Position Papers for

SWWS' 01

The First Semantic Web Working Symposium

Stanford University, California, USA
July 30 - August 1, 2001

We appreciate the contributions from our institutional sponsors:

National Science Foundation

OntoWeb Network

DARPA DAML

INRIA

We appreciate the contributions from our industrial sponsors:

Organizers

Isabel F. Cruz
U. Illinois at Chicago
USA
(ifc@cs.uic.edu)

Stefan Decker
Stanford U.
USA
(stefan@db.stanford.edu)

Jérôme Euzenat
INRIA
France
(Jerome.Euzenat@inrialpes.fr)

Deborah McGuinness
Stanford U.
USA
(dlm@ksl.stanford.edu)

PC Members

Dan Brickley, University of Bristol, UK
Tiziana Catarci, University of Rome "La Sapienza", Italy
Vassilis Christophides, ICS-FORTH, Greece
Steve Demurjian, University of Connecticut, USA
Max J. Egenhofer, University of Maine, USA
Peter Eklund, Griffith University, Australia
Dieter Fensel, Free University of Amsterdam, The Netherlands
Asunciòn Gomez-Perez, Universidad Politecnica de Madrid (UPM), Spain
Benjamin Grosof, MIT, USA
Natasha Fridman Noy, Stanford University, USA
Nicola Guarino, CNR, Italy
Pat Hayes, University of West Florida, USA
Jim Hendler, DARPA and University of Maryland, USA
Masahiro Hori, IBM Tokyo Research Laboratory, Japan
Ian Horrocks, University of Manchester, UK
Ora Lassila, Nokia Research, USA
Raphael Malyankar, Arizona State University, USA
Massimo Marchiori, W3C, University of Venice, USA, Italy
Brian McBride, Hewlett Packard, UK
Sheila McIlraith, Stanford University, USA
Robert Meersman, Free University Of Brussels, Belgium
Eric Miller, W3C, USA
Enrico Motta, The Open University, UK
Amedeo Napoli, LORIA, France
Dimitris Plexousakis, ICS-FORTH & Univ. of Crete, Greece
Peter Patel-Schneider, Lucent Technologies, USA
Guus Scheiber, University of Amsterdam, The Netherlands
Amit Sheth, University of Georgia and Taalee Inc, USA
Steffen Staab, University of Karlsruhe, Germany
Heiner Stuckenschmidt, University of Bremen, Germany
Frank van Harmelen, Free University of Amsterdam, The Netherlands

Reviewers

Sofia Alexaki
Cecilia Bastarrica
Peter Becker
Olivier Brunet
Paul Calnan
Farid Cerbah
Ying Ding
Michael Klein
Maurizio Lenzerini
Tarcisio Lima
Donald Needham
Borys Omelayenko
Charles E. Phillips, Jr.
Jeff Z. Pan
Margie Price
Mitchell Saba
Feng Zhang

Table of Contents

SWWS 2001 - Position Paper

Supporting Knowledge Discovery on the Semantic Web by Exploiting the Semantics of Complex Relationships

K. Anyanwu, A. Sheth, *

* LSDIS Lab {anyanwu,amit@cs.uga.edu.

Department of Computer Science, University of Georgia, Athens, GA 30602, USA,

Summary

Current research efforts into the role of ontologies in the Semantic Web have focused mainly on semantic modeling, querying, information exchange and integration. Correspondingly, most of the specification languages including DAML/OIL, as well as other XML/RDF based languages, e.g. SHOE, provide support for representing basic semantic relationships between ontology concepts, typically in a single domain. The reasoning support provided for systems is limited to that supported by description logic, i.e., subsumption, which may be used to derive relationships of a hierarchical nature like is-a relationships, and sometimes instance-of relationships.

Focus of our research is on knowledge discovery and complex information requests that may involve correlating information from disparate domains. The entities in the different domains may have complex relationships e.g. causal relationships whose semantics cannot be modeled using current ontology representation languages. In this position paper, we motivate the need for support of complex relationships with a scenario. We outline how our approach for supporting inter-domain relationships together with a rich query mechanism can be used to support knowledge discovery.

Background

The modeling primitives available in present-day ontology languages allow for expressing mainly hierarchical relationships (inheritance) and relational properties like transitivity, symmetry etc. This affords a reasonable level of reasoning capability, but is somewhat limited in the nature of questions that can be answered by systems. For example, questions of the is-a, part-of, instance-of, nature may be readily answered by such systems. In contrast, a system like InfoQuilt that supports semantic information correlation [2] with more complex relationships both within and across domains can be used to answer more exploratory questions. For example, one question that has recently become of interest to geography researchers is ìDo Nuclear Tests cause Earthquakes?î. Considering the wealth of information on the web, it should be possible for a Semantic Web solution to correlate and analyze information from federated (heterogeneous and autonomous) information sources containing information on nuclear tests and those containing information on earthquakes, and come up with a preliminary answer to that question. The capability that is needed here, and is provided by InfoQuilt, goes beyond the traditional approach of providing integrated views of multiple data sources. It involves the ability to express and represent complex inter-domain relationships that can be exploited by the system to perform useful correlation amongst the different domain sources, as well as a rich information request mechanism that allows users to express more meaningfully their information need.

In the next section, we will outline how the unique features of the InfoQuilt system can be used to support complex relationships and knowledge discovery.

Complex Relationship Support In InfoQuilt

The aforementioned example explores the hypothetical causal relationship between Nuclear Tests and Earthquakes. In the InfoQuilt system, the user can explicitly represent detailed semantics of this relationship, using attributes/properties from the Nuclear Tests and Earthquake ontologies, and a library of operators and functions. The notation that is used below is a concise representation of the domain ontologies, containing domain attributes, attribute properties, and rules.

NuclearTest (testSite, explosiveYield, waveMagnitude, testType, eventDate, conductedBy, latitude, longitude, waveMagnitude > 0, waveMagnitude < 10, testSite -> latitude longitude);

Earthquake (eventDate, description, region, magnitude, latitude, longitude, numberOfDeaths, damagePhoto, magnitude > 0);

The following is a representation of the causal relationship.

NuclearTestCausesEarthquake : <= dateDifference(NuclearTest.eventDate, Earthquake.eventDate) < 30 AND *distance(NuclearTest.latitude, NuclearTest.longitude, Earthquake,latitude, Earthquake.longitude) < 100*

This relationship can be verbalized as follows: A nuclear test may be said to have caused an earthquake, if the earthquake occurred within thirty days and 100 miles of the test explosion.

Additionally, InfoQuilt provides a construct called an Iscape or Information Scape, which is used represent a userís information request. An Iscape is semantically richer than traditional keyword based or attribute based (e.g., SQL) query mechanisms because it contains information from the domain ontologies as well as any specified relationships between the domains to process and evaluate an information request. Consider the following text version of an Iscape.

ìFind all nuclear tests conducted by USSR or US after 1970 and find any information about any earthquakes that could have potentially occurred due to these tests.î

To answer this question, the query planner extracts information from only relevant nuclear test information sources, as well as relevant information sources on earthquakes. Then using information about the *NuclearTestCausesEarthquake* relationship, will compare dates and locations of nuclear tests and earthquakes and eliminate those that do not meet the relationshipís constraint [1,3].

Knowledge Discovery Using InfoQuilt

The framework provided by InfoQuilt can be used to support knowledge discovery either by formulating complex information requests. Alternatively a user may pose a hypothesis involving complex relationships between data from heterogeneous and autonomous Web-accessible information sources. Corresponding results can help either justify or falsify their hypothesis and guide further requests. For example, to explore the aforementioned relationship, we can try the following sequence of Iscapes

"*Find when the earliest recorded nuclear test was conducted.*"

We find from the results that nuclear testing began in 1950. So we use a few more Iscapes whose results show that there is a sudden increase in the number of earthquakes since 1950, and that in the period 1900-1949, the average rate of earthquakes was 68 per year and that for 1950-present was 127 per year, that is, it almost doubled. Next, we try to analyze the same data grouping the earthquakes by their magnitudes.

"*For each group of earthquakes with magnitudes in the ranges 5.8-6, 6-7, 7-8, 8-9, and magnitudes higher than 9 on the Richter scale per year starting from year 1900, find the average number of earthquakes.*"

The results show that the average number of earthquakes with magnitude greater than 7 on the Richter scale have remained practically constant over the century (about 19).

We can therefore deduce that the earthquakes caused by nuclear tests usually are of magnitudes less than 7 on the Richter scale. We can then try to explore the data at a finer level of granularity by trying to look for specific instances of earthquakes that occurred within a certain period of time after a nuclear test was conducted in a near by region.

Conclusion

InfoQuilt provides support for representing and utilizing (1) domain knowledge including concepts, relationships, domain rules and data dependencies, (2) complex inter-ontology relationships, (3) a semantically rich information request mechanism, (4) modeling of information resources which captures the nature of content present, and (5) a library of operators and functions (user-defined) that are useful in defining semantic relationships as well as resolving syntactic heterogeneities.

We believe that taken together, these components support *deeper* semantics and provide a framework for supporting (defining, sharing, executing) semantic information correlations and complex semantic relationships between data managed by Web-accessible heterogeneous and autonomous information, such as InfoQuiltís Iscapes, should be investigated to future enrich the rapidly evolving Semantic Web.

References (also see: http://lsdis.cs.uga.edu/proj/iq/iq_pub.html)

1. S. Patel, and A. Sheth, "Planning And Optimizing Semantic Information Requests Using Domain Modeling And Resource Characteristics ". to appear in 6th Intl Conf on Cooperative Information Systems (CoopIS 2001), Italy. Spetember 2001.
2. K. Shah and A. Sheth, "Logical Information Modeling of Web-accessible Heterogeneous Digital Assets", Proc. of the Advances in Digital Libraries," (ADL'98), Santa Barbara, CA, May 28-30, 1998.
3. S. Thacker, A. Sheth, and S. Patel, Complex Relationships for the Semantic Web, to appear in Creating the Semantic Web, D. Fensel, J. Hendler, H. Liebermann, and W. Wahlster (eds.) MIT Press, 2001.

Towards a Knowledge-Level Software Platform

A position paper for the International Semantic Web Workshop
Stanford, July, 2001

Avron Barr, Shirley Tessler and Buddy Kresge*

A software platform is, to students of the software industry, what volcanoes are to geologists – dramatic manifestations of powerful subterranean forces, often colossal, sometimes threatening. There have been many software platforms over the years – questions as to their relative importance or goodness must be left to a longer discussion. As the Semantic Web community works out a way to author, encode, interpret and manage machine-readable knowledge that is applicable across a range of systems and subject domains, we may be witnessing the emergence of a major new software platform.

Historically, software platforms have varied widely in terms of their impact on the evolution of computing, measured, for example, by the range of applications that run on them, and thus the number of lives they touch. To become an OS360 or Windows behemoth, a platform must not only embody the right mix of technologies at the right time, but must additionally serve the needs of three disparate constituencies:

- **Solve the real needs of customers.** For example, Windows established itself among PC manufacturers by allowing them great flexibility in components and bus designs, while largely insulating software publishers and corporate application developers from the complexity that resulted. Market share on new and installed PCs in turn attracted hundreds of independent software publishers. Customers care about well-supported software that is compatible with other systems already in place. Products should also reduce customers' operating costs, comply with standards, and have lots of utility (applications), largely supplied by the independent publishers and corporate applications developers.
- **Create opportunities for software publishing companies** and, to a lesser extent, software services firms. Independent software publishers decide whether to publish to the platform – without them you don't actually have a platform, just a tool. Publishers care about market size, primarily, as well as the talent pool, tools, support/training, long-term platform strategy, and fairness – a level playing field. Of course, the platform also creates opportunities for systems integrators and for software developers.
- **Win the hearts and minds of software development teams**. One dimension of the importance of a software platform is surely its acceptance by a large community of software professionals (not just programmers, but the whole team). Software developers, who mostly go where the money is, will additionally be attracted to platforms with minimal stupidity, powerful tools, language-independence, and a rational, if not elegant, design. They also like openness, so that they can build their own tools.

Whatever technology is involved, a software platform results from a complex social arrangement among these stakeholders.

---

* Barr and Tessler are strategy consultants for software companies, software startups, investors, and government planners through their firm, Aldo Ventures. Kresge is a knowledge engineer at Essential Logic and a specialist in the healthcare benefits knowledge domain.

While it may seem premature to inquire about the impact of the Semantic Web on the structure of the software industry, history has shown that the eventual outcome of the platform-formation process is often shaped early in the evolution of the platform. Breadth of vision, choice of partners, openness, and dumb luck are all major success factors.

Who Touches The Knowledge?

The Semantic Web involves a nexus of technologies: distributed computing, HTTP, XML, RDF, frame-like KR, agents, inference engines, and more. To see the direction matters must take to produce a new software platform, consider the following application space in U.S. corporate healthcare benefits (healthcare providers, insurers, corporate HR, and employees):

- Insured employees can inquire about their benefits and options via their company's portal, at work or at home.
- Corporate HR departments can do benefits planning, get bids from insurance carriers, and easily exchange real-time information with business partners, like enrollment and eligibility outsourcers, based on their unique, complex employee group structure.
- Insurers, service providers and insured patients can communicate about claims and bills.
- Healthcare providers can get approvals in real time. Physicians can know instantly which alternative medications are covered by a particular patient's insurance plan.

Until all the players' websites are semantically enhanced, the full value of the shared knowledge cannot be realized. And there are thousands of domains like healthcare benefits, where knowledge-enabled applications make sense. The huge number of different applications and knowledgebases that must be built and maintained by a wide variety of people, make this look to us like the makings of a major software platform: these software developers will want to share tools and infrastructure to "raise the level" of their knowledge-level programming.

In addition to tools for programmers, the effective use of machine-readable knowledge in an industry requires that a large number of knowledge editing tools be created, so that different people in different roles can author, test, share, audit, and maintain the knowledgebases in a natural way. (Out-of-date knowledge is worse than useless.) The platform and infrastructure must support the construction of these tools. It must also support the evolution of abstractions that tie the whole industry together. In healthcare, for example, the concept of "benefit" is associated with a set of services, a cost formula, a payment mechanism, a liability, and so on, depending on one's perspective.

