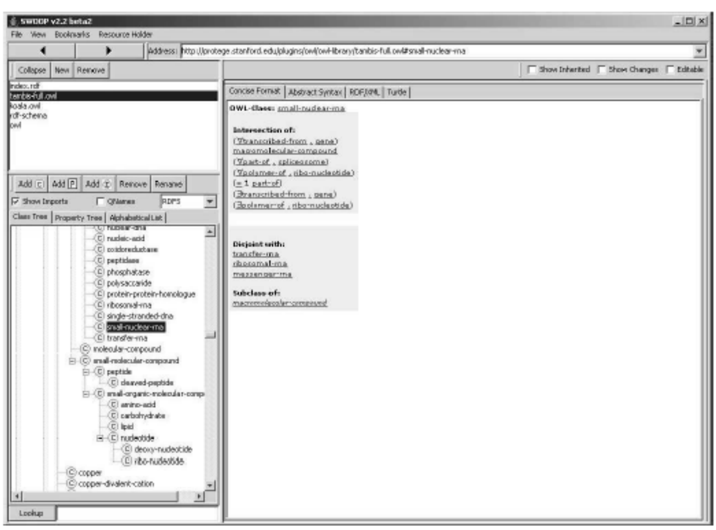
4. semantic consistency information (whether the concept is satisfiable or not, again using an OWL reasoner);
5. provenance information (source location of a particular axiom, its author, creation date, etc.);
6. entity annotations (human-readable comments made on the entity);
7. changes (a log of changes made to the entity definition); and
8. usage of an entity (references in other Semantic Web documents).

Thus, there is an array of entity-related information that could be displayed as a single Web document that pertains to any OWL entity. Currently, Swoop supports all but the provenance information and usage views (#5,#8) listed above, making clear distinctions between the various view types displayed. For instance, inferred axioms are italicized, inconsistent classes have red icons, and changes pending are shown in green (see Figures 1 and 2). The other two open some complex research issues that are being explored by our research group and others.

Orthogonal to the above levels of detail is the syntax (format) used to render the ontology. Currently on the Semantic Web, a wide range of OWL presentation syntaxes exist—the raw RDF/XML serialization, the more triple-oriented Turtle language (Beckett, 2004), and the OWL Abstract Syntax (Patel-Schneider, Hayes, & Horrocks, 2004), to name a few. It is important to support as many as possible of these different syntaxes while designing an open, Semantic Web ontology engineering environment. One reason for this is that people tend to have strong biases toward different notations and simply prefer to work in one or another. A second is that some other tool might only consume one particular syntax (with the RDF/XML syntax being the most typical), but that syntax might not be an easy or natural one for a particular user.

A third is that it is important to support the "view source" effect, allowing cut-

Figure 1. Web-browser UI reflected in Swoop



and-paste reuse into different tools, including text editors, markup tools, or other semantic Web tools. We have observed that the easy, direct data transformation between any two formats feels very powerful to the user, especially if they need to use more than one format for a particular task. The challenge here is that the formats should be treated as similarly as possible—that is, any task that can be done in one format should be allowed in any other so that people can stick with the syntax they prefer for both browsing and editing.

Swoop uses a plug-in-based mechanism for renderers. The architecture supports two types of renderers, a coarse-grained type for viewing the ontology as a whole (i.e. class/property tree, graphs, lists etc.) and a fine-grained type for viewing the description of a single ontological entity (i.e., an OWL class, property, or individual). Other levels of granularity can be achieved by filtering out information from the above main types. All of these formats use URIs (and various URI abbreviations) throughout. Swoop renders those URIs as hyperlinks, allowing for essentially the same hypertext-based navigation, no matter what format is being used.

Also, the layout of the ontology and entity renderers resembles a familiar frame-based Web site viewed through a Web browser. As shown in Figure 1, a navigation sidebar on the left contains the multiple ontology list and class/property hierarchies for each ontology, and the center pane contains the various ontology/entity renderers for displaying the core content.

Currently, Swoop bundles in six renderers; two Ontology Renderers—Information and Species Validation; and four Entity Renderers—Concise Format, OWL Abstract Syntax, Turtle, and RDF/XML. Besides these, there exists a class/property hierarchy renderer for each ontology, along with an alphabetical list of entities present in the ontology. Here we discuss only the Concise Format renderer, since its motivation, design, and subsequent functionality is unique to Swoop.

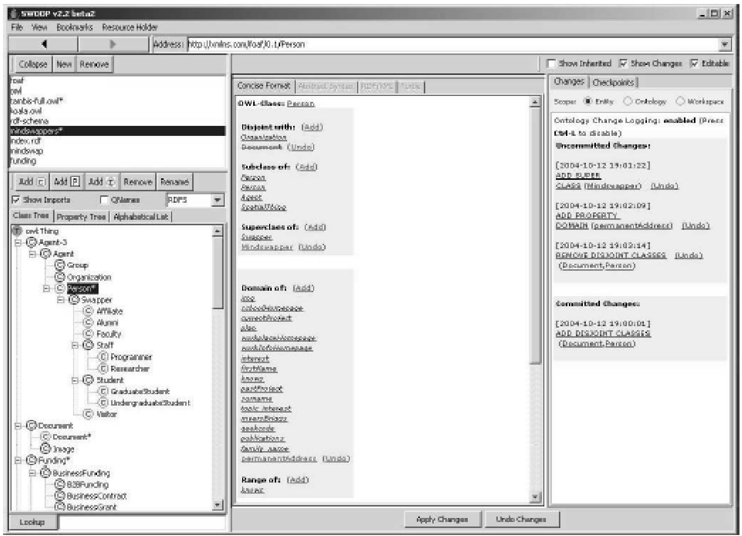
The Concise Format entity renderer is a non-standard presentation syntax in Swoop (see Figure 2). The idea here is to generate a "Web document" that displays all information related to a particular OWL entity concisely in a single pane. Items are divided into logical groups and rendered in a linear fashion. So taking an OWL Class for example, its OWL enumerations if any, i.e., intersectionOf, unionOf, and oneOf—are listed in one group, while the OWL properties related to it (through domain or range) are listed in another group. Standard description logic (DL) operators are used whenever they occur in class expressions to make the representation more concise. Here again, all entity references are made hyperlinks using their URIs as the identifiers. Thus, clicking on an OWL entity link in a particular document causes the view to shift directly to the linked entity's document. This is in keeping with the look and feel of traditional Web-like viewing and navigation of documents.

Editing

Editing OWL entities in a multiple ontology engineering environment can be challenging. Some of the issues that arise include:

1. The scope of a change (should editing be restricted to the local ontology alone or can the imported ontology be (directly or indirectly) altered as well?).
2. The types of changes allowed (i.e., atomic vs. composite change strategies

Figure 2. Editing OWL entities in Swoop (concise format view)



as discussed in Bechhofer, Lord, & Volz, 2003).

3. The level at which changes are made (in the abstract representations or directly in the source code).
4. How to display the effects of changes before they are committed (direct vs. inferred effects on related entity definitions).
5. The degree of rollback possible (for how long changes can be "undo"ne).

Issues 1 and 2 are dealt with in detail in subsequent sections; we consider the remaining here. All ontology editing in Swoop is done in line with the renderer pane. This way, context is maintained while editing a particular entity. Also, effects of change on any of the related entities can be easily observed (a single click away) by switching back and forth between the current entity and the related ones by following hyperlinks and use of the history buttons.

Swoop allows ontology editing either at the concise representation level or directly in the code (currently only RDF/XML code editing is supported). There are some fundamental differences between editing in these two modes. For instance, in the concise format, all information related to an entity is displayed in a single pane. As noted earlier, this information is further subdivided into various logical groups, each of which can be edited separately. The changes enacted in this mode are identifiable, and hence can be recorded and undone. Also, the axioms related to a particular entity may not be located in a single region of the code.

Thus, directly editing all references of a single entity in RDF/XML (for example) might be cumbersome. Moreover, given the arbitrary manner in which the RDF/XML code can be edited, it is not easy to capture and record changes easily. On the other hand, direct code editing can be faster and certain changes can be made easily, for example renaming all references of a single class in the entire ontology can be done using the find/replace functionality of an editor. Given the need for both types of changes, Swoop supports both forms of editing.

Another important consideration in Swoop is the manner in which changes are effected. Swoop provides two options for this: either a change can be applied immediately (upon enacting it), or a set of proposed changes can be set aside and collectively committed at a later stage. While the former approach gives immediate results, the latter has numerous advantages. It speeds up alteration of large ontologies, where enforcing multiple changes one at a time would take considerably more time. Additionally, it provides a composite change record that is especially useful for ontology versioning. Finally, it gives a basis for implementing Issue 4 noted above—displaying change effects before they are committed.[4]

Searching, Comparing, Reusing

In a distributed Web ontology setting, numerous engineering tasks—such as comparing entities with a view to understanding semantic differences, mapping entities to ensure semantic interoperability, or simply reusing entities to prevent reinventing the wheel—requires a search/browse process involving disparately located entities. The ontology engineering client can play a large role in making this process efficient.

We take inspiration from the hyperlink-based search and cross-referencing utility present in a programming IDE such as Eclipse (www.eclipse.org). All named entities in the code are identified, and one can easily obtain (and jump directly to) useful related information such as all its references.

During an extended search and browsing routine, the user of Swoop may come across numerous interesting results (OWL entities) that may need to be set aside and revisited. In Swoop we have a provision to store and compare OWL entities via a resource holder panel. Items can be added to this placeholder at any time and they remain static there until the user decides to remove or replace them at a later stage. Upon adding an entity, a time-stamped snapshot of it is saved (with hyperlinks and all), thus providing a reference point for future engineering tasks. These include, but are not limited to, tracking changes made to a particular entity; storing entities for reuse in another ontology; comparing differences in definitions of a set of entities; and determining semantic mappings between a specific pair of entities. We are working to further improve the resource holder by adding automatic dynamic tracking for selected entities, color coding diffs between different entity definitions, and providing support for the editing of mapping terms, such as "owl:equivalentTo" between terms in different resource panes.