Trilogy Corp., an early success story in the use of explicit knowledge to support e-commerce, uses proprietary knowledge modeling tools, and scores of Stanford graduates, to maintain a single, proprietary knowledgebase of electronic components. For the Semantic Web, thousands of companies will be maintaining similarly complex knowledgebases. The tools must support sharing of industry abstractions and knowledge maintenance by non-PhD's! Eventually, there will likely be several layers of tools – another indication that a software platform might emerge.

Of course, there is a lot of money to be made by companies who enable this next level of e-commerce, where applications can offer increased functionality by accessing companies' published knowledgebases. But the infrastructure that must be built to enable this kind of knowledge sharing may have implications beyond commerce: explicating what we think we all know is a first step toward coming to a common understanding.

The wireless environment differs from the wired environment by more than just the lack of wires. The aspect of mobility adds new constraints and different characteristics. The wireless devices themselves add physical constrains regarding input and output. The arduous methods of input dictate that wireless users enter a minimum amount of information. The small displays prevent users from browsing information, and suggest the need for more precise searches.

Based on aspects of mobility, wireless applications must also maintain the following characteristics: 1) simplicity 2) dynamism and 3) awareness. While the input/output constraints are device related, these three characteristics address how the mobile users themselves normally interface with the mobile devices.

In a general context, wireless users follow the 5-step RDCAP methodology to complete a task:

1. **R**ecognition of general task
2. **D**ecomposition of general tasks into sub-tasks and questions
3. **C**ompletion of sub-tasks and questions (recursively completed through RDCAP process)
4. **A**ggregation of results
5. **P**rovision of the solution

The RDCAP methodology realizes that wireless users search for information for more than just the information itself. The wireless users need the information to solve a more general task. The specific questions and services are aggregated to provide a **solution** for the general task. To address the physical and mobile characteristics of wireless devices, applications must provide solutions that follow the RDCAP methodology.

At this point, current Internet technologies prevent developers from producing applications that can provide the desired solutions because the Internet technologies require hard-coded links from solutions to services. This approach makes the solutions either 1) too general or 2) too specific. A "general solution" can provide all users a limited solution that only addresses part of the general task, and a "specific solution" addresses the entire general task for only a few users. Neither solution type effectively addresses the wireless environment.

The semantic web solves this dilemma by enabling developers to establish a general framework in which the specific links to sites are determined at runtime. Developers specify the desired **types** of sites and services, but not the specific sites and services. At runtime, an agent determines the most appropriate services based on user preferences, input, location, etc, and the agent delivers the user a personalized aggregation of services.

For example, consider a travel solution for a user who wants to travel from Pittsburgh to Sao Paulo. The user needs ground transport from his home to the Pittsburgh International Airport, a flight from Pittsburgh to Sao Paulo, a hotel in Sao Paulo, and transportation from Sao Paulo's airport to the user's hotel. To provide a complete solution without the semantic web, developers would need to hardcode links to ground transportation for every city in the world. This method is doomed to fail.

Instead, the Semantic Web solution specifies the framework needs flight, hotel, and transportation services (the meaning of these terms is clarified by the use of ontologies). At runtime, an agent determines the most appropriate services for this framework based on the solution environment. In the travel example, the agent finds and uses services that provide ground transportation in Sao Paulo without foreknowledge of the services.

In this manner, the semantic web wireless application personalizes each solution returned to the user while insulating the user from unnecessary complexities. Without the need to hard-code solutions, developers can create solution frameworks that provide the specificity needed by users and the generality needed for actual implementation. This framework addresses the entire

RDCAP process and provides a solution for the user's general task that minimizes user input and maximizes simplicity.

The PIA Project: Learning to Semantically Annotate Texts from an Ontology and XML-Instance Data

Nigel Collier
National Institute of Informatics (NII)
National Center of Sciences, 2-1-2 Hitotsubashi
Chiyoda-ku, Tokyo 101-8430, Japan
collier@nii.ac.jp

Koichi Takeuchi
National Institute of Informatics (NII)
National Center of Sciences, 2-1-2 Hitotsubashi
Chiyoda-ku, Tokyo 101-8430, Japan
koichi@nii.ac.jp

Keita Tsuji
National Institute of Informatics (NII)
National Center of Sciences, 2-1-2 Hitotsubashi
Chiyoda-ku, Tokyo 101-8430, Japan
keita@nii.ac.jp

ABSTRACT

The development of the XML and RDF(S) standards offer a positive environment for machine learning to enable the automatic XML-annotation of texts that can encourage the extension of Semantic Web applications. After reviewing the current limitations of information extraction technology, specifically its lack of portability to new domains, we introduce the PIA project for automatically XML-annotating domain-based texts using example XML texts and an ontology for supervised training.

1. INTRODUCTION

PIA aims to develop a domain and language portable information extraction (IE) system. In contrast to other Web-based technologies such as information retrieval (IR) which are characterized by strong portability, no such system as yet exists for IE. We consider that the main factors which have prevented this are: (1) A focus within the IE community on general news-based IE, exemplified by systems that resulted from the message understanding conferences (MUCs) [6], and, (2) Despite recent moves towards machine learning for low level IE tasks such as named entity recognition there is still a strong reliance on large lexical resources such as term lists, and an emphasis on hand-built rules and patterns. The problem we see with this direction is that it promotes the development of rather inflexible IE systems that cannot easily be ported to new domains without substantial efforts to customize the system with domain-specific knowledge resources, e.g. the collection of domain dictionaries, writing domain-specific rules etc. Perhaps the greatest problem is that since there is no prior understanding between the IE system developer and the domain knowledge provider about the encoding of the knowledge that will be used to train the IE system, there is no guarantee that the type of knowledge that the system needs will be available in the new domain.

While the MUCs have made great advances in promoting the formalisation of IE tasks and evaluation, the MUC-style of IE technology provides a relatively sterile semantic environment. Semantics is limited to the border between syntax and semantics that occurs at the word and term level but largely ignores higher-level relations between the classes themselves and class-relations are ‘forced’ into disjoint relations as far as possible. The markup of text, while conforming to SGML, makes no use of an explicit ontology and relatively little use of meta-data.

Recent IE projects have looked beyond news to the molecular biology domain, e.g. [4] [5]. Some projects have implicitly incorporated simple taxonomies (is-a hierarchy) into the annotation guidelines for domain experts. To the best of our knowledge these projects still largely ignore explicit properties of classes and class relations that could be contained in the ontology and their potential contribution to automatic annotation. There seems to be great potential in this technical domain for incorporating ontologies into the learning model since a large amount of research has taken place on their development for gene-product databases such as SwissProt [1].

We believe then that with the advent of standards for the annotation of semantic content such as XML [3] for document structure, RDF [7] for defining objects and their relations, and RDFS [8] for defining the object model for describing RDF, that sources of domain knowledge will become widely available in electronic form and that these resources can and should be used for supervised training of a portable IE system which we call PIA-Core. Crucially these sources of knowledge will be available in a predictable format allowing PIA-Core to be rapidly deployed in a new domain.

2. PIA-CORE

We consider the W3C standardization process of XML and RDF(S) to offer a positive environment for machine learning of expert knowledge. Although ontologies in RDF are likely to emerge primarily as a result of (human) expert introspection we cannot expect that XML-instances of the defined concepts, such as technical terms, proper nouns, quantity and time expressions and their relations, will be annotated by experts for every document due to the high

Semantic Web Workshop (SWWS) 2001 Stanford University, USA

cost. This is one of the bottlenecks in the extension of Semantic Web [2] applications to the majority of documents that can be viewed on the Web and Intranets today. What is missing in the current focus on formalisation is a consideration about how the actual instantiation of the concepts defined in the ontologies will take place.

Actual semantic annotation of terminology and relations involves considerable time by domain experts and for this reason we believe it is worth investing in machine learning as a way to reliably replicate the capabilities of experts. This is the goal of PIA-Core. The scenario is that experts will develop a domain model in RDFS and a relatively small set of example annotated texts using an integrated XML and ontology editor. From this knowledge PIA will learn how to automatically XML-annotate unseen texts in the same domain. By focussing on domain-based learning we hope to make use of the ontology as a valuable knowledge resource and also to reduce the problems of ambiguity that developers of general IE systems must face.

In combination with robust domain-independent natural language processing (NLP) modules such as part of speech taggers, chunkers and shallow parsers, as well as general linguistic resources such as thesauri, PIA-Core will be used to XML-annotate texts that are consistent with those in the training set. We hope that PIA-Core can provide rapid acquisition of domain-knowledge and provide functionality that can be used at the heart of an IE system or within XML-tagging tools for computer-aided annotation. This can then serve as the basis for the deployment of 'smart' applications providing intelligent services that we hope to see emerge on the 2nd generation World Wide Web (Web), i.e. the 'Semantic Web'. The application we want to apply PIA-Core to is domain-based question-answering (e.g. [9]) in English and Japanese.

3. DISCUSSION

Some of the key questions that we need to consider are:

- How should we integrate the ontology into a statistically-based machine learning (ML) model? For example, how should we make use of concepts that appear in the ontology but don't appear in the training set? How can statistical evidence from sub-classes be shared through ontological relations to help overcome data-sparseness problems?

- XML is in some respects quite limited in its ability to represent complex object structure and relations as it is designed to encode serialization. We need to explore the limits of this representation for practical annotation of terminology and relations.

- Ontologies change over time - they may be revised, expanded or incorporated into other (shared) ontologies. How do we update the knowledge base that was extracted from the training set based on the original ontology? Should the model be retrained every time a change is made to the ontology? How can changes in the ontology be reflected automatically at the text markup level? E.g. the introduction of a new subclass.

- How will the issue of multi-linguality affect the design of ontologies in RDF for PIA?

4. CONCLUSION

We briefly presented a critical analysis of the current status of IE research and proposed a new project called PIA based on domain-based learning through XML-annotated texts and domain models described in RDF. We also considered some of the key research issues. From now we intend to implement PIA-Core and apply it to the task of question answering in technical domains such as molecular-biology.

5. ADDITIONAL AUTHORS

Gareth Jones (Department of Computer Science, University of Exeter, UK, email: `gareth@dcs.exeter.ac.uk`), Jun'ichi Fukumoto (Department of Computer Science, Ritsumeikan University, Japan, email: `fukumoto@cs.ritsumei.ac.jp`), Norihiro Ogata (Faculty of Language and Culture, Osaka University, Japan, email: `ogata@lang.osaka-u.ac.jp`), Chikashi Nobata (CRL, Japan, email: `nova@crl.go.jp`), Kinji Ono (National Institute of Informatics, Japan, email: `ono@nii.ac.jp`).

6. REFERENCES

[1] A. Bairoch and R. Apweiler. The SWISS-PROT protein sequence data bank and its new supplement TrEMBL. *Nucleic Acids Research*, 25:31–36, 1997.

[2] T. Berners-Lee. *Weaving the Web*. Harper, San Francisco, 1999.

[3] Namespaces in xml, world wide web consortium recommendation. http://www.w3.org/xml/TR/REC-XML, 14th January 1998.

[4] N. Collier, H. Park, N. Ogata, Y. Tateishi, C. Nobata, T. Ohta, T. Sekimizu, H. Imai, and J. Tsujii. The GENIA project: corpus-based knowledge acquisition and information extraction from genome research papers. In *Proceedings of the Annual Meeting of the European chapter of the Association for Computational Linguistics (EACL'99)*, June 1999.

[5] M. Craven and J. Kumlien. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the 7th International Conference on Intelligent Systemps for Molecular Biology (ISMB-99)*, Heidelburg, Germany, August 6–10 1999.

[6] DARPA. *Proceedings of the Sixth Message Understanding Conference(MUC-6)*, Columbia, MD, USA, November 1995. Morgan Kaufmann.

[7] Resource description framework (rdf) model and syntax specification, world wide web consortium recommendation. http://www.w3.org/xml/TR/REC-rdf-syntax, 22nd February 1999.

[8] Resource description framework (rdf) schema specification 1.0, w3c candidate recommendation. http://www.w3.org/xml/TR/2000/CR-rdf-schema-20000327, 27th March 2000.

[9] R. Srihari and W. Li. Information extraction supported question answering. In *Proceedings of the Eighth Text REtrieval Conference (TREC-8), Gaithersburg, Maryland, USA*. National Institute of Standards and Technology (NIST), November 17–19 1999.

Delivering Configurable Problem-Solving Services to Web Users

Monica Crubézy and Mark A. Musen

Stanford Medical Informatics
MSOB X-215, 251 Campus Drive
Stanford, CA 94305-5479 USA
+1 650 723-6979
{crubezy,musen}@smi.stanford.edu

The World-Wide Web is suitable not only for distributing static data resources but also for providing users with dynamic services. Currently, a growing number of online services are appearing, offering functionality such as booking a performance ticket, organizing a meeting, or comparing experimental genomic data against reference databases of gene sequences. Existing services, however, tend to be integral: The underlying reasoning process that these services perform is neither modifiable, nor configurable for different domains, nor can it be integrated with other services to produce new functionalities. As a result, the level of functionality that users can expect from Web services is limited to performing simple, predefined tasks. Such tasks typically do not handle large amounts of knowledge. Meanwhile, the current evolution of the Web into a Semantic Web encourages communities of users to create and publish shared domain knowledge conceptualizations—ontologies—with which they describe the data that they exchange. We anticipate that user communities soon also will need to exchange the processes and services that operate on their shared data and knowledge to solve complex tasks in their domain. For instance, developers and users of image-processing techniques already try to share either software modules or experience with specific tools. In the context of the Semantic Web, the goal of our research is to develop the technologies and frameworks needed to provide sophisticated online reasoning facilities, configurable for different domains and applications. For instance, we can imagine making available online a generic heuristic classification technology, which different communities (e.g., paleontologists, geologists, biologists) or individual users can adapt for specific applications. We believe that the Semantic Web will provide the grounding infrastructure to deliver *intelligent problem-solving services* to users willing to achieve knowledge intensive tasks in their domain.

At Stanford Medical Informatics[1], we have been developing knowledge systems for different application domains and purposes for many years. Our approach consists in building intelligent systems by assembling reusable knowledge components, namely *domain ontologies* and *problem-solving methods*. Problem-solving methods are domain-independent problem-solving strategies, that provide standard ways of addressing stereotypical problems, or *generic tasks*, such as diagnosis, design, and classification. Our major contribution in the area of knowledge engineering is incarnated by the Protégé[2] suite of tools that supports developers in building a running application system by creating, configuring and assembling reusable components. Protégé offers a generic, full-fledged knowledge-modeling and user-interface support, extensible and customizable for different purposes and users. In particular, Protégé supports the process of selecting a problem-solving method suitable for the user problem, and configuring the method by integrating its generic components with domain-specific information through the use of explicit architectural elements—declarative *mapping relations*.

Delivering problem-solving services to Web users involves developing the necessary infrastructure and technologies that can enable service identification and interoperability on the Web. In

[1] http://www.smi.stanford.edu/

[2] http://protege.stanford.edu/

particular, the competence of each Web service needs to be specified and advertized so that brokering agents can then reason about these competence descriptions to match services to user needs and to combine services together. Ongoing work in the IBROW project[3] is addressing these issues and aims to produce a brokering agent that can locate and configure problem-solving services on the Web to solve specific tasks requested by users. In particular, the IBROW consortium has developed the Uniform Problem-solving Method development Language (UPML[4])—a modeling framework and a markup language (in the form of an RDF Schema[5]) for specifying problem-solving components on the Web. UPML provides a comprehensive ontology to specify the properties of each kind of component involved in a knowledge system. For instance, problem-solving methods define a set of input and output roles, as well as set of logical formulas, which provide a functional specification of their competence. UPML also provides explicit ways of specifying the assembly of components in a knowledge system. In the context of IBROW, we have developed a special-purpose editor, based on Protégé, that enable developers to model and advertize problem-solving services on the Web, as resources marked-up in UPML.

Jointly with the Knowledge Media Institute[6] at the Open University (UK), we also have developed the IBROW Internet Reasoning Service (IRS), a tool that enables developers to prototype knowledge systems quickly by configuring reusable reasoning components from online structured libraries. The process of configuring problem-solving components into a running application may involve several activities: mapping generic tasks and methods to a domain model (e.g., mapping a generic classification framework to a database of archeological artifacts to produce an artifact-classification application), mapping methods to tasks (e.g., selecting a particular abstraction method for performing a data-abstraction task), or, in general, refining existing components (e.g., specializing a data-to-solutions matching component by introducing fuzziness in the matching process). Based on the knowledge-level descriptions of the problem-solving components in UPML, the IRS provides different levels of user support, from interactive browsing and navigation facilities, which enable manual selection and configuration of reasoning components, to intelligent, semi-automated assistance in building an executable application. We have implemented a version of the IRS as an extension to Protégé, that interfaces domain knowledge bases to UPML-compliant libraries of problem-solving components.

The ultimate aim of our work is to make sophisticated problem-solving technology available to a wider audience and provide the level of intelligent support needed to allow rapid generation of web-based reasoning services. A side-effect of achieving this goal will be simply to make much artificial intelligence technology available online, thus making its use more widespread. The Semantic Web is a vehicle that will allow us to perform a large evaluation experiment of our approach, by delivering problem-solving services directly to users. At the same time, the heterogeneous, distributed and versatile nature of the Web will challenge us to incorporate more automated support, with simplified procedures and customized interactions, so that less-experienced users also can benefit from advanced distributed problem-solving services. We anticipate a future where the Internet will make libraries of problem solvers available to any person with a Web browser. The challenge is to describe and index these problem solvers meaningfully, and to develop the infrastructure that allows users to locate these problem solvers easily and to link them to specific data sets and knowledge bases in simple, intuitive ways. Our research is exactly a step towards a Semantic Web in which users can access not only knowledge but also intelligent problem solvers.

---

[3] http://www.swi.psy.uva.nl/projects/ibrow/home.html

[4] http://www.cs.vu.nl/~upml/

[5] http://www.w3.org/TR/rdf-schema/

[6] http://kmi.open.ac.uk/

Semantic Web Workshop

Karl D'Adamo

Since I got my first UNIX account on the undergraduate-operated computer cluster at Caltech, I have been exposed to computer science, the web and the limitless possibilities of the future of computers and information. Exposure to the historical adoption of standards in computer science and the web whetted my interest, but I remained a user of the technology in stead of a creator of new technologies. Many interests and projects in my life relate to the Semantic Web. It is difficult to pare this down to two pages, but I will try.

I have been working at Applied Materials since I graduated from college (2 years). I was educated as a chemical engineer at Caltech, and my education included little practical engineering, and a lot of mathematics and computer science. I realized very quickly that I could never be happy being a chemical engineer when what had excited me in school was not the results of the solutions to the differential equations of fluid, mass and heat transfer, but the solutions themselves. In stead of feeling cheated at not receiving the engineer they thought they were getting, my employer was happy to accomodate my interests. Applied Materials makes semiconductor processing equipment, but the division that I work in is responsible for offering service to our customers and empowering the 10,000 technicians that fix our often broken tools with the information that they need to do their jobs. I am also very lucky to be in a position where I get to make recommendations and decisions about what kinds of IT solutions we should use to maximize the productivity of our employees.

We, as a company, have come to the realization that the delivery of information to our technicans in the field is vital to improving their productivity. Our industry presents perhaps a worst case scenario because our technicians need access to gigabytes of information (installation and repair manuals, up-to-date best known methods), our customers span hundreds of countries, requiring documentation and information be available in dozens of languages, and most of the people in the field require wireless access to our networks since they work in customer fabs which do not allow their computers to connect to the Applied Materials network. As a first step to solve this problem, we are attempting a redesign of how our documents are written and distributed. Authoring will require extensive use of meta-tagging, and retrieval will require a more intelligent system of searching and delivery. The next step would be intelligent logic that push the information to the end users based on their roles, responsibilities and interests. The first step has begun; the second step is being planned. To implement and design the solutions, we require an understanding of the basic ideas and technology available.

A second major area I am involved with is the optimization and automation of our customers' fabs. If there is one thing that I have come to realize recently, it is that operating a $2B fab and ensuring costs are kept down while yield is maximized is a dauntingly complex task. All aspects of operations, from material monitoring to resource planning to failure escalation require tremendous steps forward in terms of the data that is collected as well as in terms of how that data is organized and analyzed. We are in a position to define standards in terms of data collection and analysis, but at Applied Materials, as elsewhere, we are still unsure of how to proceed.

In addition to the challenges that I face at work, there are other areas of my life that relate to the Semantic Web. Since I first learned of predicate logic and its power, I have been interested in its limitations as well. Understanding Godel's Incompleteness theorem brought a new clarity to logic as well as a new set of questions to my mind. I have struggled constantly with attempts to formalize my thought. Although my lack of success can probably be attributed to my own mind, I think that RDF is an imensely interesting framework to do further analysis of predicate logic, its power and its limitations. I would love to be one of the people who finds out what happens when we use RDF and the Semantic Web define the how the predicate of "predication" can be applied to resources.

In addition to these high-minded applications of the Semantic Web, I have become somewhat of an information junkis in my 7 or so years of using the web. The practical knowlege and intuition that I have built up has been invaluable in sorting through the junk that is out there in order to get at the information that I want. Once the underlying logic has been established, I can only imagine the possibilities that become reality.

I am also keenly aware and interested in the law and how it applies to the internet. I realize that while all information should be available to everyone, it is important to have patent-type protection on the organization of information. It should be legal to copy a white pages and resell it. But I think there need to be laws protecting the novel and unique organization of information. The Semantic Web offers to create a much more powerful framework for the organization and manipulation of information, and I think it will force people to rethink these laws.

I don't want to sort through 100 webpages to find what I want anymore. I am tired of writing python scripts to search webpages to find and extract structured information. Most of all, I am in awe of the possibilities that the Semantic Web will provide, and I want to learn more, and help the effort.

Extension of RDF(S) with Contextual and Definitional Knowledge

Alexandre Delteil, Catherine Faron-Zucker, Rose Dieng
INRIA, ACACIA Project, 2004 route des Lucioles, BP 93,06902 Sophia Antipolis, France
{Alexandre.Delteil,Catherine.Faron,Rose.Dieng}@sophia.inria.fr

1 Introduction

The need of a Semantic Web is now well recognized and always more emphasized. The huge amount of information available on the Web has become overwhelming, and knowledge based reasoning now is the key to lead the Web to its full potential. In the last few years, a new generation of knowledge based search engines has arisen which rely on extensions of HTML to annotate Web documents with semantic metadata, thus enabling semantic content guided search. For interoperability on the Web, the importance of widely accepted standards is emphasized. RDF is the emerging standard proposed by the W3C for the representation and exchange of metadata on the Semantic Web. RDF Schema is the standard dedicated to the representation of ontological knowledge used in RDF statements.

ACACIA is involved in the CoMMA European IST project, dedicated to corporate memory management through agents. The memory is materialized by the electronic documents of the organization which are described by RDF annotations. These are the key for knowledge based information retrieval on the Intranet by using the inference engine *CORESE* implemented in our team (Corby *et al.*, 2000). However the expressivity of RDF(S) appears too much limited to represent the ontological knowledge of the corporate memory. Inference rules representing domain axioms, class and property definitions are crucial for intelligent information retrieval on the Web. The need for axiomatic knowledge is well-known since the first information retrieval systems on the Semantic Web. It is the key to discover implicit knowledge in Web page annotations so that information retrieval be independent of the point of view adopted when annotating.

When compared to object-oriented knowledge representation languages, description logics, or conceptual graphs, RDF(S) does not enable to declare class, property and axiom definitions. We have specified an extension of RDF(S) with class, property and axiom definitions based on the similarity between the RDF and the Conceptual Graphs models. We call it DRDF(S) for *Defined Resource Description Framework* (Delteil *et al.*, 2001a). Other extensions of RDFS have been proposed, such as DAML and OIL. DAML provides useful primitives for declaring intersection, disjunction, complementary of classes, … . OIL stems from a description logic; as DRDFS, it provides a way of expressing class and property definitions. However DRDFS and OIL have two incomparable expressivities, in the sense that none can be considered as a fragment of the other.

What DRDFS provides also is the ability to express contextual knowledge on the Web (Delteil *et al.*, 2001b). The RDF philosophy consists in letting anybody free to declare anything about any resource. Therefore the knowledge of who and in which context a certain annotation has been stated is crucial. DRDF(S) enables to assign a context to any set of annotations. We hope that DRDF(S) will contribute to the ongoing work of the W3C committee for improving RDFS.

2 RDF(S) and its Limitations

RDF is a data model provided with an XML syntax. RDF knowledge is positive, conjunctive and existential. A set of statements is viewed as a directed labeled graph: a vertex is either a resource or a literal; an arc between two vertices is labeled by a binary property. RDFS is dedicated to the specification of schemas representing the ontological knowledge used in RDF statements. A schema consists in declarations of classes and properties. It is defined by refining the core RDFS: domain specific classes and properties are declared as instances of the *Class* and *Property* resources; the *subClassOf* and *subPropertyOf* relations enable the representation of class and property hierarchies.

2.1 A Triple Model.

The RDF data model is a triple model: an RDF statement is a triple (resource, property, value). When asserted, RDF triples are clustered inside annotations. An annotation can thus be viewed as a graph, subgraph of the great RDF graph representing the whole set of annotations on the Web. However, there is no distinction between the statements made in a single sentence and the statements made in separate sentences. Let us consider two different annotations relative to two different research projects which the employee 46 of T-Nova participates to:

- {(employee-46, worksIn, T-Nova), (employee-46, project, CoMMA), (employee-46, activity, endUser)} and
- {(employee-46, worksIn, T-Nova), (employee-46, project, projectXX), (employee-46, activity, developer)}.

The whole RDF graph does not distinguish between these two clusters of statements. Employee 46 is both endUser and developer: the knowledge of which activity he is implicated in inside of a project is lost.

DRDF(S) enable to represent independent clusters of RDF statements through the *context* feature.

2.2 RDF Reification.

The RDF model is provided with a reification mechanism dedicated to higher order statements about other statements. A statement (r, p, v) is reified into a resource s described by the four following properties: the *subject* property identifies the resource r, the *predicate* property identifies the original property p, the *object* property identifies the property value v, the *type* property describes the type of s; s is an instance of *rdf:Statement*.
However, the reification of a set of statements requires the use of a container to refer to the collection of the resources reifying these statements. This leads to quite complicate graphs. Moreover a statement containing an anonymous resource can not always be reified: the values of the properties *subject* and *object* must have an identifier. The notion of context we introduce in DRDF(S) enable to reify a set of statements much more easily.

2.3 Existential quantification through anonymous resources.

The RDF model focuses on the description of identified resources but allows a limited form of existential quantification through the anonymous resource feature. It is handled by RDF parsers by automatically generating an ID for the anonymous resource. However, it is a limited solution and a graph containing a cycle with more than one anonymous resource can not be represented in the XML syntax of RDF. DRDFS enable to represent every existential, positive and conjunctive statement, without any restriction.

The roots of DRDF(S) stand in the correspondence between RDF(S) and the conceptual graph (CG) model (Sowa, 2001). The CG model provides a direct way of expressing independent pieces of knowledge through graphs. It thus enables the representation of contexts for various applications (quotations, viewpoint, …). CGs are particularly useful as definitional contexts enabling the definition of concepts or axioms (Delteil *et al.*, 2001a). An in-depth comparison of both models is studied in (Corby *et al.*, 2000).

3 Extending RDFS with Contexts, Existential Quantification and Coreference

In DRDF(S), a resource of type *Context* expresses the clustering of statements - much more easily than an RDF container. A context identifies a sub-graph of the whole RDF graph. A context is defined from a resource G of type *Context* as the largest subgraph of the whole RDF graph whose all internal nodes excepted G are anonymous resources $c\varnothing_i$. A context is an abstraction that enables to talk about representations of resources (through anonymous resources) rather than directly about resources. Anonymous resources are "externally identified" by the *referent* property.
DRDF(S) is provided with a general mechanism for existential quantification handling. It is represented by an anonymous resource described by a *referent* property whose value is an instance of *Variable*, a new RDF class we introduce. The scope of a variable is the context it belongs to, just like in FOL, where the scope of a variable is the formula it belongs to. We introduce a *parameter* property to link a resource of type *Variable* to a resource of type *Context*.

4 Special Contexts: Axioms, Class and Property Definitions

This general feature of context can be used for representing axioms, and class and property definitions. DRDF(S) class and property definitions are descended from type definitions in the CG model; DRDF(S) axioms are descended from Conceptual Graph rules. A class definition is a monadic abstraction, i.e. a context whose one resource of type Variable is considered as formal parameter; a property definition is a diadic abstraction. An axiom is a couple of lambda abstractions, i.e. two contexts representing the hypothesis and the conclusion.