Why Not a Web Site?

In principle, the entire Swoop interface and functionality could have been provided as a Web site, or on top of a more full fledged Web browser such as Mozilla. Indeed, a very common first question we get when we show people Swoop is, "Why not do it as a Web site?" There are several examples of current Web site-based ontology tools such as Ontosaurus (Farquhar, Fickas, & Rice, 1996) and WebODE (Arpírez, Corcho, Fernández-López, &

Gómez-Pérez, 2001), and new ones are being developed such as pOWL (powl.sourceforge.net). However, we have found that using a standard Web-based server-client architecture for ontology engineering suffers from being slow (especially for large ontologies, and depending on network traffic), and cumbersome for maintaining consistency while editing (e.g., trapping input errors, changing/deleting objects but reloading from browser cache, etc.). In addition, such tools can be difficult to extend to new functionalities via plug-in architectures (such as the one used in Swoop). Finally, most Web site-based ontology editors use distinct HTML pages (perhaps dynamically generated) not just for each entity, but for each view of those entities. This indirection puts an uncomfortable distance between the user and the ontology itself. For these reasons, Swoop is developed as a separate Java application that attempts to provide the look and feel of a browser-based application, but with its specialized architecture designed to optimize OWL browsing and to be extensible via a plug-in architecture.

MULTIPLE ONTOLOGIES: FROM MANY, MANY

OWL's Web-based features open up the Web ontology engineering environment to multiple ontologies which can, and often do, refer to each other in a number of ways or share terms. This has ramifications for a number of aspects of ontology-editing that have been largely ignored in many earlier AI-based ontology tools. Swoop assumes the use of multiple ontologies and supports this use in a number of ways.

Display and Navigation

Being an open multiple ontology engineering environment, Swoop has a no-holds-barred approach for pulling different Web ontologies into its model. Depending on the nature and context of the task being performed, ontologies are brought into Swoop seamlessly, that is, no additional user intervention is required and the UI treats all ontologies similarly. For example, consider the scenario in which the user is browsing a particular OWL class, say A, in a Web ontology that has an OWL class B related to it by an axiom (say rdfs:subClassOf). Also, B is not defined in the same ontology; instead it has a separate physical Web location and has a number of URIs that share no common prefix with the rest of A's URIs. Clicking on the class B hyperlink causes Swoop to directly load the external ontology referenced and select class B in it. Thus, no distinction is made in terms of UI between navigation across entities in a single ontology or those present in multiple ontologies. Also, the back and next buttons can be used to jump between OWL entities in different ontologies on a single click, ensuring the familiar Web browser experience.

Besides the aforementioned scenario, there are various other situations that can drive Swoop to load more than one ontology. For example, multiple ontologies can be loaded at any point by entering their Web location URLs in the address bar. Alternately, the bookmarks feature can be used to store, categorize, and reload ontologies directly. Finally, if a particular OWL ontology has imported ontologies (defined using owl:imports), loading it causes all its imports under transitive closure to be loaded into Swoop directly.

Living with Imports

The use of owl:imports reveals numerous open issues in Web ontology engineering. Two interrelated issues are considered here—UI issues in distinguishing between the definitions and semantics of imported OWL axioms, and editing support for axioms defined in the importing ontology. Consider the case when an OWL class A is related by an axiom (say owl:disjointWith) to another class B. Suppose A and B have been defined in different ontologies, OA and OB respectively, and moreover, OA imports OB. (In OWL, an entity reference is defined in an ontology using rdf:ID, and it can be further referenced in the same or any other ontology using rdf:about—thus allowing cross-referencing of terms between ontologies.)

Now, the owl:disjointWith axiom can be defined in either ontology OA or ontology OB (or both!). Either way, the semantics of owl:imports, and the fact that OA imports OB, ensures the axiom is present in ontology OA. Yet, it is important to display to the user the exact source of axiom definition. This is especially important when the user wishes to delete this axiom. Obviously, the axiom cannot be deleted in the importing ontology; instead, the user must delete the axiom at the location at which it is originally defined (i.e., imported ontology). Hence, in our case, if the axiom is defined in OB, even though it is displayed in OA as well, it can only be deleted in OB. Swoop needs to make these distinctions since it does viewing and editing axioms inline. Currently, this is accomplished by italicizing all imported axioms (but if an axiom is also local, that overrides).

Also, given that we use the URI of a class as its identifier in a hyperlink, there is an ambiguity of a URI when the class is referenced in different ontologies in terms of what class definition needs to be displayed when the hyperlink is clicked. So consider the above case involving classes A and B, but here the owl:disjointWith axiom is present in OA and not OB. Now, if the user is viewing the axiomatic definitions of class A and clicks on the hyperlink corresponding to class B, there are two possibilities:

1. Swoop jumps to the class definition B in ontology OB (imported ontology), and here the disjointWith axiom is neither defined nor displayed.
2. Swoop jumps to the class definition B in ontology OA itself (importing ontology), and here all imported axioms from OB are displayed along with the owl:disjointWith axiom.

Note how the two views hold different semantics and rightly so, reiterating the point that the meaning of an OWL entity is defined in a specific context (ontology). To solve the URI ambiguity problem, Swoop provides labels next to the hyperlinks as an indicator to the jump location.

Beyond Imports?

Current research makes it clear that owl:imports is not the last word in combining (or referencing) Web-based ontologies; in fact, problems with the use of this mechanism were pointed out as part of the OWL documents as an important area for future standardization. Recent work, for example, has been looking at using concepts from foreign ontologies without resorting to the all or nothing approach that owl:imports demands (Borgida & Serafini, 2003; Kutz, Lutz, Wolter, & Zakharyaschev, 2003; Cuenca-Grau, Parsia, & Sirin, 2004). We have discovered in Swoop that the problem of "where to go" when following a URI in an OWL document is not unique to owl:imports and arises in many different contexts during the editing of multiple, linked ontologies. Different collections of axioms seem to define (or characterize) different concepts. The RDF(S)/OWL Full view of concepts (or properties) as entities which may have varying definitions (and extensions) associated with them in different contexts—even in situations where there is no disagreement, but mere normal use—is helpful, especially when coupled with some explicit identification mechanism for various definitions. In our work we have observed that the OWL Full view is more helpful at the Web infrastructure level than, as far as we can currently see, at the logic level. Classes as instances can be a USEFUL Ontological modeling tool (Noy, 2004), but it might be that in the Semantic Web context, much of their value lies outside their use in characterizing a domain. For this reason, Swoop supports OWL Full, and the concise view displays both the class and instance properties of an entity in the same panel. However, these are separated visually to allow the user to more easily identify cases where this occurs.

ANNOTATIONS

When browsing or building ontologies that live on the Web, it is almost as important to have information about the ontologies as it is to have the ontologies themselves. OWL allows for the associating of variously structured information with its core entities (e.g., classes and properties).

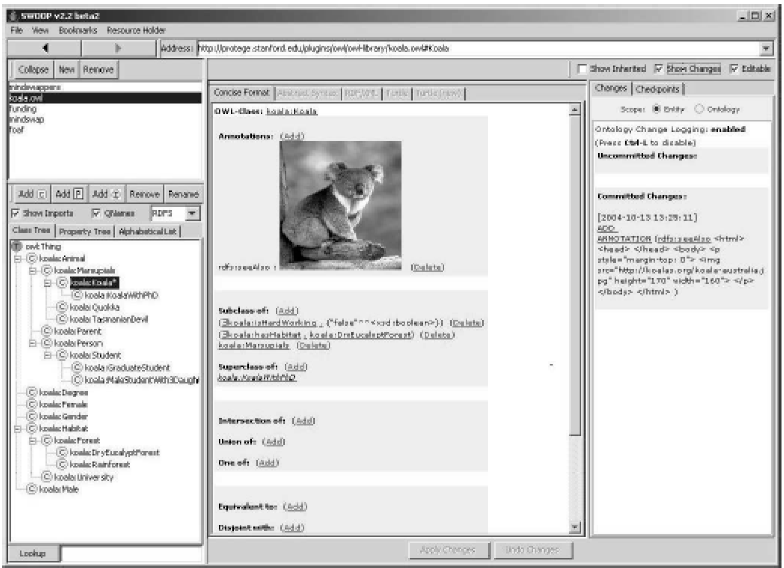
Swoop supports the editing and display of textual or HTML-formatted comments, and of photos and other multimedia (both via HTML and independently) as part of ontologies (see Figure 3). Since OWL ontologies can reference and import other ontologies, one can separate annotations about ontologies from the core ontologies themselves. The Annotea framework (Kahan, Koivunen, Prud'Hommeaux, & Swick, 2001) takes this idea and provides both a specific RDF-based, extensible annotation vocabulary, and a protocol for publishing and finding out-of-band annotations. Swoop uses the Annotea framework as the basis of collaborative ontology development.

Annotea support in Swoop is provided via a simple plug-in whose implementation is based on the standard W3C Annotea protocols (Swick, Prud'Hommeaux, Koivunen, & Kahan, 2001) and uses the default Annotea RDF schema to specify annotations. Any public Annotea server can then be used to publish and distribute the annotations created in Swoop. The default annotation types (comment, advice, example, etc.) seem an adequate base for human-oriented ontology annotations. One extension we have begun experimenting with is "Prototypical Illustration," that is, a photo or drawing that represents a typical or canonical instance of the class.

Change Annotations

We have extended the Annotea Schema with the addition of an OWL ontology for a new class of annotations—ontology changes (similar to Klein & Noy, 2003). The "Change" annotation defined by the Annotea projected was designed to indicate a proposed change to the annotated document, with the proposal described in HTML-marked-up natural language. In our extended ontology, change individuals correspond to specific, undoable changes made in Swoop during editing.

Figure 3. Annotating OWL entities—"Prototypical Illustration" of classes



Swoop uses the OWL API (Bechhofer et al., 2003) to model ontologies and their associated entities, benefiting from its extensive and clean support for changes. The OWL API separates the representation of changes from the application of changes. Each possible change type has a corresponding Java class in the API which is subsequently applied to the ontology (essentially, the Command design pattern). These classes allow for the rich representation changes, including metadata about the changes.

The Swoop change annotations can be published and retrieved by Annotea servers or any other annotation distribution mechanism. The retrieved annotations can then be browsed, filtered, endorsed, recommended, and selectively accepted. It is thus possible to define "virtual versions" of an ontology, by specifying a base ontology and a set of changes to apply to it. This is a fairly new addition to Swoop, and we are just beginning to explore the implications of change tracking, coupled with annotations for the development of large, curated ontologies by collaborative groups of scientists or other ontology definers.

CONCLUSION

We have built a Web (ontology) browser and editor, Swoop, which takes the standard Web browser as the UI paradigm, believing that URIs are central to the understanding and construction of Semantic Web ontologies. The familiar look and feel of a browser emphasized by the address bar and history buttons, navigation side bar, bookmarks, hypertextual navigation, and so forth are all supported for Web ontologies, corresponding with the mental model people have of URI-based Web tools based on their current Web browsers.

All design decisions are in keeping with the OWL nature and specifications.

Thus, multiple ontologies are supported easily, various OWL presentation syntaxes are used to render ontologies, and an OWL reasoner can be integrated for consistency checking. A key point in our work is that the hypermedia basis of the UI is exposed in virtually every aspect of ontology engineering—easy navigation of OWL entities, comparing and editing related entities, search and cross-referencing, multimedia support for annotation, and so forth—thus allowing the Swoop user to take advantage of the Web-based features of OWL significantly more easily than the user of other ontology-editing tools.

In this article, we discuss some of the key issues that our work in Swoop has identified as being important in Web ontology tools. Topics we are currently exploring, not yet implemented in Swoop, are dealing with the ad hoc modification of ontologies by one or more users working on the ontology over time. These are issues exploring the editing of imported ontology data, and the use of annotated ontology change sets for ontology versioning as described above. Currently, we have preliminary solutions for these issues implemented in Swoop, but we are investigating alternate approaches that may be more powerful and better integrated with emerging Web standards. For example, one such approach is the use of the XPointer framework (DeRose, Maler, & Daniel, 2002) to enable efficient syntactic filtering of ontological code, in order to reduce ontology modification time and effort.

REFERENCES

Arpírez, J., Corcho, O., Fernández-López, M., & Gómez-Pérez, A. (2001). WebODE: A scalable ontological engineering workbench. *Proceedings of the 1st International Conference on Knowledge Capture* (K-CAP).

Bechhofer, S., Horrocks, I., Goble, C., & Stevens, R. (2001). OilEd: A reason-able ontology editor for the Semantic Web. *Proceedings of KI2001, Joint German/Austrian Conference on Artificial Intelligence*.

Bechhofer, S., Lord, P., & Volz, R. (2003). Cooking the Semantic Web with the OWL API. *Proceedings of the International Semantic Web Conference*.

Beckett, D. (2004). *Turtle—Terse RDF Triple Language*. Retrieved from www.ilrt.bris.ac.uk/discovery/2004/01/turtle/.

Borgida, A., & Serafini, L. (2003). Distributed description logics: Assimilating information from peer sources. *Journal of Data Semantics*.

Cuenca-Grau, B., Parsia, B., & Sirin, E. (2004). Working with multiple ontologies on the Semantic Web. To appear in *Proceedings of the 3rd International Semantic Web Conference* (ISWC).

Dean, M., & Schreiber, G. (2004). *OWL Web Ontology Language Reference. W3C Recommendation.* Retrieved from www.w3.org/tr/owl-ref/.

DeRose, S., Maler, E., & Daniel Jr., R. (2002). *XPointer xpointer() Scheme.* W3C Working Draft . Retrieved from www.w3.org/tr/xptr-xpointer/.

Farquhar, A., Fickas, R., & Rice, J. (1996). The Ontolingua server: A tool for collaborative ontology construction. *Proceedings of the 10th Banff Knowledge Acquisition for Knowledge-Based System Workshop* (KAW95).

Kahan, J., Koivunen, M.R., Prud'Hommeaux, E., & Swick, R. (2001). Annotea: An open RDF infrastructure for shared Web annotations. *Proceedings of the WWW10 International Conference*.

Klein, M., & Noy, N. (2003). A component-based framework for ontology evolution. *Proceedings of the Workshop on Ontologies and Distributed Systems at IJCAI.*

Kutz, O., Lutz, C., Wolter, F., & Zakharyaschev, M. (2003). E-connections of description logics. *Proceedings of the 2003 International Workshop on Description Logics.*

Noy, N. (2004). *Representing classes as property values on the Semantic Web.* W3C Working Draft. Retrieved from *www.w3.org/tr/2004/wd-swbp-classes-as-values-20040721/*

Noy, N., Sintek, M., Decker, S., Crubezy, M., Fergerson, R., & Musen, M. (2001). Creating Semantic Web contents with Protégé-2000. *IEEE Intelligent Systems*.

Patel-Schneider, P., Hayes, P., & Horrocks, I. (2004). *OWL Web Ontology Language semantics and abstract syntax. W3C Recommendation.* Retrieved from *www.w3.org/tr/2004/rec-owl-semantics-20040210/*

Stojanovic, L., Maedche, A., Motik, B., & Stojanovic, N. (2002). User-driven ontology evolution management. *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web* .

Sure, Y., Erdmann, M., Angele, J., Staff, S., Studer, R., & Wenke, D. (2002). OntoEdit: Collaborative ontology development for the Semantic Web. *Proceedings of the International Semantic Web Conference* (ISWC).

Swick, R., Prud'Hommeaux, E., Koivunen, M.R., & Kahan, J. (2001). *Annotea protocols*. Retrieved from *www.w3.org/2001/Annotea/User/Protocol.html*

ENDNOTES

1 Visit the SWOOP Web site at *http://www.mindswap.org/2004/SWOOP* to obtain the latest information on the tool, to download a free copy of the source code or binary release, and/or to try out the online demo.

2 A URI is a generalization of the more common URL, roughly composed of a naming scheme or protocol indicator (http, ftp, mailto, etc.) a unique indicator (a domain space name for http, a mail address for mailto) and a "fragment id" which is a hash mark followed by a set of characters—thus, for example, an owl class called "person" from an ontology on a University server might be named by the URI *http://www.thisuniversity.edu/OntologyLib/csontology#person*.

3 Bookmarks are another example of hidden URIs, at least in their most common form. Browsers typically have many ways to review bookmarked URIs.

4 We plan to extend our ontology evolution/versioning framework based on related work such as Stojanovic, Maedche, Motik, and Stojanovic (2002) in a specific project or working set. This practice is highly beneficial in understanding and debugging code.

Aditya Kalyanpur is a PhD student at the University of Maryland and a GRA at MINDSWAP. His research interests broadly lie in ontology engineering for the Semantic Web with a focus on tool support for editing, refactoring, annotating, searching, and debugging ontology data. He is also the principal author of SWOOP.

Bijan Parsia is a research philosopher at the University of Maryland specializing in the Semantic Web, description logics, epistemology, oppression theory, AI planning, and the philosophy of mathematics and computer science.

James Hendler is a professor at the University of Maryland and director of Semantic Web and agent technology at the Maryland Information and Network Dynamics Laboratory. He has joint appointments in the Department of Computer Science, the Institute for Advanced Computer Studies and is affiliated with the Institute for Systems Research. One of the inventors of the "Semantic Web," Hendler was the recipient of a 1995 Fulbright Foundation Fellowship and is a fellow of the American Association for Artificial Intelligence.

An Ontology-Based Multimedia Annotator for the Semantic Web of Language Engineering

Artem Chebotko, Wayne State University, USA
Yu Deng, Wayne State University, USA
Shiyong Lu, Wayne State University, USA
Farshad Fotouhi, Wayne State University, USA
Anthony Aristar, Wayne State University, USA

ABSTRACT

The development of the Semantic Web, the next-generation Web, greatly relies on the availability of ontologies and powerful annotation tools. However, there is a lack of ontology-based annotation tools for linguistic multimedia data. Existing tools either lack ontology support or provide limited support for multimedia. To fill the gap, we present an ontology-based linguistic multimedia annotation tool, OntoELAN, which features: (1) the support for OWL ontologies; (2) the management of language profiles, which allow the user to choose a subset of ontological terms for annotation; (3) the management of ontological tiers, which can be annotated with language profile terms and, therefore, corresponding ontological terms; and (4) storing OntoELAN annotation documents in XML format based on multimedia and domain ontologies. To our best knowledge, OntoELAN is the first audio/video annotation tool in the linguistic domain that provides support for ontology-based annotation. It is expected that the availability of such a tool will greatly facilitate the creation of linguistic multimedia repositories as islands of the Semantic Web of language engineering.