5 References

Corby, O., Dieng, R., Hebert, C. (2000). A Conceptual Graph Model for W3C Resource Description Framework. ICCS'00, LNCS 1867, 2000.
Delteil, A., Faron, C., Dieng, R. (2001a). Extensions of RDFS Based on the Conceptual Graph Model. ICCS.
Delteil, A., Faron, C., Dieng, R. (2001b). Extensions of RDFS with Contextual and Definitional Knowledge. WebNet.
Sowa, J.F. (2001). Conceptual Graphs: DpANS. http://www.bestweb.net/%7Esowa/cg/cgstand.htm, 2001.

Implications of Semantic Web Technology for Wireless Handheld Computing

Charles Earl

My interest in the semantic web is as a framework for facilitating exchange of data between wireless handheld computing devices and web-connected information repositories. In my view, it is important that the notion of "web" be broadly construed to include both the larger World Wide Web and local, "emergent" webs that ebb and flow in an addhoc fashion. That is, a web that exists for facilitating the exchange of information between handheld devices at a conference or concert – a web limited in time and space – is nevertheless the kind of web that will be of considerable important to users of handheld devices. In this paper, I will outline some of the challenges faced connecting handheld devices to webs, discuss some roles for semantic web technology in addressing these challenges, and discuss some work to be done.

The term "handheld computer" or "personal digital assistant" (pda) typically refers to PalmOS devices (e.g. Palm V, Sony Clie, etc.), PocketPC devices (e.g. Compaq ipaq), and RIMM Blackberry pager. This category of device is ever broadening to include so-called "smartphones" (e.g. Nokia Communicator,), and eBooks. These devices – my interest is in those with wireless connectivity – are thus being developed in a variety of form factors (from eBooks to wristwatch size devices) and are being developed with variety of targeted functionality (e.g. a Samsung smartphone supports web-browsing and MP3 playing). With a limited form factor comes the problem of presenting the range of web data on such devices. The proliferation of special function devices is another argument for on the fly content adaptation. For example, a document having multimedia content should be delivered differently to a smartphone than it would be to an eBook.

Semantic web technology has already been used -- albeit in a limited sense -- to address the "multiple device delivery" problem. Work on image and document transcoding for example, is relatively mature. Given a "desktop-viewable" document sufficiently annotated with semantic tags (e.g. DAML+OIL) it is possible to transform that document (e.g. assuming appropriate style sheets, or transformation rules) into a form viewable on a limited screen handheld device. A significant challenge remaining in this area is that of adoption: how do we facilitate the transition of existing repositories over to machine readable form; how to we enable users to develop new content using ontologies and other tools developed in support of the semantic web. Tools such as Annotea (http://www.w3.org/2001/Annotea/) and web data migration services such as those provided by 2Roam (http://www.2roam.com/) are a beginning.

The trends of increased processing and storage resources, and increased wireless connectivity options for handheld devices drive another important use for semantic web technology. Although the processing power and storage capacity of these devices lag those of desktop machines by an order of magnitude, the resources now available on high end handheld devices have scaled to support storage and manipulation of significant amounts of heterogeneous data. A number of wireless communication technologies have been adopted in these devices, making it possible to connect them to the larger web (e.g. a PalmOS device using the Omnisky service) or to add-hoc local networks (e.g. several PalmOS devices connected via Bluetooth radios and infrastructure). In either of these networking scenarios, the challenge of limited bandwidth and limited connectivity still remain. Most importantly, users of such devices will inevitably what to exchange data among each other.

For example, imagine a scenario in which a wireless local network has been setup to run during the course of a music concert. Assuming for a moment that the legal issues have been addressed, how is the exchange of information – anything from MP3s to videos to addresses -- among concert-goers using these devices – some having smartphones, some having Palm devices with 802.11b cards, others with different configurations – to be achieved? I think that semantic web technology can figure into the solutions.

I believe that one of the most important and interesting uses of semantic web technology for wireless handhelds will be in facilitating community in public spaces (e.g. libraries, concerts, sporting events.) How can users of such devices search for and make available large amounts and varieties of data in ways that construct and enhance communities of shared interest?

This raises three large problems for designers of such systems:

- Search: If I am a visitor to such a place, how do I locate information that I'm interested in? How do I locate people with similar interests?
- Broadcast: How do I make the information available on my device available to the right people in the local area? How do I control who can see it?
- Indexing: How do I maintain the information on my device so that I can make it intelligible, useful to others interested in the information?

To understand the tools necessary to support this functionality, I'll make the simplifying assumption that participants in a shared community will make use of ontologies developed by and for members of that community. Making this assumption, the following technologies still need to be developed:

- Adaptable search engines for semantic webs. What are the characteristics of search engine that would run in a library and locate everyone having present in the library having similar research interests? or having a similar bibliographies stored on their devices. If we assume that the information on local information repositories follows consistent semantic organization, then the search task becomes easier. Efficiency tradeoffs between peer-to-peer and centralized search must also be explored.

- Tools for ontology construction. Assume that the users of such devices belong to communities that want to share information among one another. What are tools that allow them to construct ontologies useful for the members of that community.
- Indexing tools. Tools that allow information to contained on such devices to be easily indexed according to shared ontologies. How can index construction algorithms scale to the capabilities of such devices?

As these problems are addressed, I'm sure that new technologies and synergies will emerge.

Applications of the Semantic Web for document retrieval

Andreas Eberhart
International University in Germany
D-76646 Bruchsal
E-mail: eberhart@i-u.de

1. Introduction

Finding the right information in the Internet or corporate Intranet remains one of the biggest problems in our digital everyday world. Most full-text search engines offer the user the ability to combine search terms with Boolean expressions and to limit the search space. A popular augmentation of this approach is to categorize the area of interest à la Yahoo. This definitely yields a high precision of queries like: give me a list of all Canadian universities, but obviously, due to the organizational and administrative overhead involved, the categories are very coarse grained. Furthermore, one has to make a lot of compromises when designing such a taxonomy of categories and it will definitely not fit everybody's view.

We also face these problems in the area of online learning. The idea here is to define a fixed set of metadata for the documents stored in the document repository. The Learning Object Metadata standard, for instance, suggests storing keywords, the technical document format, difficulty level, etc. for each unit. This information is very useful for the management of content, however, when used for a search interface, too many search parameters confuse the user.

We observe some fundamental problems: Information retrieval must have a more personalized character with the system being able to leverage context information. A search for a JDBC tutorial should be treated differently if it is issued by a software professional or by an undergraduate student. The interface into the system should be flexible and not only allow a fixed set of attributes to look for. Finally, the system should not be restricted to only one source of information in order to avoid the problems mentioned with a Yahoo-like approach.

2. Points of interest

From keywords and taxonomies to ontologies

We think that there are two aspects of using keywords and taxonomies that are noteworthy. The first point is that it is crucial that every user of a metadata-based content management system shares the same interpretation of the taxonomy terms and the keywords. This is a noteworthy point even if it is implicitly clear to us, since natural language is used. Secondly, a content management system could be viewed as an application that uses a very simple ontology about categories, subcategories, keywords, documents, etc. Document management and retrieval systems implicitly share this ontology and implement it in the application logic..

It is quite clear that a knowledge supported retrieval system is only useful if it bases on a large ontology and if it has access to a large base of tagged content. We believe that this ontology does not have to provide a deep understanding of the domain being searched. After all we want to build a smart librarian that retrieves a document with the answer, not an expert that produces an answer. The "shallow" nature of the ontology should enable the integration of information from several external sources. For instance, an ontology on Java by Sun Microsystems could be used by an online learning system on distributed applications.

From our point of view, it is better to sacrifice some level of detail and expressiveness in ontologies if it makes them easier to integrate and more useable, rather than having islands of complex ontologies for isolated applications. The lessons we will learn from integrating simple ontologies will then also be the foundation for more complex endeavors.

Agent scenario

Today's search engines work in a brute force fashion. Let us look at how we obtain information in our daily lives: an important aspect could be characterized under the term "ask the expert". We know that Jim is the database guru in our company, therefore he could probably point us to a good tutorial on JDBC. What is important here is that Jim also knows me, thus he knows which level of difficulty would be appropriate. If someone encounters a software setup problem while working on a term project, the right person to ask would probably be an experienced computer user who is taking the same class. Chances are that this person has already encountered and solved the same problem. We would probably also find the relevant information on the Internet, but in such a personalized environment, the search precision is much higher. We believe that relevant context information such as the user's background and experience can be exchanged in a flexible way using semantic web techniques.

People usually engage in a conversation where the experts tries to find out more. A user saying: I am having trouble setting up software X might prompt the expert's question: which operating system are you working on? Finally this interaction is ended with the expert providing an answer, pointing to a document or another expert, or saying: I don't know. Obviously these processes are very complex and several quite fuzzy heuristics are involved in every step. However, we feel that even a partial mapping of this model to a search agent system will solve many of the shortcomings of traditional systems.

Integration of information services

Another interesting point in this agent based information retrieval example is the integration of traditional information systems. The information which student is enrolled in which course is probably stored in a university ERP system. Data from the ERP system can be highly relevant when students search for classes. Work done in the area of semantic description of web services will be a valuable foundation for our application example.

POSITION STATEMENT
Semantic Web Workshop (SWWS)

Update Semantics for Cooperative Ontologies

Peter Eklund, Peter Becker and Philippe Martin

DSTC Pty Ltd
Griffith University
PMB 50 Gold Coast MC QLD, 9726, Australia
peklund@dstc.com

An assumption of the *Semantic Web* is that knowledge producers will generate knowledge (as meta-data or content descriptors) that can be automatically compared. A domain ontology system must aim at helping knowledge consumers and producers to use unambiguous descriptors. For example, when I use the term "switch", a domain ontology system should know about the various meanings of "switch": (i) a mechanical, electrical device; (ii) a flexible instrument for punishment; (iii) a substitute (iv) a basketball maneuver, etc. The domain ontology server intervenes to help refine the term until it is classified to a category with its intended meaning. The category can then be compared with other identical categories of "switch" used as meta-data descriptors associated with documents.

Often when terms are used in context, they can be disambiguated automatically. "Switch on a wall" is enough context to discount a number of meanings of switch since the signature of the spatial relation "on" provides a restriction on the category type[1]. Furthermore, a domain ontology system needs facilities for individuals or teams to work together, to create and refine categories, maintain meaningful views and give access security of the categories that they own. The number of categories for "switch" (and their intended meanings) for a company (or industry) that manufactures switches, will need to complement the more general meanings given earlier. For this reason a mechanism to describe categories and their place in a type hierarchy is necessary, as are filters to accommodate multiple views.

To experiment with a ontology that is meaningful and large enough to simulate the difficulties of an ontology server for the World Wide Web we knowledge engineered the natural language ontology WordNet and our own top-level ontology into a object-relational database called FASTDB[2]. We call the resulting ontology server and inference system WEBKB-2[4, 5]. Our ontology contains 94,500 nouns, 66,000 categories referred to by nouns, 21,000 adjectives and 7,900 categories referred to by adjectives. WEBKB-2 is an ontology server but also a inference engine in the classic knowledge base sense. It can also be used to store and retrieve conceptual graphs [6, 7].

[1] In this case the presence of "on" excludes the last two meanings of switch (given above).
[2] http://www.ispras.ru/ knizhnik/fastdb.html

Graph matching is also permitted. The graphs shown below are an interpretation of the original graph placed into the knowledge base. The original graph was of the form,

```
[philippe.martin@gu.edu.au,
  agent of: (the renting,
     object: (an apartment,
              part: 1 bedroom,
              location: Southport),
     instrument: 140 Australian_dollars,
     period: a calendar_week,
     beneficiary: pm#Spirit_Of_Finance)](pm);
```

but would be retrieved by the query "*?[a renting]*" as,

```
[philippe.martin@gu.edu.au,
   pm#agent of: (some #renting,
     pm#object: (some #apartment,
        pm#part: 1 #bedroom,
        pm#location: QLD#Southport),
     pm#instrument: 140 #Australian_dollar,
     #time_period: some #calendar_week,
     pm#beneficiary: pm#Spirit_Of_Finance)]];
```

here we observe that each of the terms in the initial graph have been unambiguously resolved to categories by WEBKB-2. This interpretation of terms as categories is indicative of the ontology domain system service.

Our interest is to experiment with engineering of semantic web-style applications using this ontology server and inference engine and furthermore researching the practical and theoretical difficulty involved in an update semantics for cooperative ontologies. We are also interested in re-engineering our ontology server to interoperate with XML Schema, RDFS and DAML+OIL. WEBKB can be found at http://www.webkb.org

References

[4] P. Martin and P. Eklund *Embedding Knowledge in Web Documents.* Proceedings of WWW8, 8th International World Wide Web Conference, pp. 324–341, Elsevier, 1999.

[5] P. Martin and P. Eklund *Knowledge Indexation and Retrieval and the Word Wide Web, IEEE Intelligent Systems,* July, pp. 18-25, 2000.

[6] P. Martin "Conventions and Notations for Knowledge representation and Retrieval" *Proceedings of the 8th International Conference on Conceptual Graphs*, LNAI 1867, pp. 41–54, Springer-Verlag, 2000.

[7] J.F. Sowa: Conceptual Structures: Information Processing in Mind and Machine. Addison-Wesley, 1984.

Peer-to-Peer Infrastructure Supporting the Semantic Web

Johannes Ernst
R-Objects Inc.

1 CONTEXT AND MOTIVATION

Peer-to-peer technologies have recently garnered a lot of attention. Currently, most broadly adopted peer-to-peer initiatives focus on file sharing (Napster, Gnutella, many others), distributed computation (Seti@Home, Porivo, etc.), or collaboration (Groove Networks, Aimster, etc.). However, it has been argued [2][3] that the killer benefit of peer-to-peer computing will be the ability of individual users to gain back local control from centrally maintained websites, and innovate locally in spite of taking advantage of the network.

On the other side, decentralized control and innovation is what the Semantic Web needs to realize much of its potential. A "walled garden" – which is the best way to describe many web sites today – will not be able to provide the types of interoperability benefits that visionaries such as Berners-Lee foresee if/when the Semantic Web gets broadly adopted [1].

In fact, the need for decentralization is even more important for the Semantic Web than for Today's Web: not only do content authors need the ability to link from their content to another site's content, as they do on the web, but in addition:

- it needs to be easy for innovative projects/companies to link new "semantics definitions" to existing content on the web, even if the authors of the content and the meta-information are not the same; and it needs to be easy to develop software agents that use the new meta-information with existing information. If this wasn't possible, lacking a business case, the vast majority of Today's Web would remain without semantics for a long time; further, where information was indeed published with semantics, those "semantics definitions" would be essentially frozen, disallowing a lot of innovation, such as the ability by a third party to discover and represent interesting new relationships between two existing pieces of information. It is essential that this can be done without additional work by the data publisher, otherwise it won't be done in practice (imagine what it would mean for Yahoo to change their HTML code every time someone creates an addition/change to an ontology). This is a non-trivial problem as it requires the solution of a "reverse pointer traversal" problem.
- as a user, I need to be able to access semantic information published by several independent content providers, and create new semantic relationships among it for my personal use. Further, I need to be able to easily publish those relationships to people, and to software agents by no more than the push of one button. This is generally not feasible on Today's Web, and will not be easier for the Semantic Web.

Over the past several years, we have developed a pure-play peer-to-peer infrastructure that addresses these issues in a fairly novel way, which we will describe in brief in the following. Our goal with this position paper is to solicit feedback on the suitability of this approach compared to other approaches to the Semantic Web as well as the pros and cons of different approaches to integration, prior to us releasing our source code and developer's kit to the public.

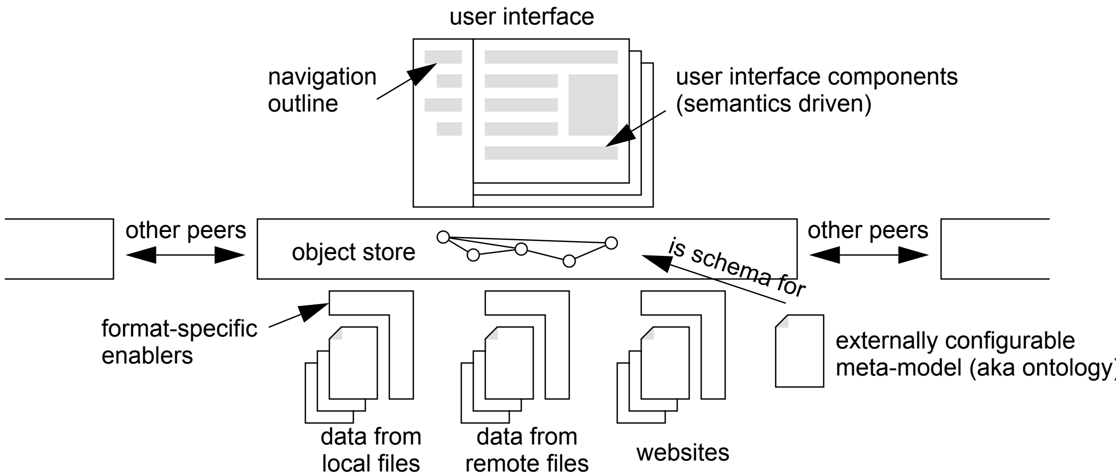
2 CORE FEATURES AND IMPLEMENTATION OVERVIEW

The R-Objects system is a pure peer system. The user interacts with a peer, implemented in Java, that runs locally on his computer. A schematic overview is shown below.

User Interface. The peer application has a component-based, fully-functional user interface which is driven by the underlying meta-model (our term for "ontology" or "semantics definition"). At run-time, the available user interface components "announce" which types of semantic objects they can interact with by identifying concepts from the meta-model. As the user accesses information, the user interface constructs itself dynamically depending on the semantics of the accessed information. This means a high degree of appropriateness of the user interface for the job at hand. It also means that the user as well as third parties can easily extend the application, while maintaining the look-and-feel and without conflicting with other parties' concurrent extensions.

Communications. A peer communicates with other peers through a partly synchronous, partly asynchronous protocol that is currently implemented on top of CORBA. This protocol allows several users to share the same object (edits are subject to distributed locking), but it also allows any user individually to relate any local object to other, local or remote objects, subject to the rules set by the underlying meta-model. This protocol also allows the identification of any object in the R-Objects network by a URI. It also provides smart replication of distributed semantic objects, once the user traverses (his, or someone else's) relationships from his local objects to remote objects. In addition to performance benefits, smart replication allows offline use of the software, which was one of our design goals.

Meta-models. R-Objects supports arbitrary meta-models, which are defined using a simple entity-relationship-attribute modeling language. This allows the use of standard information modeling tools (e.g. UML or ER tools). Any user can locally define his desired meta-model. However, a locally defined meta-model can propagate to other users when another user makes use of one of its concepts.



URLs and Today's Web. A user in the R-Objects network has access all data on the web simply by entering the appropriate URL. As this is a peer-to-peer system, the user also has access to all data on his local computer. Unlike the web itself, files from the web are represented as a web of instance of his local meta-model. There are many benefits to this, one of which is it allows the user to relate arbitrary objects in arbitrary files (not just files, or only objects in XML files). The translation between Web and R-Objects is performed in software components we call enablers, which are uni or bidirectional software converters that can be easily built and inserted.

User-specific information. In addition to accessing information from the (Semantic or Today's) Web, the user can also create new objects simply by instantiating his meta-model locally. For example, the user can instantiate arbitrary meta-relationships between (local or remote) objects, as long as the semantics of the meta-model permit a certain meta-relationship between a certain pair of objects. In particular, the user can create a relationship between objects that have been published by different authors (e.g. different websites, or between one website and an object that the user created locally). Within the R-Objects network, these relationships are always bidirectional, their integrity is guaranteed, and, crucially, their creation does not require the cooperation of the original content author(s).

Active Information and Agents. R-Objects supports both "passive information" and "active information". An object becomes active if its attributes, or its relationships to other objects depend on the attributes or relationships of one or more other objects. For example, an AlgebraicExpression meta-entity is active, and its Value meta-attribute represents the current result of evaluating the expression with respect to its arguments that are related to the AlgebraicExpression. For a user of the infrastructure, there is no difference between active and passive information, allowing the creation of complex structures of information dependencies not all that different from what Gelernter foresaw in [4], except that it driven by an explicit definition of the semantics of the information. In particular, this allows a user to create "agents" (i.e. "active objects") which derive semantic information from raw data. A simple example for that would be a meta-entity that grabs dollar numbers from HTML pages. This provides a migration path from Today's to the Semantic Web.

3 LOOKING FORWARD

R-Objects can access and use all data that is on the web today. An extension to the R-Objects systems to access 3rd-party semantics definition, such as RDF, DAML etc. (by translating it into its own internal representation) seems straightforward, although we have not implemented this at this point, and we would like to explore these opportunities further with participants of the workshop.

Most importantly, the R-Objects system has been built with decentralization in mind not only for users, but also for projects to extend the technology independently without running into central bottlenecks (such as the R-Objects development organization). This is reflected in the user interface (independent implementation of user interface components supported), the meta-model (independent extensions supported), and the types of data and meta-data that can be accessed through enablers (independent extensions supported) and our planned, upcoming release of our Java source code.

We hope to be able to work with other Semantic Web pioneers and technologies to gain critical mass for the Semantic Web to solve these challenging problems together going forward.

4 REFERENCES

[1] T. Berners-Lee, J. Hendler, O. Lassila: The Semantic Web. Scientific American May 2001.

[2] J. Ernst: Is Peer Computing Real? Invited Talk, SCI 2001, Florida, July 2001.

[3] Gartner Consulting: The Emergence of Distributed Content Management and Peer-to-Peer Content Networks. Engagement #010022501, January 2001

[4] D. Gelernter: Mirror Worlds. Oxford Uni Press, 1992.

5 ABOUT THE AUTHOR

Johannes Ernst has been excited about well-defined semantics since his time in CDIF ('93-97) that defined an integrated meta-model for complex engineering information. He discovered the natural fit with peer computing in 1998, when he started developing this described technology. He was recognized as a Technology Pioneer by the World Economic Forum in Davos in January 2001.

Logics for the Semantic Web

James Farrugia
Department of Spatial Information Science and Engineering
National Center for Geographic Information and Analysis
University of Maine, Orono, Maine USA 04469-5711

Abstract. Logic clearly has a role to play in the development of the Semantic Web, most obviously in the model-theoretic formalization of semantics to allow for automated reasoning. But logic can do more than provide model-theoretic underpinnings for the Web. We believe the Semantic Web will evolve to accommodate many kinds of reasoning, based on a variety of logics. We discuss how different logics' languages, axioms, and inference rules can support different kinds of reasoning. Our position is that the research being done on using multiple logics for different kinds of reasoning should be exploited by the Semantic Web community.

1. Overview

Popular expositions of the Semantic Web often stress that *semantic* in this context means "machine processable" or "machine understandable" [1,18]. But the data on the Web are already being processed and in some sense understood by machines. So there must be something more to this notion of *semantic*. This something more is usually taken to be logical (model-theoretic) semantics. We accept the utility of model theory [11,15,16] in providing a formal semantics for the Web. But the emphasis on model-theoretic foundations tends to overlook an interesting possibility: that the Semantic Web might evolve to support many kinds of reasoning and so will need to employ a variety of different logics. The focus on general model-theoretic foundations for the Semantic Web hides important differences among logics that are relevant for different kinds of reasoning.

For example, on certain occasions we may need to reason about what might possibly be the case (e.g., "Is this document possibly related to that one?"). On other occasions we may be interested in a different sort of question (e.g., "Is every *Research Paper* also a *Document*?"). In both cases, a particular logical syntax can be used to govern the formal deductive steps, and an appropriate model theory can be used to formalize the semantics. But whereas the logic used in the first case might be designed specifically to reason about possibility, the logic used in the second case might be designed specifically to reason about subsumption. Thus, different logics would be used for different kinds of reasoning.

The questions of how to use different logics to support different styles of reasoning are being actively investigated within the logic community [2,3,4,5,6]. Recently, similar issues have been taken up by researchers more closely affiliated with the Semantic Web [9,14]. Our position is that the research being done on using multiple logics for different kinds of reasoning should be exploited by the Semantic Web community.

2. Many logics

Model-theoretic semantics allows us to analyze “the validity of inference processes" [9], which in turn paves the way for automated reasoning, widely considered to be *the* major goal of the

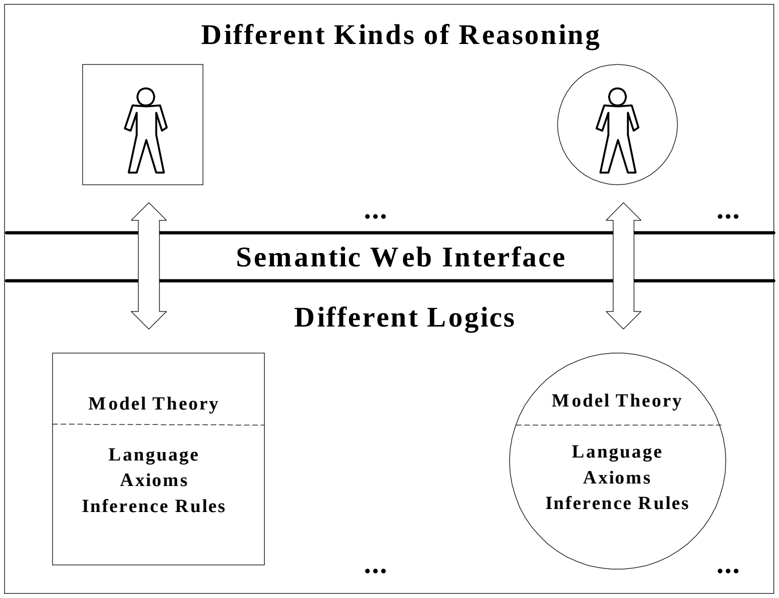


Figure 1 - Different Logics for Different Kinds of Reasoning

Semantic Web. But a model-theoretic foundation for automated reasoning on the Web is only one part of a larger picture (Figure 1). The more interesting parts of this picture come into view once we consider that different kinds of reasoning require different kinds of logical support. Then we can view logic not as a single monolithic foundation for machine-processable semantics, but rather as a collection of different reasoning systems constructed for particular purposes.

3. How Logics Differ

We consider a *logic* to consist of a syntax (or deductive system [8]) and an appropriate semantics (or model theory). We follow the exposition of Epstein [4] to highlight how differences in the language, axioms, or inference rules of a logic support different kinds of reasoning. According to Epstein, "What we pay attention to in reasoning determines which logic is appropriate." He elaborates: "Each logic, other than classical logic, is based on some aspect of propositions in addition to form and truth-value; different aspects give rise to different structural conditions on the semantics, yielding a spectrum of semantics" [4].

For instance, one aspect of reasoning deemed important by the creators of the Ontology Inference Layer (OIL) is the ability to verify that one concept or class subsumes another [12]. This aspect of reasoning is reflected in OIL's language by the use of *concept* as an important primitive [12]. The choice of *concepts* as primitives in OIL's language, along with their role in reasoning about subsumption, demonstrates the connection between the elements of a logic (more specifically, the elements of a logic's language) and the reasoning purposes to which the logic is put.

But reasoning about subsumption is only one of many conceivable purposes for which one might use the Semantic Web. For instance, to reason about what might possibly be the case, it may be useful to employ a modal logic, which adds to the language of classical propositional logic specific modal operators dealing with possibility and necessity, and which also modifies the axioms and inference rules of propositional logic [7,13].

These are just two examples to suggest some of the relevant issues. As people recognize the appropriateness of other kinds of reasoning on the Semantic Web, many issues dealing with multiple semantics will arise and will deserve a more thorough treatment.

4. Conclusion

The Semantic Web will need to employ multiple logics to support a variety a reasoning tasks. The current focus on model-theoretic semantics overlooks the potential benefits of using multiple logics to support multiple kinds reasoning on the Web. The Semantic Web community can benefit by exploiting the research done in the logic community concerning the use of multiple logics for different reasoning tasks.