Keywords: annotation; general multimedia ontology; GOLD; multimedia; ontology; OWL; Semantic Web

INTRODUCTION

The Semantic Web (Lu, Dong, & Fotouhi, 2002; Berners-Lee, Hendler, & Lassila, 2001) is the next-generation Web, in which information is structured with well-defined semantics, enabling better cooperation of machine and human effort. The Semantic Web is not a replacement, but an extension of the current Web, and its development greatly relies on the availability of ontologies and powerful annotation tools.

Ontology development and annotation management are two challenges of the development of the Semantic Web, as we discussed in Chebotko, Lu, and Fotouhi

(2004). In this article, although we use our developed General Multimedia Ontology as the framework and the GOLD ontology developed at the University of Arizona as an ontology example for ontology-based annotation of linguistic multimedia data, our focus will be on addressing the second challenge—the development of an ontology-based multimedia annotator *OntoELAN* for the Semantic Web of language engineering.

Recently, there is an increasing interest and effort for preserving and documenting endangered languages (Lu et al., 2004; The National Science Foundation, 2004). Many languages are in serious danger of being lost, and if nothing is done to prevent it, half of the world's approximately 6,500 languages will disappear in the next 100 years. The death of a language entails the loss of a community's traditional culture, for the language is a unique vehicle for its traditions and culture.

In the linguistic domain, many language data are collected as audio and video recordings, which impose a challenge to document indexing and retrieval. Annotation of multimedia data provides an opportunity for making the semantics explicit and facilitates the searching of multimedia documents. However, different annotators might use different vocabulary to annotate multimedia, which causes low recall and precision in search and retrieval. In this article, we propose an ontology-based annotation approach, in which a linguistic ontology is used so that the terms and their relationships are formally defined. In this way, annotators will use the same vocabulary to annotate multimedia, so that ontology-driven search engines will retrieve multimedia data with greater recall and precision. We believe that even though in a particular domain, it can be very difficult to enforce a uniform ontology that is agreed on by the whole community, ontology-driven annotation will benefit the community once ontology-aware federated retrieval systems are developed based on ontology techniques such as ontology mapping, alignment, and merging (Klein, 2001).

In this article, we present an ontology-based linguistic multimedia annotation tool, *OntoELAN*—a successor of EUDICO Linguistic Annotator *(ELAN)* (Hellwig & Uytvanck, 2004), developed at the Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands, with the aim to provide a sound technological basis for the annotation and exploitation of multimedia recordings. Although *ELAN* is designed specifically for linguistic domain (analysis of language, sign language, and gesture), it can be used for annotation, analysis, and documentation purposes in other multimedia domains. We briefly describe the features of *ELAN* in the section, "An Overview of OntoELAN," and refer the reader to Hellwig and Uytvanck (2004) for details. *OntoELAN* inherits all *ELAN*'s features and extends the tool with an ontology-based annotation approach. In particular, our main contributions are:

- *OntoELAN* can open and display ontologies, specified in OWL Web Ontology Language (Bechhofer et al., 2004).
- *OntoELAN* allows the creation of a language profile, which enables a user to choose a subset of terms from a linguistic ontology and conveniently rename them if needed.
- *OntoELAN* allows the creation of ontological tiers, which can be annotated with profile terms and, therefore, their corresponding ontological terms.
- *OntoELAN* saves annotations in XML (Bray, Paoli, Sperberg-McQueen, Maler, & Yergeau, 2004) format as class in-

stances of the General Multimedia Ontology, which is designed based on the XML Schema (Fallside, 2001) for *ELAN* annotation files.
- *OntoELAN*, while annotating ontological tiers, creates class instances of corresponding ontologies linked to annotation tiers and relates them to instances of the General Multimedia Ontology classes.

This paper extends the presentation of *OntoELAN* in Chebotko et al. (in press), with more details on ontological and architectural aspects of *OntoELAN* and with a premier on OWL. Since *OntoELAN* is developed to fulfill annotation requirements for the linguistic domain, it is natural that, in this article, we use linguistic annotation examples and link the General Ontology for Linguistic Description (GOLD) (Farrar & Langendoen, 2003) to an ontological tier. To our best knowledge, *OntoELAN* is the first audio/video annotation tool in the linguistic domain that provides support for ontology-based annotation. It is expected that the availability of such a tool will greatly facilitate the creation of linguistic multimedia repositories as islands of the Semantic Web of language engineering.

RELATED WORK

In the following, first we identify the requirements for linguistic multimedia annotation, then we review existing annotation tools with respect to these requirements. We conclude that these tools do not fully satisfy our requirements, and this motivates our development of *OntoELAN*.

Linguistic domain places some minimum requirements on multimedia annotation tools. While semantics-based contents such as speeches, gestures, signs, and scenes are important, color and shape are not of interest. To annotate semantics-based content, a tool should provide a time axis and the capability of its subdivision into time slots/segments, multiple tiers for different semantic content. Obviously, there should be some multimedia resource metadata such as title, authors, date, and time. Additionally, a tool should provide ontology-based annotation features to enable the same annotation vocabulary for a particular domain.

As related work, we give a brief description of the following tools: *Protégé* (Stanford University, 2004), *IBM MPEG-7 Annotation Tool* (International Business Machines Corporation, 2004), and *ELAN* (Hellwig & Uytvanck, 2004).

Protégé is a popular ontology construction and annotation tool developed at Stanford University. *Protégé* supports the Web Ontology Language through the OWL plug-in, which allows a user to load OWL ontologies, annotate data, and save annotation markup. Unfortunately, *Protégé* provides only simple multimedia support through the Media Slot Widget. The Media Slot Widget allows the inclusion and display of video and audio files in *Protégé*, which may be enough for general description of multimedia files like metadata entries, but not sufficient for annotation of a speech, where the multimedia time axis and its subdivision into segments are crucial.

The *IBM MPEG-7 Annotation Tool* was developed by IBM to assist annotating video sequences with MPEG-7 (Martínez, 2003) metadata based on the shots of the video. It does not support any ontology language and uses an editable lexicon from which a user can choose keywords to annotate shots. A *shot* is defined as a time period in video in which the frames have similar scenes. Annotations are saved based on MPEG-7 XML Schema (Martínez, 2003). Although the *IBM MPEG-7 Annotation Tool* was specially designed to annotate video, shot and lexicon-based annotation does not provide enough flexibility for linguistic multimedia annotation. In particular, the shot approach is good for the annotation of content-based features like color and texture, but not for time alignment and time segmentation required for semantics-based content annotation.

ELAN (EUDICO Linguistic Annotator), developed at the Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands, is designed specifically for linguistic domain (analysis of language, sign language, and gesture) to provide a sound technological basis for the annotation and exploitation of multimedia recordings. *ELAN* provides many important features for linguistic data annotation such as time segmentation and multiple annotation layers, but not the support of an ontology. Annotation files are saved in the XML format based on *ELAN* XML Schema.

As a summary, existing annotation tools such as *Protégé* and the *IBM MPEG-7 Annotation Tool* are not suitable for our purpose since they do not support many multimedia annotation operations such as multiple tiers, time transcription, and translation of linguistic audio and video data. *ELAN* is the best candidate for becoming a widely accepted linguistic multimedia annotator, and it is already used by linguists throughout the world. *ELAN* provides most of the required features for linguistic multimedia annotation, which motivates us to use it as the basis for the development of *OntoELAN* to add ontology-based annotation features such as the support of an ontology and a language profile.

AN OVERVIEW OF ONTOELAN

OntoELAN is an ontology-based linguistic multimedia annotator, developed on the top of *ELAN* annotator. It was partially sponsored and developed as a part of Electronic Metastructure for Endangered Languages Data (E-MELD) project. Currently, *OntoELAN* source code contains more than 60,000 lines of Java code and has several years of development history started by the Max Planck Institute for Psycholinguistics team and continued by the Wayne State University team. Both development teams will continue their collaboration on *ELAN* and *OntoELAN*.

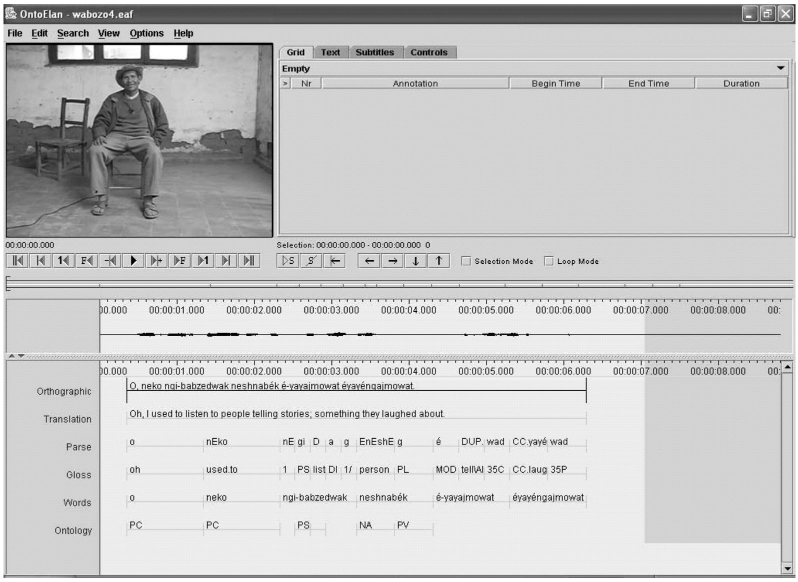
OntoELAN has a long list of detailed descriptions of all its technical features, including the following features that are inherited from *ELAN*:

- display a speech and/or video signals, together with their annotations;
- time linking of annotations to media streams;
- linking of annotations to other annotations;
- unlimited number of annotation tiers as defined by a user;
- different character sets; and
- basic search facilities.

OntoELAN implements the following additional features:

- loading of OWL ontologies;
- language profile creation;
- ontology-based annotation; and
- storing annotations in the XML format based on the General Multimedia Ontology and domain ontologies.