References

(All URL references available as of May 31, 2001)

1. T. Berners-Lee. Semantic Web - XML 2000. http://www.w3.org/2000/Talks/1206-xml2k-tbl/Overview.html.
2. W. Carnielli, M. Coniglio. A Categorial Approach to the Combination of Logics. *Manuscrito*, vol, XXII, pp. 64-94 (1999), draft available at http://www.cle.unicamp.br/prof/carnielli/articles_published.htm
3. M. Coniglio, W. Carnielli. A Model Theoretic Approach to Translations Between Logics. *Proceedings of the 7th WoLLIC '2000*, pp. 55-65, draft available at http://www.cle.unicamp.br/prof/carnielli/articles_published.htm.
4. R. Epstein. *Propositional Logics: The Semantic Foundations of Logic*, 2nd ed. Belmont, CA: 2001
5. D. Gabbay (ed.) *What is a logical system?* New York: Oxford, 1994.
6. D. Gabbay. *Fibring Logics*. Clarendon Press: Oxford, 1999.
7. D. Gabbay, A. Kurucz, F. Wolter, M. .Zakharyaschev, *Many-Dimensional Modal Logics: Theory and Applications*. Draft of Chapter 2 available at http://www.dcs.kcl.ac.uk/staff/mz/GKWZ/gkwz.html.
8. H. Gaifman. *First Order Logic, Preparatory Supplement for The Incompleteness Results*. Available at http://www.columbia.edu/~hg17/FOL-supplement.pdf.
9. P. Hayes. Email to the www-rdf-logic list, May 16, 2001. http://lists.w3.org/Archives/Public/www-rdf-logic/2001May/0163.html.
10. P. Hayes. Email to www-rdf-logic list May 16, 2001. http://lists.w3.org/Archives/Public/www-rdf-logic/2001May/0109.html.
11. W. Hodges. *A Shorter Model Theory*. Cambridge: Cambridge University Press, 1997.
12. I. Horrocks et al. The Ontology Inference Layer OIL. http://www.cs.vu.nl/~dieter/oil/Tr/oil.pdf.
13. G. Hughes, M. Cresswell. *A New Introduction to Modal Logic*. New York: Routledge, 1996.
14. R. Kent, Foundation Ontology. http://suo.ieee.org/IFF-Foundation-Ontology.pdf.
15. M. Manzano. *Model Theory*. New York: Oxford University Press, 1999.
16. A. Pillay. Model Theory. *Notices of the American Mathematical Society*, 47 (12) December, 2000.
17. G. Restall. *An Introduction to Substructural Logics*. New York: Routledge, 2000.
18. W3C®. Semantic Web Activity Statement. http://www.w3.org/2001/sw/Activity.

Position Paper

Semantic Web Workshop

Christian Fillies, cf@sc4.org
Dr. Frauke Weichhardt, fweichhardt@fweichhardt.de

SC4 is a StartUp company founded January 2001. Our mission is to deliver tools and consulting solutions which help a wide community of users to publish their knowledge on the Semantic Web. We believe that the Semantic Web will change the way people and machines are communicating dramatically. The most important critical success factor (beyond standards) of the Semantic Web is, that the devices which people use to enter their knowledge into the Semantic Web are easy to use and resulting models are easy to understand.

One major problem we found at most of our customers is the need for enterprise or organization wide ontologies, with domain or - especially in Europe - country specific variants. Those ontologies are needed for Knowledge Management Systems, Data Warehouse Systems and the specification of new software. Other applications are CRM and EAI solutions.

Our product SemTalk™ is integrated into MS Office. It uses MS Visio for graphical specification of knowledge. SemTalk™ reads and writes a subset of RDFS. It is integrated in Word XP in a way that it analyzes each sentence you type and matches it with an RDFS model. In order to do this we have implemented a simple model crawler. We expect from this workshop to get partners / ideas to solve especially this issue with a better compliance to standards.
It is our vision that the Semantic Web and SemTalk™ will increase the use of modeling in the daily work significantly. We just attach a RDF model as a graphical summary of its contents. We do not assume every letter or mail in a newsgroup, but in a lot of cases a simple drawing sketching out the most important statement will be worth to spend this small effort. Sharing knowledge and documents will become simple and cheap.

SemTalk™ is implemented basically in Visual Basic and MSXML. It has an integrated object engine to take care of object representation and some reasoning. Visio 2000 is used for graphical presentation. RDFS models are stored in a distributed, hyperlinked fashion as XML flat files.

For the simplified creating of ontologies we use threefold strategy in our workshops:

- Linguistic / statistical analysis of source documents (together with Univ. Leipzig)
- Direct linking of source documents to graphical models and vice versa
- Business process model to reduce the complexity of the modeling space

SemTalk is currently beta software. Pilot projects are going to start soon with a major Swiss bank and a Japanese toy vendor. SemTalk™ has been presented the first time at the DFKI / Univ. Karlsruhe knowledge management conference in Baden-Baden March 2001.

Our partners are currently Microsoft / Visio, IMG AG Switzerland, a major consulting company related to

Hochschule St. Gallen, and Intraware AG, a BPM tool vendor. Intraware's most important US customer for its product Bonapart is Mitre Corp.

We have opened Bonapart to the Semantic Web by an http based R/W RDF(S) interface.
SemTalk™ can be used as a client for BPM tools as System Architect, Bonapart and in the near future with OntoWeb's Sesame. One of the challenges of the workshop might be to specify standard query interfaces to Semantic Web repositories, possibly following Arjohn Kampman's proposal for Sesame. We are also in touch with Persist AG and Ontoprise.

Since our background is BPM and especially simulation, we are very interested in agent technology, DAML and other ways to build appliances using the Semantic Web.

Homepage: www.sc4.org
Beta version: www.cfillies.de
Some slides: www.cfillies.de

Position Paper for the International Semantic Web Workshop (Infrastructure and Applications for the Semantic Web)

Title: Creating the Semantic Web: the Role of an Agricultural Ontology Server (AOS)

Author: Fisseha, Frehiwot; Hagedorn, Kat; Keizer, Johannes and Katz, Stephen (FAO, Library and Documentation Systems Division)

"Knowledge management is vital for effective decision-making.....It is therefore essential to maintain and improve the coverage, quantity, utility, timeliness and accessibility of the information collected and disseminated." (citation from the Strategic Framework for FAO).

FAO is a huge content provider for the World Wide Web. The FAO website has more than 6 gigabytes of information that contains knowledge created by more than 4000 FAO staff working in the world to combat hunger and to help people create a better life.

The semantic web is based on knowledge representation systems. Creating infrastructure for the semantic web is not only an encoding problem. Topic maps need underlying ontologies, as do any RDF description of web sites. Ontologies are emerging as a key aspect of information management in many areas, from the interchange of engineering data to corporate knowledge management.

To create knowledge representation systems (ontologies), knowledge about the represented domains is needed. **This knowledge is not with the developers of encoding systems and software but with the producers and providers of content.**

There is no realistic hope of globally classifying all concepts, terms and relationships; we need to be able to manage and interrelate knowledge representation systems (ontologies) project by project, domain by domain, so that scalability is achieved without either runaway complexity or over-simplification.

FAO, together with partners and other stakeholders in the area, has been developing and maintaining knowledge representation systems in the basic form as represented in the AGRIS/CARIS subject categories and the AGROVOC thesaurus for nearly two decades. The advent of the internet and the World Wide Web gives us the possibility to extend the concepts behind these systems.

We are planning to develop an Agricultural Ontology Server (AOS):

- To allow domain knowledge to be defined and described
- To communicate among domains without semantic ambiguity
- To enable reuse of domain knowledge
- To share the structure and meaning of agricultural information among users and tools
- To provide foundation to build other specific ontologies

And last, but not least:

- To provide more effective dissemination and access to knowledge for users

Briefly defined, the Agricultural Ontology Server (AOS) functions as a central common reference tool for serving ontologies. An *ontology* is a system that contains terms and the definitions of those terms, and the specification of relationships among those terms. It can be thought of as an enhanced thesaurus—it provides all the basic relationships inherent in a thesaurus, plus it defines and enables the creation of more formal and more specific relationships. It is designed to serve as a central focal point for the vocabulary of a particular domain, and to codify and standardise the knowledge within that domain. It enables better communication within and across domains, and structures meaning contained in the domain.

In essence, the AOS provides the "building blocks" that assist in developing and maintaining other ontologies. It will contain the core vocabulary and definitions (multilingual) and the core relationships (including common richer relationships) which subsets of the knowledge domain will use in building and maintaining their own ontologies. For instance, in this case, the AOS provides the reference for all the terminology of the agricultural domain. Knowledge domain subsets, including forestry, fishery, plant biology, sustainable development, organic agriculture and nutrition, will use this reference tool to build their own ontologies. Once these ontologies are created, they can be used to inform knowledge bases—and can be re-used and enhanced by other knowledge bases. This is an iterative process that grows and maintains the ontologies.

The existence of a common ontology server guarantees that common concepts are clearly defined by unique identifiers, and basic relations are used throughout the domains.

A thesaurus has equivalence (USE/UF), broader term (BT), narrower term (NT) and related term (RT) relationships. These relationships provide structure for the terms. For instance, knowing that a broader term for “cereals” is “plant products” and that narrower terms are “maize” and “rye” provides a structure that defines the scope of those terms.

Recently, there has been considerable discussion relating to extending this core set of relationships. In the late 1990s, the American Library Association Subcommittee on Subject Relationships/Reference Structures examined over 165 relationships within the English language alone and from these produced a checklist of twenty candidate subject relationships for information retrieval.

We can use a richer set of relationships to develop tools that provide more granular and more consistent indexing, and more effective searching and browsing for users. With ontologies we can more fully define these relationships—creating rules for developing specific relationships—and thereby provide a means for better knowledge sharing. However, this would have to be balanced with the need for compatibility with existing systems and future interoperability.

Since the server will be the central reference resource for vocabulary control and relationship structure of agricultural terminology, we will need to utilize multiple different sources to build it. The main source will be the AGROVOC thesaurus, which already has the appropriate scope and basic relationships to serve as a base for the AOS. Other sources will include:

- classifications—lists of terms often using hierarchical relationships
- controlled vocabularies—controlled lists of preferred and variant terms based on concepts
- thesauri—controlled vocabularies containing hierarchical relationships
- authority files—controlled lists of preferred and variant names
- glossaries—lists of terms with definitions
- gazetteers—dictionaries of place names
- subject headings lists—broad categorizations of knowledge domains

A key aspect of the AOS is that it will be multilingual. For users in all countries who need access to resources, we need to provide the ability to index and find information in any language needed. The AOS should collect and coordinate terminology, definitions and relationships in the five official languages of the FAO—English, French, Spanish, Arabic and Chinese. Additional languages can be added if necessary by those developing ontologies, if working in the mother tongue of the country is beneficial.

We will need to develop a suite of ontology tools to be used in accessing the AOS and its set of ontologies. This suite should contain tools that allow:

- description—discovery of overlap in terminology and mapping of common terms and definitions
- relationship building—creation of ontologies using common relationships and building new relationships
- coding—storage of terms, definitions and relationships in a standard, interoperable format
- indexing—using ontologies to index resources
- discovery—searching and browsing by users in the AOS or in an ontology
- maintenance—ontology collection, storage, dissemination and evaluation by managers

We will need to incorporate current state-of-the-art standards (RDF and XTM) in the encoding of the ontology server for the KOSs to communicate with each other effectively.

We believe that other stakeholders in the area of agriculture, forestry, fishery and nutrition and the like, who have or need Knowledge Organization Systems for their information ensembles would benefit widely from the AOS. We believe also that the development of tools and software could be inspired by the existence of a large structured system of knowledge representation.

A proposal of infra-structural needs on the framework of the Semantic Web for ontology construction and use[1]

Asunción Gómez-Pérez
Facultad de Informática, Universidad Politécnica de Madrid.
Campus de Montegancedo s/n. Boadilla del Monte, 28660. Madrid. Spain.
asun@fi.upm.es

Three extremely important factors contribute to the construction of the Semantic Web: (1) a common language in which the resources implied can be formally specified, (2) ontologies, which provide a shared knowledge model and description of the domain resources, (3) a workbench for (semi)automatic construction, evaluation, evolution and maintenance of ontologies, and for supporting the selection and use of ontologies for the Semantic Web. We call them the *syntactic*, *semantic* and *technological* dimensions. This position paper only covers the semantic and technological dimensions.

Semantic Dimension

The semantic dimension is related with ontologies. The construction of large and consensuated ontologies for the Semantic Web is difficult, time consuming and expensive to build. Currently, a few domain ontology servers (Ontolingua [1], Ontosaurus [2], Protégé2000 [3], WebODE [4], WebOnto [5], etc.) provide libraries with a few number of knowledge representation ontologies, common-sense ontologies, upper-level ontologies, generic ontologies that could be reusable across domains, domain ontologies, etc. However, the maturity level of such ontologies is insufficient for the construction of the Semantic Web. Efforts exist such as the IEEE Standard Upper Ontology (SUO) Working Group[2], which is working for the construction of a unified SUO ontology.

Therefore, a "hot" issue for the Semantic Web is the construction of a kind of **multilingual and multi-domain reference ontology** that could be used as a shared resource not only for the Semantic Web, but also for Natural Language applications, Intelligent Information Extraction, Intelligent Information Integration, e-commerce, Knowledge Management, etc.

The proposed multilingual and multi-domain reference ontology should provide formal and detailed knowledge models that will allow the vertical intra-operability of systems in specialized domains and also the horizontal inter-operability of application in different domains.

The approach consists of structuring the ontologies in several layers. Figure 1 shows a *multilayered content networks* that can be established between the ontologies that are present in the architecture. The following types of ontologies are needed:

- Several *Knowledge Representation Ontologies*, which formally define the primitives used to represent knowledge under a given knowledge representation paradigm (frames, description logic, etc.).
- *Upper Level Ontologies,* which define the common terms used in the communication between systems, providing a unified upper-level vocabulary for all the systems accessing the ontology.
- *Generic domain ontologies* provide broad, coarse-grained vocabulary in a given domain.
- More specialised ontologies in a given domain (*regional domain ontologies*) can be created. These ontologies can be organised in as many layers as the ontology developers consider necessary.

To speed up the construction of ontologies, existing upper-level ontologies and also standards and initiatives could be automatically processed and enriched. With the current state of affairs, it is more suitable to establish ontological mappings between well-formed existing ontologies and between standards and initiatives than to pretend to build *the* unified knowledge model from scratch.

From the methodological point of view, we also need:

- Methodologies for integrating and merging ontologies.
- Methodologies for evaluating ontologies.
- Methodologies for collaborative construction of ontologies.

Technological Dimension

In the last years, there has been a great number of tools developed for building ontologies (OILed[3], OntoEdit[4], Ontolingua [1], Ontosaurus [2], Protégé2000 [3], WebODE [4], WebOnto [5], etc.). There also exist some tools for merging ontologies (Chimaera [6], Ontomorph [7], PROMPT [8]) and for translating ontologies between different languages. The main problems that arise are:

---

[1] This paper is an extension of the position paper presented to the Program Consultation Meeting on Knowledge Technologies, held in Brussels on 27/04/01.
[2] http://suo.ieee.org/
[3] http://img.cs.man.ac.uk/oil/
[4] http://ontoserver.aifb.uni-karlsruhe.de/ontoedit/

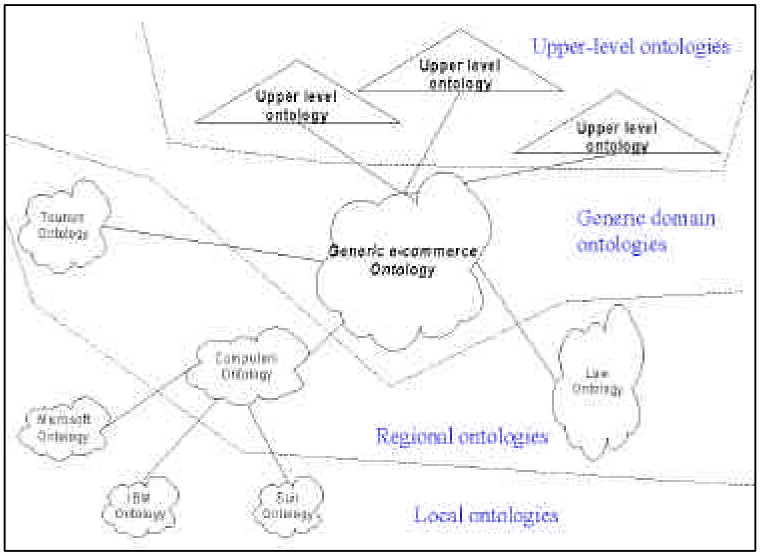


Figure 1. Multilingual and multi-domain reference ontology.

- A correspondence between existing methodologies for building ontologies and environments for building ontologies, except for METHONTOLOGY [9] and WebODE [4], does not exist.
- There exist a lot of "similar" ontology development tools that allow building ontologies, but neither they do interoperate nor they do cover all the activities of the ontology life cycle.
- Most of the tools only give support for designing and implementing the ontologies, but they do not support all the activities of the ontology life cycle.
- The lack of interoperability between all these tools provokes important problems when a given ontology is going to be integrated into the Ontology Library System of a different tool, or if two ontologies built using different ontology tools are integrated using the merging tools.
- None of these tools provide specialized modules that facilitate the (semi)automatic construction, evaluation and configuration management of ontologies.

Consequently, we need **a workbench for ontology developers**, as shown in figure 2, that facilitates:

- Ontology development construction during the whole ontology life cycle, including: knowledge acquisition, edition, browsing, integration, merging, ontological mappings, reengineering, evaluation, translation to different languages and formats, interchange of content with other tools, etc.
- Ontology management: configuration management and evolution of isolated ontologies as well as of ontology libraries.
- Ontology support: scheduling, documentation, etc.
- Workbench Administration.

A methodology for building ontologies using the workbench is also needed.

However, the ontology developers workbench should be accompanied by **a workbench for supporting the use of ontologies (ontology middleware services).** It should include:

- Software that helps to locate the most appropriate ontology for a given application.
- Formal metrics that compare the semantic similarity and semantic distance between terms of the same or different ontologies.
- Software that allows incremental, consistent and selective upgrades of the ontology which is being used by a given application.
- Query modules to consult the ontology.
- Remote access to the ontology library system.
- Software that facilities the integration of the ontology with legacy systems and databases.
- Administration services.

Finally, a wide transfer of this technology into companies, with the subsequent development of a large number of ontology-based applications in the Semantic Web context, will be achieved by the creation of **ontology application development suites**, which will allow the rapid development and integration of existing and future applications in a component based basis.

WebODE

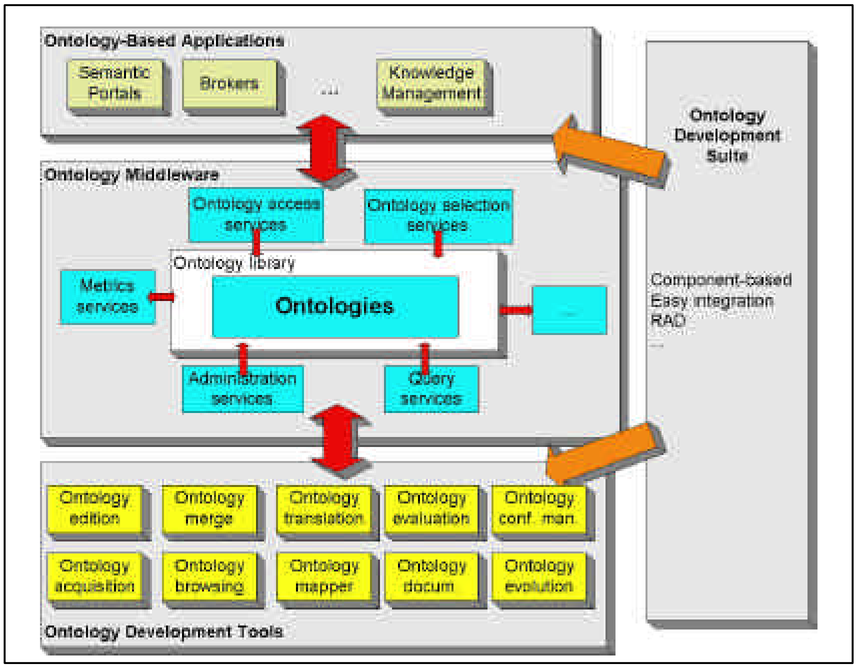


Figure 2. An ontological engineering workbench.

WebODE [4] is not an isolated tool for the development of ontologies, but an advanced ontological engineering workbench that provides varied ontology related services and covers and gives support to most of the activities involved in the ontology development process. In more detail, WebODE covers the following aspects of the workbench previously presented:

- Ontology development.- It offers an ontology editor, ontology translation into several languages, ontology evaluation, ontology documentation and ontology browsing capabilities.
- Ontology middleware.- It offers a well-defined ontology access API and an inference engine implemented in Prolog.
- Ontology-based applications.- Currently, several applications are being developed using the WebODE infrastructure, in the domains of Knowledge Management and e-commerce.

References

[1] Farquhar A., Fikes R., Rice J., *The Ontolingua Server: A Tool for Collaborative Ontology Construction*. 10th Knowledge Acquisition for Knowledge-Based Systems Workshop, Banff, Canada. 1996.

[2] Swartout, B.; Ramesh P.; Knight, K.; Russ, T. *Toward Distributed Use of Large-Scale Ontologies*. AAAI Symposium on Ontological Engineering. Stanford. USA. March, 1997.

[3] *Using Protégé-2000 to Edit RDF*. Technical Report. Knowledge Modelling Group. Stanford University. February, 2000. http://www.smi.Stanford.edu/ projects/protege/protege-rdf/protege-rdf.html

[4] Arpírez, J.C., Corcho, O., Fernández-López, M., Gómez-Pérez, A. *WebODE: a Workbench for Ontological Engineering*. Submitted to K-CAP 2001.

[5] Domingue, J. *Tadzebao and Webonto: Discussing, Browsing and Editing Ontologies on the Web*. 11th Knowledge Acquisition of Knowledge-Based Systems Workshop (KAW98) Banff, Canada. 1998.

[6] McGuinness, D., Fikes, R., Rice, J., Wilder, S. *The Chimaera Ontology Environment*. 17th National Conference on Artificial Intelligence (AAAI-2000). Austin, Texas. August, 2000.

[7] Chalupsky, H. *OntoMorph: A Translation System for Symbolic Knowledge*. KR-2000. 471-482. 2000.

[8] Fridman, N., Musen, M. *PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment*. 17th National Conference on Artificial Intelligence (AAAI-2000). Austin, Texas. August, 2000.

[9] Fernández, M.; Gómez-Pérez, A.; Pazos, J.; Pazos, A. *Building a Chemical Ontology using methontology and the Ontology Design Environment*. IEEE Intelligent Systems and their applications. #4 (1):37-45. 1999.

SWWS Position Paper : Semantic Web and Adaptive Multimedia Access

Yoshihisa Gonno[1, 2], Fumihiko Nishio[1, 2], Tomohiro Tsunoda[1], Yasuaki Yamagishi[1, 2]

[1] Sony Corporation
6-7-35, Kitashinagawa, Shinagawa-ku, Tokyo, 141-0001 Japan
phone:+81-3-5448-6443

[2] Information Broadcasting Laboratories, Inc.
1-1-1, Nishi-asakusa, Taito-ku, Tokyo, 111-0035 Japan
phone:+81-3-5326-7383

email: ygonno@sm.sony.co.jp

Introduction

Digital broadcasting systems have been coming into wide use all over the world. Efficient use of bandwidth allows us to transport hundreds of TV programs through a single broadcasting satellite to millions of user terminals simultaneously. On the other hand millions of multimedia contents, including TV-like live programs, are streaming over the Web. In order to make the most use of the digital information infrastructure, various forms of information must be well-structured in common ways between TV and the Web environment from every point of view such as authoring, delivery, browsing, searching and retrieval. Well-structured information will allow us to navigate in floods of services between TV and the Web seamlessly. In these senses, we have reported basic ideas on a future framework for integration of TV and the Web[1][2][3][4] and a further study on the "Integrated Broadband Environment for Personalized TV Experience (IBEX)"[5] as a TV-Anytime[6] service platform. We are expecting that the idea of the semantic web will help us to improve future multimedia access environment and our IBEX will be a good application for the semantic web in this area.

Our Positions on the Semantic Web

What is our view on the Semantic Web? What is the interest of our organization in the Semantic Web?

Our interests lies on the adaptive multimedia access through the integrated broadband environment. As a way of providing consumers adaptive access to favorite contents and metadata customized not only for user preferences but also terminal capabilities, we are expecting that the Semantic Web will be one of the most important ideas.

Are we planning to provide services and machine readable data on the Web? How? Access requirements?

Currently we can't appoint the exact date. It is because there is a great difficulty in involving multimedia content industry who can provide high quality contents such as TV programs and movies. In most cases such content industry has their customs much different from the consumer industry. We are expecting that numerous amount of efforts must be made to interconnect each other.

Which languages and tools are we currently using?

Our initial proposals[2][3][4] for describing metadata of TV programs were based on the RDF data model because we considered that it was easy to understand and enough to represent semantic relationships between and within TV programs. However, as the later trend moved to XML Schema, MPEG, which is ISO working group in charge of multimedia format, decided the use of XML Schema as a multimedia description definition language, and then TV-Anytime, which is TV industry forum working on TV metadata standardization, is following MPEG. Considering such background, currently our prototype system will be mostly based on the MPEG and TV-Anytime specifications although we are still expecting the use of RDF for our future development in approapriate manners.

What do we envision to be the most important practical uses of the Semantic Web in a few years ?

Because of the difficulty in involving content industry, it is difficult to provide practical services in a few years. But we are sure that the adaptive multimedia access through heterogeneous networks with heterogeneous devices will be one of the most beneficial uses of the Semantic Web.

What applications in our organization would improve by making use of the Semantic Web?

As functions of the IBEX - Integrated Broadband Environment for personalized TV experience

- Adaptation of program selection based on user preferences
- Adaptation of content and metadata presentation based on device capabilities
- More to come..

What new research and tools need to be done to support our use of the Semantic Web?

- Efficient production of metadata because it is resource consuming
- Establish practical metadata schemas to provide beneficial services for users
- Digital rights management of content and metadata
- Interconnecting efforts between content industry and consumer industry rather than technical research

References

1. Y. Gonno, "A View to the Broadcasting and the Internet". W3C Workshop on "Television and the Web", June 1998, <http://www.w3.org/Architecture/1998/06/Workshop/paper03/>.
2. Y. Gonno, et al., "Metadata Structuring of Audiovisual Data Streams on MPEG-2 Systems". Metastructures98, August 1998.
3. F. Nishio, et al., "Transporting RDF Metadata Associated with Structured Contents". Metastructures98, August 1998.
4. Y. Gonno, et al., "Data Models for Distributed Multimedia Contents Associated with Structured Metadata". ICCC99, September 1999.
5. Y. Gonno, et al., "White Paper on Integrated Broadband Environment for Personalized TV Experience (IBEX) - Preliminary Edition", ACM Multimedia2000, November 2000, <http://www1.acm.org/sigs/sigmm/MM2000/ep/gonno/>.
6. TV-Anytime Forum, <http://www.tv-anytime.org/>, Metadata and Content Referencing Specifications also available.

Semantics for Scientific Data: Smart Dictionaries as Ontologies

Syd Hall, Nick Spadaccini, Doug du Boulay and Ian Castleden, Crystallography Centre and Department of Computer Science and Software Engineering, University of Western Australia, Crawley, 6009, Australia (syd@crystal.uwa.edu.au)

The Challenges

Phenomenal growth in scientific databases over the past decade, as well as the enormous disparity in data expression across long-standing and important taxonomic collections poses serious challenges to existing data handling methodologies. These will undoubtly be met with an array of approaches - common data protocols, better collaboration, global ontologies and active knowledge bases. Some scientific disciplines are advanced in removing the communication and access barriers, and there is a general recognition that a much higher level of data semantics is a key objective.

Existing data handling approaches apply semantic knowledge (meta-data) which is encoded as part of highly customised software (e.g. existing database, procurement, and inventory systems) lack the generality or extensibility needed for the easy addition of new data structures or methods. The major challenge met by our research is the development of a generic object-oriented approach to coalescing dictionary meta-data and instantiated data into executable processes that will underpin active knowledge bases.

A Generic Dictionary Approach

We embed machine-executable meta-data into dictionaries as simple text attributes capable of representing complex data relationships. These "smart" dictionaries provide the knowledge framework for generating data manipulation and interpretation tools targeted at local data needs. The presence of extensive meta-data in a dictionary affects future archival practices, in that only non-derivable data (i.e. measurements, etc.) need be archived - the rest can be generated from current knowledge. The effect of this paradigm shift on database management systems will be ubiquitous. In science, data can be arranged and interpreted according to derivation dependencies and the semantic content of dictionaies will proffer a level of flexibility and generality that is unattainable with current approaches.