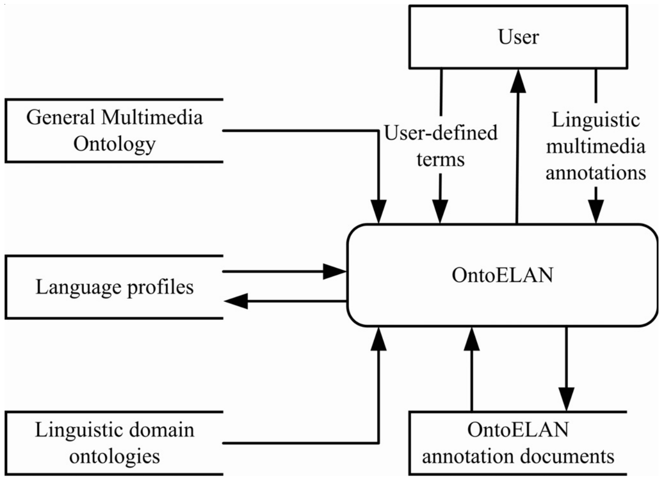
Figure 1. A snapshot of the OntoELAN main window



The main window of *OntoELAN* is shown in Figure 1. *OntoELAN* has the video viewer, the annotation density viewer, the waveform viewer, the grid viewer, the subtitle viewer, the text viewer, the timeline viewer, the interlinear viewer, and associated with them controls and menus. All viewers are synchronized so that whenever a user accesses a point in time in one viewer, all the other viewers move to the corresponding point in time automatically. The video viewer displays video in "mpg" and "mov" formats, and can be resized or detached to play video in a separate window. The annotation density viewer is useful for navigation through the media file and analysis of annotations concentration. The waveform viewer displays the waveform of the audio file in "wav" format; in case of video files, there should be an additional "wav" file present to display waveform. The grid viewer displays annotations and associated time segments for a selected annotation tier. The subtitle viewer displays annotations on selected annotation tiers at the current point in time. The text viewer displays annotations of a selected annotation tier as a continuous text. The timeline viewer and the interlinear viewer are interchangeable, and both display all tiers and all their annotations; only one viewer can be used at a time. In this article, we will mostly work with the timeline viewer (see Figure 1), which allows a user to perform of operations on tiers and annotations. Because a significant part of the *OntoELAN* interface is inherited from *ELAN*, the reader can refer to Hellwig and Uytvanck (2004) for detailed description. *OntoELAN* uses and manages several data sources:

- *General Multimedia Ontology (OWL):* ontological terms for multimedia annotations.

Figure 2. OntoELAN data flow diagram



- *Linguistic domain ontologies (OWL):* ontological terms for linguistic annotations.
- *Language profiles (XML):* a selected subset of domain ontology terms for linguistic annotations.
- *OntoELAN annotation documents (XML):* storage for linguistic multimedia annotations.

A data flow diagram for *OntoELAN* is shown in Figure 2. We do not specify names of most data flows, as they are too general to give any additional information. Two data flows from a user are user-defined terms for language profiles and linguistic multimedia annotations.

In the following sections, we will give more details on *OntoELAN* data sources and data flows. We focus more on the description of features that make *OntoELAN* an ontology-based multimedia annotator, like OWL support, linguistic domain ontology and the General Multimedia Ontology, a language profile, ontological annotation tiers, and so forth.

SUPPORT OF OWL

OWL Web Ontology Language (Bechhofer et al., 2004) is recently recommended as the semantic markup language for publishing and sharing ontologies on the World Wide Web. It is developed as a revision of DAML+OIL language and has more expressive power than XML, RDF, and RDF Schema (RDF-S). OWL provides constructs to define ontologies, classes, properties, individuals, data types, and their relationships. In the following, we present a brief overview of the major constructs and refer the reader to Bechhofer et al. (2004) for more details.

Classes

A class defines a group of individuals that share some properties. A class is defined by *owl:Class*, and different classes can be related by *rdfs:subClassOf* into a class hierarchy. Other relationships between classes can be specified by *owl:equivalentClass*, *owl:disjointWith*, and so forth. The extension of a class can be specified by *owl:oneOf* with a list of class members or by *owl:intersectionOf*, *owl:unionOf* and *owl:complementOf* with a list of other classes.

Properties

A property states relationships between individuals or from individuals to data values. The former is called *ObjectProperty* and specified by *owl:ObjectProperty*. The latter is called *DatatypeProperty* and specified by *owl:DatatypeProperty*. Similarly to classes, different properties can be related by *rdfs:subPropertyOf* into a property hierarchy. The domain and range of a property are specified by *rdfs:domain* and *rdfs:range*, respectively. Two properties might be asserted to be equivalent by *owl:equivalentProperty*. In addition, different characteristics of a property can be specified by *owl:FunctionalProperty*, *owl:InverseFunctionalProperty*, *owl:TransitiveProperty*, and *owl:SymmetricProperty*.

Property Restrictions

A property restriction is a special kind of a class description. It defines an anonymous class, namely the set of individuals that satisfy the restriction. There are two kinds of property restrictions: *value constraints* and *cardinality constraints*. Value constraints restrict the values that a property can take within a particular class, and they are specified by *owl:allValuesFrom*, *owl:someValuesFrom*, and *owl:hasValue*. Cardinality constraints restrict the number of values that a property can take within a particular class, and they are specified by *owl:minCardinality*, *owl:maxCardinality*, *owl:cardinality*, and so forth.

OWL is subdivided into three species (in increasingly-expressive order): OWL Lite, OWL DL, and OWL Full. OWL Lite places some limitations on the usage of constructs and is primarily suitable for expressing taxonomies. For example, *owl:unionOf* and *owl:complementOf* are not part of OWL Lite, and cardinality constraints may only have a 0 or 1 value. OWL DL provides more expressivity and still guarantees computational completeness and decidability. In particular, OWL DL supports all OWL constructs, but places some restrictions (e.g., class cannot be treated as an individual). Finally, OWL Full gives maximum expressiveness, but not computational guarantee.

OntoELAN uses the Jena 2 (Hewlett-Packard Labs, 2004) Java framework for writing Semantic Web applications to provide OWL DL support. On the language profile creation stage, *OntoELAN* basically uses class hierarchy information based on *rdfs:subClassOf* construct. However, while annotating data with ontological terms (by means of a language profile), *OntoELAN* generates dynamic interface for creating instances, assigning property values, and so forth.

LINGUISTIC DOMAIN ONTOLOGY

As a linguistic domain ontology example, we use the General Ontology for Linguistic Description (GOLD) (Farrar & Langendoen, 2003). To make things clear from the beginning, *OntoELAN* does not have GOLD as a component; both are independent. The user can load any other linguistic domain ontology, therefore

OntoELAN can be used as a multimedia annotator in other domains that require similar features. Moreover, the user can load several different ontologies for distinct annotation tiers to provide multi-ontological or even multi-domain annotation approaches. For example, a gesture ontology can be used for linguistic multimedia annotation, as a speaker's gestures help the audience understand the meaning of a speech better. Therefore, linguists can use GOLD in one tier and the gesture ontology in another tier to capture more semantics.

The General Ontology for Linguistic Description is an ongoing research effort led by the University of Arizona to define linguistic domain-specific terms using OWL. GOLD is constantly under revision, and the ontology changes with introduction of new classes, properties, and relations; its structure also changes. Current information about GOLD is available at www.emeld.org, and the ontology is also downloadable from www.u.arizona.edu/~farrar/gold.owl. We briefly describe GOLD content in the next few paragraphs and refer the reader to Farrar and Langendoen (2003) and also to Farrar (2004) for more details.

GOLD provides a semantic framework for the representation of linguistic knowledge and organizes knowledge into four major categories:

- *Expressions:* Physically accessible aspects of a language. Linguistic expressions include the actual printed words or sounds produced when someone speaks. For example, *Orthographic Expression*, *Utterance*, *Signed Expression*, *Word*, *WordPart*, *Prefix*.
- *Grammar:* The abstract properties and relations of a language. For example, *Tense*, *Number*, *Agreement*, *PartOf Speech*.
- *Data Structures:* Constructs that are used by linguists to analyze language data. A linguistic data structure can be viewed as a structuring mechanism for linguistic data content. For example, a lexical entry is a data structure used to organize lexical content. Other examples are a phoneme table and a syntactic tree.
- *Metaconcepts:* The most basic concepts of linguistic analysis. The example of a metaconcept is a language itself.

Through the article we will use only simple GOLD concepts like *Noun*, *Verb*, *Participle*, *Preverb*. They are subclasses of *PartOfSpeech*, and their meaning is easy to understand without special training. Additionally, we will use the concepts *Animate* (living things, including humans, animals, spirits, trees, and most plants) and *Inanimate* (non-living things, such as objects of manufacture and natural "non-living" things), which are two grammatical genders or classes of nouns.

GENERAL MULTIMEDIA ONTOLOGY

Although *OntoELAN* is an ontology-based annotator, a user may not use ontological terms for annotation. In fact, for linguistic multimedia annotation there should usually be several annotation tiers whose annotation is not based on ontological terms. For example, a speech transcription and a speech translation into another language do not use an ontology. Consequently, *OntoELAN* needs to save not only instances of classes created for ontology-based annotations, but also other text data created without ontologies. One solution is to use XML Schema definitions to save an annotation file in the XML format—this is what *ELAN* does. Being consistent in using an ontological approach and, therefore, building the Semantic Web, we provide another solution—the multimedia ontology.

We have developed the multimedia ontology that we called General Multimedia Ontology and that serves as a semantic framework for multimedia annotation. In contrast to domain ontologies, the General Multimedia Ontology is a crucial component of the system. *OntoELAN* saves its annotations in the XML format as class instances of the General Multimedia Ontology and class instances of linguistic domain ontologies that are used in ontological tiers.

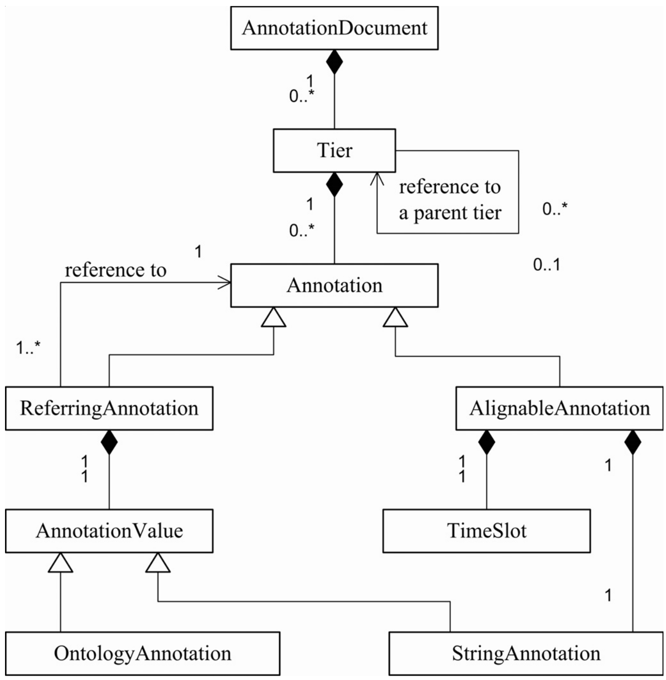
The General Multimedia Ontology is expressed in Web Ontology Language and is designed based on *ELAN* XML Schema for annotation. The General Multimedia Ontology contains the following classes:

- *AnnotationDocument*, which represents the whole annotation document.
- *Tier*, which represents a single annotation tier/layer. There are several types of tiers that a user can choose.
- *TimeSlot*, which represents a concept of a time segment that may subdivide tiers.
- *Annotation*, which can be either *AlignableAnnotation* or *Referring Annotation*.
- *AlignableAnnotation*, which links directly to a time slot.
- *ReferringAnnotation*, which can reference an existing *Alignable Annotation*.
- *AnnotationValue*, which has two subclasses *StringAnnotation* and *Ontology Annotation* that represent two different ways of annotating.
- *MediaDescriptor*, *TimeUnit* and others.

Relationships among some important General Multimedia Ontology classes are presented in Figure 3. In general, *AnnotationDocument* may have zero or many *Tiers*, which, in turn, may have zero or many *Annotations*. *Annotation* can be either *AlignableAnnotation* or *ReferringAnnotation*, where *Alignable Annotation* can be divided by *TimeSlots*, and *ReferringAnnotation* can refer to another annotation. *ReferringAnnotation* may refer to *AlignableAnnotation*, as well as to *ReferringAnnotation*, but the root of the referenced annotations must be an *AlignableAnnotation*. Each *Annotation* has one *AnnotationValue*, which can be either a *StringAnnotation* or an *OntologyAnnotation*. *StringAnnotation* represents any string that a user can input as an annotation value, but values, represented by *OntologyAnnotation*, come from a language profile and, consequently, from an ontology. Note that the General Multimedia Ontology allows *Ontology Annotation* to be used only with *ReferringAnnotation*. In other words, tiers with *AlignableAnnotations* do not support an ontology-based approach. This limitation is due to software development issues—*OntoELAN* does not support annotation with ontological terms in alignable tiers. We intentionally emphasize this constraint in the ontology, although conceptually it should not be the case. Among our contributions is the introduction of the OWL class *Ontology Annotation*, which serves as an annotation unit for an ontology-based annotation. *OntologyAnnotation* has restrictions on the following properties:

- *hasOntAnnotationId:* The ID of the annotation. The property cardinality equals one (*owl:cardinality = 1*).

Figure 3. Relationships among some General Multimedia Ontology classes (UML class diagram)



- *hasUserDefinedTerm*, which relates *OntologyAnnotation* to a term in a language profile (described in the next section). The property cardinality equals one (*owl:cardinality = 1*).
- *hasInstances*, which relates *Ontology Annotation* to a term (represented as an instance) in an ontology used for annotation. The property cardinality is greater than zero (*owl:minCardinality = 1*).
- *hasOntAnnotationDescription:* Descriptions/comments on the annotation. The property cardinality is not restricted.

The General Multimedia Ontology is available at database.cs.wayne.edu/proj/OntoELAN/multimedia.owl. We will add new concepts to the ontology in case if *OntoELAN* needs them for annotation. We have developed the General Multimedia Ontology especially for *OntoELAN* and have not included most concepts in multimedia domain. In particular, we did not include multimedia concepts such as those related to shapes, colors, motions, audio spectrum, and so forth. Our small ontology focuses on high-level multimedia annotation features and can be used for similar annotation tasks.

LANGUAGE PROFILE

A language profile is a subset of ontological terms, possibly renamed, that are used in the annotation of a particular multimedia resource. The idea of a language profile comes from the following practical issues related to an ontology-based annotation.

A domain ontology defines all terms related to a particular domain, and the number of terms is usually considerably large. However, to annotate a concrete data resource, an annotator usually does not need all terms from an ontology. Moreover, an experienced annotator can identify a subset of ontological terms that will be useful for a given resource. Speaking in terms of a linguistic domain, an annotator will only use a subset of GOLD to annotate a particular language and may need a different subset for another language.

Linguists have been annotating multimedia data for years without standardized terms from an ontology. They have their individual sets of terms that they are accustomed to using for annotation. It will be difficult to come to a consensus about class names in GOLD so that every linguist is satisfied with it. Additionally, linguists widely use abbreviations like "n" for "noun" which is concise and convenient. Finally, linguists whose native language is, for example, Ukrainian may prefer to use annotation terms in Ukrainian rather than in English.

More formally, a language profile is defined as a quadruple: ontological terms; user-defined terms; a mapping between ontological terms and user-defined terms; and a reference to an ontology, which contains the structural information about terms (like subclass relationship). In summary, a language profile in *OntoELAN* provides convenience and flexibility for a user to:

- select a subset of ontological terms useful for a particular resource annotation;
- rename ontological terms, for example, use another language, give an abbreviation or a synonym;
- combine the meaning of two or many ontological terms in one user-defined term (e.g., ontological terms "Inanimate" and "Noun" may be conveniently renamed as "NI").

OntoELAN allows ontology-based annotation by means of a language profile. A user opens an ontology, creates a profile, and links it to an ontological tier. Annotation values for an ontological tier can only be selected from a language profile. A language profile in *OntoELAN* is represented as a simple XML document (see Figure 4) with a specified schema, which basically maps ontological terms to user-defined terms, and has a link to the original ontology and some metadata. A user can easily create, open, edit, and save profiles with *OntoELAN*.

Figure 4 presents an example language profile, created by the author Artem and linked to GOLD ontology at URI www.u.arizona.edu/~farrar/gold.owl. In

Figure 4. An example of the language profile XML document

```
<?xml version="1.0" encoding="UTF-8"?>
<PROFILE AUTHOR="Artem" DESCRIPTION="" VERSION="1.0"
SOURCE= "http://www.u.arizona.edu/~farrar/gold.owl">
<USER_DEFINED_TERM DESCRIPTION="" NAME="NI">
   <ONTOLOGY_TERM NAME="Noun"/>
   <ONTOLOGY_TERM NAME="Inanimate"/>
</USER_DEFINED_TERM>
</PROFILE>
```

this example, there is only one user-defined term "NI" that maps to ontological terms "Noun" and "Inanimate." This is a one-to-many mapping, but a mapping can be many-to-many as well. For example, we can add another user-defined term "IN" that maps to the same ontological terms "Noun" and "Inanimate." In general, a mapping can be one-to-one, one-to-many, many-to-one, or many-to-many.

ANNOTATION TIERS AND LINGUISTIC TYPES

OntoELAN allows a user to create an unlimited number of annotation tiers. Multiple-tier feature is a must for linguistic multimedia annotation. For example, while annotating an audio monolog, a linguist may choose separate tiers to write a monolog transcription, a translation, a part of speech annotation, a phonetic transcription, and so forth.

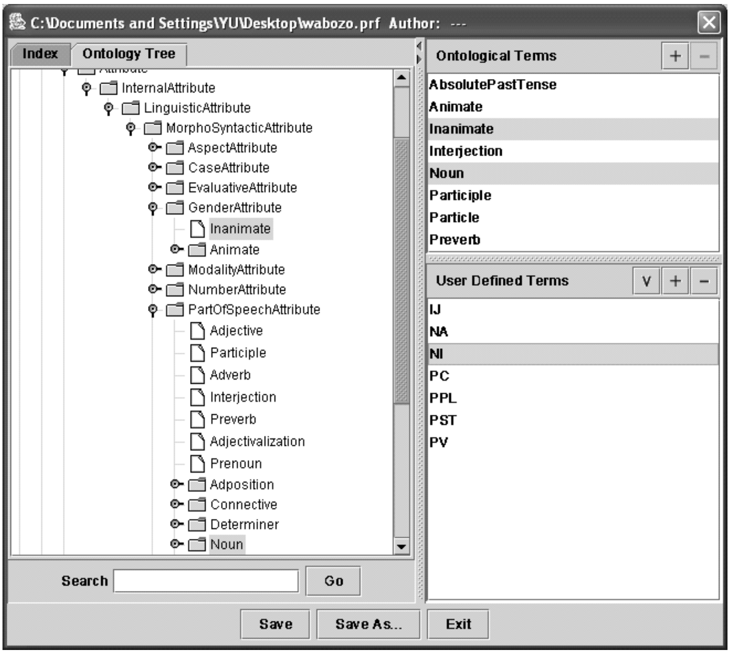
An annotation tier can be either *alignable* or *referring*. Alignable tiers are directly linked to the time axis of an audio/video clip and can be divided into segments (time slots); referring tiers contain annotations that are linked to annotation on another tier, which is also called *a parent tier* and can be alignable or referring. Thus, tiers form a hierarchy, where its root must be an alignable tier. Following the previous example, the speech transcription could be an independent time-alignable tier that is divided into time slots of the speaker's utterances. On the other hand, the translation-referring tier could refer to the transcription tier, so that the translation tier inherits its time alignment from the transcription tier.

After a tier hierarchy is established, changes in one tier may influence other tiers. Deletion of a parent tier is cascaded: all its child tiers are automatically deleted. Similarly, this is true about annotations on a tier: deletion of an annotation on a parent tier causes the deletion of all corresponding annotations on its child tiers. Alteration of the time slot on a parent tier influences all child tiers as well.

Each annotation tier has associated with it linguistic type. There are five predefined linguistic types in *OntoELAN* which put some constraints on tiers assigned to them. The first four of them are described in Hellwig and Uytvanck (2004), and we also give their definitions here:

- *None:* The annotation on the tier is linked directly to the time axis. This is the only type that alignable tiers can have.
- *Time Subdivision:* The annotation on the parent tier can be subdivided into smaller units, which, in turn, can be linked to time slots. They differ from annotations on alignable tiers in that they are assigned to a slot that is contained within the slot of their parent annotation.
- *Symbolic Subdivision:* Similar to the previous type, but the smaller units cannot be linked to the time slots.
- *Symbolic Association:* The annotation on the parent tier cannot be subdivided further, so there is a one-to-one correspondence between the parent annotation and its referring annotation.
- *Ontological Type:* The annotation on such a tier is linked to a language profile. This is not an independent type, as it can be used only in combination with referring tier types such as *Time Subdivision*, *Symbolic Subdivision*, or *Symbolic Association*. To emphasize that a referring tier allows ontology-based annotation, we call it an ontological tier.

Figure 5. A snapshot of creating a language profile



Only ontological tiers allow annotation based on language profile terms; other types of tiers allow annotation with any string value.

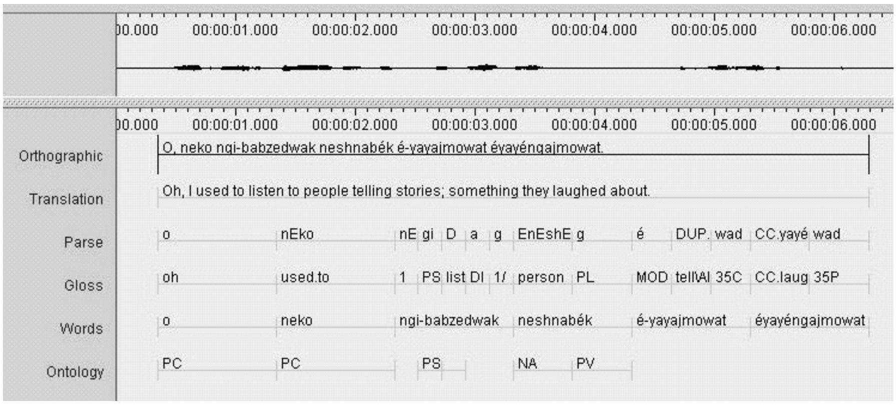
LINGUISTIC MULTIMEDIA ANNOTATION WITH ONTOELAN

In this section, we describe an annotation process in *OntoELAN* using a linguistic multimedia resource annotation example. In general, an annotation process in *OntoELAN* consists of three major steps: (1) language profile creation, (2) creation of tiers, and (3) creation of annotations. The first step is unnecessary if ontological tiers will not be defined. The second step can be completed partially for non-ontological tiers before the creation of a language profile. It is also possible to have multiple profiles for multiple ontological tiers, but there is always one-to-one correspondence between a profile and an ontological tier.

As an example, we annotate the audio file, which contains a sentence in Potawatomi, one of the North American native languages.

We first load GOLD ontology and create the Potawatomi language profile. Figure 5 presents a snapshot of the profile creation window. The tabs "Index" and "Ontology Tree" on the left provide two views of an ontology: a list view, which displays all the terms of an ontology alphabetically as a list, and a hierarchical view, which displays all the terms of an ontology in a hierarchical fashion to illustrate parent-child relationships between terms. From any of these two views, a user can select required terms and add them to the "Onto-

Figure 6. A snapshot of annotation tiers in the OntoELAN main window



logical Terms" list, and rename ontological terms as shown in the "User-Defined Terms" list. In Figure 5, we selected the ontological terms "Inanimate" and "Noun" and combine them under one user-defined term "NI."

After the language profile is ready, we define six tiers in the *OntoELAN* main window (see Figure 6):

- *Orthographic* of type "None" (linked to the time axis)
- *Translation* of type "Symbolic Association" (referring to *Orthographic*)
- *Words* of type "Symbolic Subdivision" (referring to *Orthographic*)
- *Parse* of type "Symbolic Subdivision" (referring to *Words*)
- *Gloss* of type "Symbolic Association" (referring to *Parse*)
- *Ontology* of type "Symbolic Association" and "Ontological Type" (referring to *Gloss*)

The created tier hierarchy is shown in Figure 7.

Finally, we specify annotation values on all six tiers (see Figure 6). We annotate the *Orthographic* tier first, because it is the root of the tier hierarchy, and its time alignment is inherited by other tiers. We do not divide the *Orthographic* tier into time slots, and its time axis contains the whole sentence in Potawatomi. The *Translation* tier inherits time alignment from its parent and cannot subdivide it any further (type "Symbolic Association"). The *Words* tier also inherits *Orthographic* time alignment, but in this case we subdivide it into segments that correspond to words in the sentence. Similarly, we subdivide the *Parse* tier alignment inherited from *Words*. The *Gloss* tier inherits alignment from *Parse*, and the *Ontology* tier inherits alignment from *Gloss*; both *Gloss* and *Ontology* do

Figure 7. A snapshot of the tier hierarchy

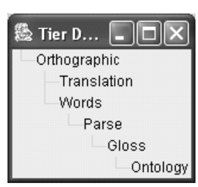


Figure 8. An example of the XML markup for the OntoELAN annotation

```
...
<media:Tier rdf:ID="Ontology">
  <media:hasTierID>Ontology</media:hasTierID>
  <media:hasParent rdf:resource="file:///C:/wabozo4.eaf#Gloss"/>
  <media:hasProfile>C:\wabozo.prf</media:hasProfile>
  <media:hasLinguisticType>
    <media:LinguisticType rdf:ID="ontology">
      <media:hasTimeAlignable>false</media:hasTimeAlignable>
      <media:hasLinguisticTypeID>ontology</media:hasLinguisticTypeID>
      <media:hasConstraint rdf:resource="file:///C:/wabozo4.eaf#Symbolic_Association"/>
      <media:hasGraphicRef>false</media:hasGraphicRef>
    </media:LinguisticType>
  </media:hasLinguisticType>
  ...
</media:Tier>
  ...
<media:RefAnnotation rdf:ID="a42">
  <media:hasAnnotationID>a42</media:hasAnnotationID>
  <media:hasAnnotationRef rdf:resource="file:///C:/wabozo4.eaf#a31"/>
  <media:hasAnnotationValue>
    <media:OntologyAnnotation rdf:ID="a42Value">
      <media:hasUserDefinedTerm>PV</media:hasUserDefinedTerm>
      <media:hasInstances
           rdf:resource="http://www.u.arizona.edu/~farrar/gold.owl#Preverb"/>
      <media:hasOntAnnotationDescription>comments</media:hasOntAnnotationDescription>
      <media:hasOntAnnotationId>e</media:hasOntAnnotationId>
    </media:OntologyAnnotation>
  </media:hasAnnotationValue>
</media:RefAnnotation>
...
```

not allow further subdivision. Correct alignment inheritance is important, because there is a semantic correspondence between segments of different tiers. For example, if we look at a Potawatomi word "neko" in the *Words* tier, we can find its gloss "used to" in the *Gloss* tier and part of speech "PC" (maps to GOLD *Participle* concept) in the *Ontology* tier.

Except for the annotations on the *Ontology* tier, which is defined as an ontological tier, all the annotations are annotated by a string value. Unlike the text annotation, the user annotates the ontological tier by selecting a user-defined term from the profile. Once the term is selected, the next step is creating individuals of the corresponding ontological term(s). The user needs to do nothing if the ontological term is defined as an instance in the ontology, to input an instance name if the ontological term is defined as a class with no restrictions, or to provide all information based on the definition of the ontological class, properties, and so forth.

The annotation is saved in the XML format as instances of the General Multimedia Ontology and, in our case, GOLD. The example of the XML markup for the *Ontology* tier instance and referring annotation instance with ID "a42" on that tier is shown in Figure 8. For the *Ontology* tier, several properties are defined such as ID, parent tier, profile, linguistic type, and so forth. For the referring annotation, *OntoELAN* has defined ID, reference to another annotation, and annotation value that includes an *OntologyAnnotation* class instance with ID, user-defined term "PV," and reference to GOLD concept *Preverb*, which is defined as an instance. The markup in Figure 8 is based on the General Multimedia Ontology, except the reference to a GOLD instance mentioned above.

CONCLUSIONS AND FUTURE WORK

In this article, we address the challenge of annotation management for the Semantic Web of language engineering. Our contribution is the development of *OntoELAN*, a linguistic multimedia annotation tool that features an ontology-based annotation approach. *OntoELAN* is the first attempt at annotating linguistic multimedia data with a linguistic ontology. Meanwhile, the ontological annotations share the data on the linguistic ontologies. Future work will improve the system and provide more channels for sharing data on the Web, such as the multimedia descriptions, the language words, and so forth. Also, a future version will improve the current searching system, which supports text searching and retrieval in one annotation document, to search, retrieve, and compare the linguistic multimedia annotation data on the Web. Additionally, we plan to integrate a text document annotation into *OntoELAN* and include semi-automatic annotation support, similar to *Shoebox* (SIL International, 2000).

ACKNOWLEDGMENTS

Developers of *ELAN* from Max Planck Institute for Psycholinguistics, Hennie Brugman, Alexander Klassmann, Han Sloetjes, Albert Russel, and Peter Wittenburg, provided us with *ELAN*'s source code and documentation. Also, we would like to thank Dr. Laura Buszard-Welcher and Andrea Berez from the E-MELD (Electronic Metastructure for Endangered Languages Data) project for their constructive comments on *OntoELAN*.

REFERENCES

Bechhofer, S., Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D., Patel-Schneider, et al. (2004). *OWL Web Ontology Language reference. W3C Recommendation*. Retrieved from *www.w3.org/TR/owl-ref/*

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*. Retrieved from *www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21*

Bray, T., Paoli, J., Sperberg-McQueen, C., Maler, E., & Yergeau, F. (2004). *Extensible Markup Language (XML) 1.0 (Third Edition). W3C Recommendation*. Retrieved from *www.w3.org/TR/REC-xml/*

Chebotko, A., Deng, Y., Lu, S., Fotouhi, F., Aristar, A., Brugman, H., et al. (in press). OntoELAN: An ontology-based linguistic multimedia annotator. *Proceedings of the IEEE 6th International Symposium on Multimedia Software Engineering (IEEE-MSE 2004)*, Miami, FL, USA.

Chebotko, A., Lu, S., & Fotouhi, F. (2004, April). Challenges for information systems towards the Semantic Web. *AIS SIGSEMIS Semantic Web and Information Systems Newsletter, 1*. Retrieved from *www.sigsemis.org/newsletter/newsletter/April2004/FINAL_AIS_SIGSEMIS_Bulletin_1_1_04_1_.pdf*

Fallside, D.C. (2001). *XML Schema part 0: Primer. W3C Recommendation*. Retrieved from *www.w3.org/TR/xmlschema-0/*

Farrar, S. (2004). *GOLD: A progress report*. Retrieved from *www.u.*

arizona.edu/~farrar/gold-status-report.pdf

Farrar, S., & Langendoen, D.T. (2003). A linguistic ontology for the Semantic Web. *GLOT International*, *7*(3), 97-100.

Hellwig, B., & Uytvanck, D. (2004). *EUDICO Linguistic Annotator (ELAN) Version 2.0.2 manual [software manual]*. Retrieved from *www.mpi.nl/tools/ELAN/ELAN_Manual-04-04-08.pdf*

Hewlett-Packard Labs. (2004). *Jena 2—a Semantic Web framework [computer software]*. Retrieved from *www.hpl.hp.com/semWeb/jena2.htm*

International Business Machines Corporation. (2004). *IBM MPEG-7 Annotation Tool [computer software]*. Retrieved from *www.alphaworks.ibm.com/tech/videoannex*

Klein, M. (2001). Combining and relating ontologies: An analysis of problems and solutions. *Proceedings of the IJCAI-2001 Workshop on Ontologies and Information Sharing*, Seattle, WA, USA.

Lu, S., Dong, M., & Fotouhi, F. (2002). The Semantic Web: Opportunities and challenges for next-generation Web applications. *International Journal of Information Research*, *7*(4). Retrieved from *InformationR.net/ir/7-4/paper134.html*

Lu, S., Liu, D., Fotouhi, F., Dong, M., Reynolds, R., Aristar, A., et al. (2004). Language engineering for the Semantic Web: A digital library for endangered languages. *International Journal of Information Research*, *9*(3). Retrieved from *InformationR.net/ir/9-3/paper176.html*

Martínez, J.M. (2003). *MPEG-7 overview (version 9). International Organisation for Standardisation*. Retrieved from *www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm*

SIL International. (2000). *The linguist's Shoebox: Tutorial and user's guide [software manual]*. Retrieved from *www.sil.org/computing/shoebox/ShTUG.pdf*

Stanford University. (2004). *The Protégé Project [computer software]*. Retrieved from protege.stanford.edu.

The National Science Foundation. (2004). *NSF 04-605. Documenting endangered languages (DEL)*. Retrieved from *www.nsf.gov/pubs/2004/nsf04605/nsf04605.htm*

Artem Chebotko is a PhD student in the Department of Computer Science at Wayne State University. His research interests include databases and the Semantic Web. He is a student member of the IEEE.

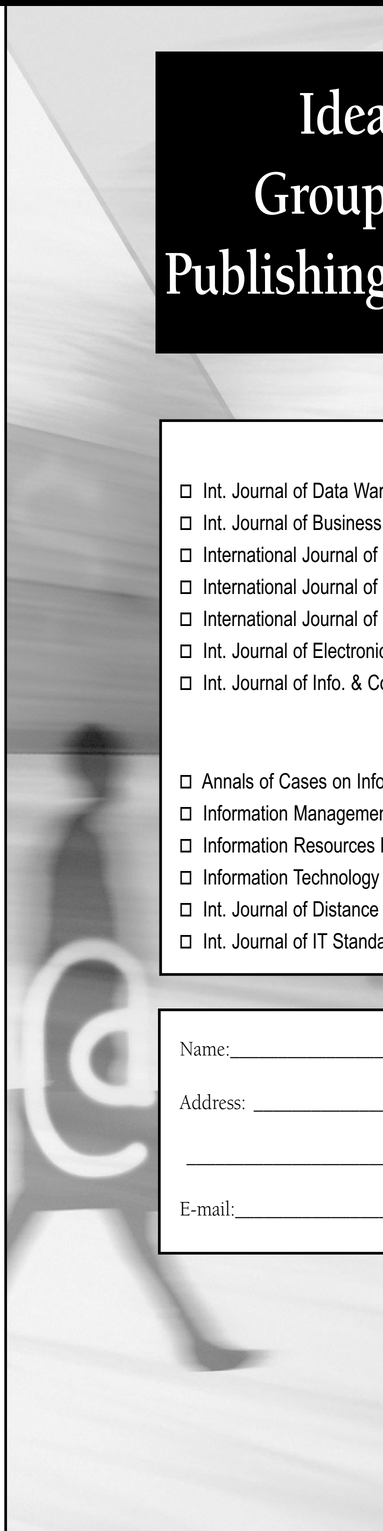
Yu Deng has graduated from Wayne State University with an MS in computer science (2004). Her research interests include databases and the Semantic Web.

Shiyong Lu received his PhD in computer science from the State University of New York at Stony Brook (2002). He is currently an assistant professor of the Department of Computer Science, Wayne State University (USA). His research interests include databases, the Semantic Web and bioinformatics. He has published more than 30 papers in top international conferences and journals in the above areas. He is a member of the IEEE.

Dr. Fotouhi received his PhD in computer science from Michigan State University in 1988. He joined the Faculty of Computer Science at Wayne State University in August 1988 where he is

currently professor and chair of the department. Dr. Fotouhi major area of research is databases, including object-relational databases, multimedia information systems, bioinformatics, medical image databases and web-enabled databases. He has published more than 80 papers in refereed journals and conference proceedings, served as a program committee member of various database related conferences.

Dr. Aristar received his PhD in linguistics from University of Texas in 1984. He was a researcher at Microelectronics & Computer Technology Corporation, 1984-1989; assistant professor, University of Australia, 1990-1991; assistant professor, Texas A&M University, 1991-1995; and associate professor, Texas A&M University, 1996-1998. He joined the Department of English at Wayne State University in 1998 where he is currently an associate professor. Dr. Aristar is the chairman of OLAC Working Group on Language Codes; moderator & founder of LINGUIST List; and organizer of several endangered languages related workshops.



NEW RELEASES

E-Commerce and M-Commerce Technologies

P. Candace Deans, University of Richmond, USA

E-Commerce and M-Commerce Technologies explores the emerging area of mobile commerce. The chapters in this book look specifically at the development of emerging technologies and their application in Internet commerce. From E-business to mobile database developments, this book offers a compilation of readings that are useful to individuals and organizations in the academic study and research surrounding mobile commerce as well as in the practical application of these technologies.

ISBN 1-59140-566-7 (h/c) • US$79.95
ISBN 1-59140-239-5 (s/c) • US$64.95
280 pages • Copyright © 2005

Innovations of Knowledge Management

Bonnie Montano, Georgetown University, USA

Innovations of Knowledge Management highlights the broad range of topics that fall under the term knowledge management, thus emphasizing the large role knowledge management plays in organizations. As a compilation of some of the most recent work in the field, the included chapters truly present innovations in how organizations can and should manage their knowledge.

ISBN 1-59140-281-6 (h/c) • US$79.95
ISBN 1-59140-229-8 (s/c) • US$64.95
364 pages • Copyright © 2005

It's Easy to Order!
www.idea-group.com
717/533-8845 x10 • Toll free 1-866-342-6657
Mon-Fri 8:30 am-5:00 pm (est) or fax 24 hours a day 717/533-8661

IRM Press
Hershey • London • Melbourne • Singapore

An excellent addition to your library!

ISSN 1552-6283