Progress So Far

Our research is aimed specifically at designing a generic knowledge-base model using the Star File (Hall, 1991; Hall & Spadaccini, 1994), and at developing supporting tools. A Star File contains data that are textual, loosely structured and self-identifying. This work will extend the recent prototype dictionary efforts (Spadaccini, Hall & Castleden, 2000) of the authors. Applications of the Star File have been widely used for a decade. In chemical science, a domain-specific version of the Star File, the crystallographic information file (CIF) (Hall, Allen & Brown, 1991) is used extensively for publication and database purposes. In biological science, the macromolecular CIF data file mmCIF (Bourne et al., 1997) has been adopted by the Protein Data Bank (PDB), the Nucleic Acid Database (NDB) and Macromolecular Structure Database group at the European Bioinformatics Institute.

The scope of Star File data is enhanced considerably when individual items are defined as meta-data stored as a collection of data attributes in dictionary files. The allowed attribute types represent the definition language of the Star dictionary (DDL), and two DDL versions (Hall & Cook, 1995; Westbrook & Hall, 1995) are in current use. The development of a prototype version of a relational expression language dREL (Spadaccini, Hall & Castleden, 2000), as part of a new dictionary language StarDDL, is the basis for the current research program.

The development of dREL and StarDDL prototypes has shown that the precision of data definitions is enhanced significantly by specifying relationships between items as symbolic expressions that can be used to compute derivative data values. In particular, this work demonstrated that Star dictionaries are made much richer semantically when the attribute set is extended to include stronger typing and executable methods. The StarDDL differs significantly from other dictionary languages that are used solely to validate the structure and content of a data file (e.g. DTD in XML). A StarDDL dictionary may be compiled into executable dictionary objects that can be injected with specific data instantiations (i.e. particular data within a file) so that related items are dynamically linked through the dictionary methods. This is a dictionary approach which is well suited to knowledge retention and reuse. The approach does, however, incorporate other languages and data handling approaches when they complement the application of StarDDL dictionaries. Although XML has no intrinsic method functionality, it is used to interface our dictionaries to other computing languages and to off-the-shelf editing/browsing software.

Objectives and Significance

The objectives of this project are directed at the most serious deficiencies in existing data handling methodologies. Most archived data in science are unsuited, and even inaccessible, to modern access tools. Biological taxonomic data are a case in point. There is an enormous and continuing effort to capture biological-species information in the many museums, herbaria and universities around the world, with almost as many databases archiving taxon-based descriptive data. There are currently CODATA and OECD (Edwards et al., 2000) efforts within the GBIF program to coordinate and integrate the coding standards, such as DELTA (Dallwitz et al., 1992), used in these collections so as to provide new data structures better suited to systematic query methods. This interest also reflects a need for on-line sharing of data across disciplines - such as the integration of taxonomic data derived from morphology with molecular and genome sequence data.

The molecular structure data in the Nucleic Acid Database (NDB) at Rutgers University, New Jersey, and the taxonomic botanical data in the Western Australian Flora (Paczkowska & Chapman, 2000) and FloraBase database at the WA Herbarium in Perth, are of special importance to the project. They are excellent exemplars of data which must be interoperable, via consistent protocols, with facilities at other sites and in other countries, and therefore provide ideal test data for our research.

SWWS 2001 Position Paper
Sandro Hawke, W3C

I view the Semantic Web as a practical architecture for a universal information system. It's an old dream: connect the world's computers into one massive system that can take advantage of its reach among locations, its computational resources, and its connections throughout humanity. We've made a lot of progress over the years. I see the Semantic Web as taking several of the next steps, building on successes in different areas, recognizing the principles of open and interoperable systems demonstrated in the history of the Internet and Web.

My interest ranges from the most central peices of the infrastructure up through various application areas. At the low levels, I want to make sure the design is simple enough that a large population can understand it and help it grow. The basic idea of message-passing peer agents presenting a relational database interface seems simple enough; the challenge is to present the real world complications (like network partioning) in the appropriate ways. We need to understand the system well enough to make it look simple.

Moving up the stack, one key to scalable open systems is an open but stable namespace for publishing information, especially meta-information like database schemas, ontologies, rules, and programs. I am very concerned about untangling the complexities of how URI-like-strings are being used as logical symbols and also to identify web pages and content.

Given a simplified relational data model with stable identifiers (symbols), the next step is to add vocabularies for more expressive communication, both for domain-specific areas and for cross-domain fields ranging from the simple (documentation of information) to the more complex (logical formulae defining some terms from others).

The essential components of the Semantic Web, then, are:

1. Layer 1: A language for making simple declarations of fact, using open identifiers which are optionally recognized by various agents. The behavior of agents receiving declarations they do not recognize must be clearly defined and adjustable for different circumstances. This language could be based on the current RDF XML syntax, SQL, KIF, or almost any formal language. It could also be based more directly on arbitrary data formats (eg XML) with a more-complex associated mapping to a relational model. With the right associated language definitions, in fact, we may be able to equivalently use any formal language.

2. Layer 2: Vocabularies for various domains of discourse, allowing Layer 1 declarations to mean something. Everybody should be able to create and disseminate vocabulary terms. Some should be standardized within certain communities for certain purposes, especially the vocabularies for

    - common terms (eg numbers)
    - common information structures (eg sequences)
    - describing vocabularies
    - sets of declarations (information packages)

3. Protocols for exchanging Layer 1 declarations in both active (send or "push") and passive (get) modes. We need protocols which work across slow networks, fast networks, between processes on a computer (possibly running sequentially), and between modules in a process.

4. Finally, as the system evolves, we will need general agent software which can efficiently handle an increasing portion of information handling and processing, following instructions in an evolving vocabulary. This kind of software can be seen as a library or a self-contained agent (or software robot, or daemon), following instructions in a stateless (eg JVM) or stateful (eg DBMS) mannter. As we develop agents which can properly handle this abstracted information processing, more application knowledge will become directly part of the Semantic Web.

All of these technologies already exist in a variety of forms, so perhaps we could say the Semantic Web already exists, but in general they are not interoperable and they are certainly not interoperating on a wide scale. Each of the above essential components needs to be revisited with a clear eye to how it interacts with the other components, across the wide variety of possible applications.

The most essential applications in my view are (1) the ones which support autocatalysis of the Semantic Web, such as rule-based systems for managing rules and discussion systems for supporting Semantic Web design discussions, and (2) the ones which bi-directionally connect existing information resources (websites, databases) with open and interoperable Semantic Web forms.

Realizing Visions for the Web

A common theme that has emerged in several recent books about science is that many of our most significant breakthroughs have resulted from insights gained while trying to answer a particular question which challenged the prevailing paradigms. The best example that comes readily to mind is Michio Kaku's description of Einstein's question which led to his theory of relativity in the book, Hyperspace.

This position paper is premised on the view that the semantic web represents a qualitative improvement to the current web. For the Semantic Web Workshop, a pressing question that needs to be asked is, "Does this activity provide us with the opportunity to better realize the visions of Vannevar Bush [1], Douglas Engelbart, Ted Nelson, and Tim Berners-Lee?"

As the semantic web community develops technology to enable background agents to bring the most relevant information to a person (the tool systems), these tools could bring us closer to Engelbart's vision for simultaneous co-evolution of human systems if explicit attention is given to the semantic web as the means to a greater end of the networked computer as the tool to augment the human intellect as first described by Engelbart [2]. A potential danger I is if the semantic web community becomes too enamored with the "gee whiz" factor of the tool system, and fails to recognize the importance that the semantic web has for truly enabling augmented human-human conversations within an improved hypertext environment. Unfortunately, this danger has a precedent within the Artificial Intelligence community, where an over-emphasis on technology stunted the broader awareness and acceptance that Engelbart's work has only recently begun to achieve. Similarly, the current web compromises the vision of Tim Berners-Lee, with the current prevailing browsers being passive renderers of HTML instead of an interactive publishing tool (W3C's Amaya is closest to his original vision [3], and does not approach the robustness of Nelson's Xanadu vision and Engelbart's Open Hyperdocument System. Both of these efforts have been reaching for ways to move the web toward a closer realization of these visions, evidenced by the New Xanadu Model for the Web [4] and the OHS Project [5], respectively..

My hope for this workshop is to bring these ideas and perspective to the forefront of the semantic web community, so we do not repeat the past compromises which has kept us from realizing the web's true potential.

Karl Hebenstreit, Jr.
President
ParadigmLeaps.org
E-mail: karlhjr@acm.org
(launching this summer)

References

1. http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm
2. http://www.histech.rwth-aachen.de/www/quellen/engelbart/ahi62index.html
3. http://www.w3.org/Amaya/
4. http://www.xanadu.net/xuWeb.html
5. http://www.bootstrap.org/ohs2/index.html

A Position Paper: Semantic Web Testbed for Manufacturing B2B

Nenad Ivezic, Michael Gruninger, Al Jones, Steve Ray
Manufacturing Engineering Laboratory
National Institute of Standards and Technology

Importance of Semantic Web

For more than twenty years, the Manufacturing Engineering Laboratory (MEL) at the National Institute of Standards and Technology (NIST) has been involved in the development of content standards for manufacturing enterprise activities ranging from product data interchange to manufacturing processes representation to enterprise integration models. These standards enable interoperability among the software applications that implement these activities.

The emerging Semantic Web carries a promise of new advances in the area of business-to-business (B2B) interoperability because it provides common infrastructures for representing and exchanging semantics of enterprise-level activities and concepts.

The Testbed Service to the Community

The NIST MEL has initiated development of a Semantic Web Testbed for Manufacturing B2B as a distributed environment for experimentation, analysis, and evaluation of emerging Semantic Web technologies.

This Semantic Web testbed will enable NIST to help the B2B community better understand the capabilities of Semantic Web; to affect the development of Semantic Web technologies and standards; and to facilitate introduction and acceptance of those standards by the B2B community.

Through an on-going interaction with manufacturing enterprises and B2B standards organizations, NIST MEL will identify and capture realistic interaction scenarios to drive testbed development. We will build and make available a repository of these scenarios encoded using a Semantic Web language (e.g., DAML+OIL). We will assist in developing extensions to the adopted Semantic Web language so that the scenarios can be captured. We will provide B2B ontologies that formally define the manufacturing concepts and behavior roles that appear within the scenarios. We will give demonstrations of interoperability using applications that utilize the advertised ontologies. Initially, the demonstrations will use machine-readable data accessible through a standard Web server; later, however, we plan to adopt some discovery and access service to advertise and offer our data to the B2B community. Ultimately, we would like to offer our service as a part of a multi-agent system capable of discovering and using ontological services (e.g., planned FIPA ontology server activity).

An Opportunity: B2B Interoperability of Enterprise Systems

We believe that B2B interoperability of enterprise systems is potentially one of the most important commercial uses of the Semantic Web. At NIST, we are investigating several essential issues in manufacturing B2B interoperability. First, we are examining alternative B2B frameworks that enable collaborative work among independent but cooperative enterprises. As part of this effort, we are evaluating the DAML+OIL semantic layer from the perspective of representing concepts needed for collaborative development of engineered products. We created a set of DAML+OIL ontologies that describe searchable manufacturing services and a basic taxonomy to describe other types of services (See http://cim4.ie.psu.edu:12/daml). We created a related set of DAML+OIL ontologies that can describe manufacturing service requirements. A web service directory (Semantic Web based search engine, e.g., RDFDB, RDF Query) may be set up, so that the user agents (client of service) can search for desired service providers.

Second, we have initiated a project to develop metrics and test methods for resolving semantic differences that result from the usage of heterogeneous ontologies in B2B scenarios. We are developing architectures and algorithms that will allow (1) semantic differences between heterogeneous agents to be recognized and resolved at runtime, and (2) the resolution to be viewed as an abductive reasoning process.

Third, we seek a process, based on individual use cases of B2B interoperability scenarios, for synthesizing formal models of coordination among the participants in the scenario. We are looking at the new family of coordination modeling methods based on pragmatic linguistic approaches (e.g., discourse analysis) coupled with formal approaches (e.g., temporal logic). We believe that such methods will be essential in the development of B2B services on the emerging Semantic Web.

Fourth, we are using a wide range of modeling languages in the development of the testbed including UML, KIF, DAML+OIL, and EXPRESS. In addition, we are partnering with a number of universities, manufactures, standards organizations, and software developers. We believe that this approach will maximize the likelihood of transitioning between research prototypes and commercially viable standards and software.

Need for A Bridge Between Research and Industrial Worlds

The research to achieve the ultimate goal of interoperable B2B enterprise systems is multi-disciplinary. Moreover, to be successful, that research must be conducted in conjunction with the real world of B2B standards, technologies, and applications. On the other hand, the technologies to make a significant impact on the B2B world must include the Semantic Web. To be successful, it will have to embrace the already deployed systems, work with evolving B2B standards, and attempt to enhance capabilities of both. The proposed testbed is one initial step in the direction of bridging the two.

THE ROLE OF SEMANTICS AND INFERENCE IN THE SEMANTIC WEB
A COMMERCIAL CHALLENGE

Rob Jasper

Anita Tyler

{Rob, Anita}@HawthornTech.Com
Hawthorn Technology Group, LLC
524 Second Avenue, Suite 500
Seattle, Washington 98104-2323

Rob Jasper is VP of Research and Technology Assessment and Anita Tyler is VP of Technology at Hawthorn Technology. They are currently working to provide a foundational architecture to a new company focused on enterprise knowledge sharing and awareness using the semantic web. Both are former members of the technical staff at FizzyLab and former researchers at Boeing's Research and Technology center in Seattle.

Many of us believe in Tim's, Jim's and Ora's semantic web vision in which "a new form of web content that is meaningful to computers will unleash a revolution of new possibilities." This vision relies on transforming web content from a form that facilitates human interpretation (i.e., HTML) to a form that's designed for interpretation, reasoning, and manipulation by computer agents (based on XML). We believe agents will eventually become the prime consumers of raw web content, acting as knowledge intermediaries, transformers, and brokers, for the end consumers, humans. In this vision, humans are consumers at the top of a "knowledge food chain" that rests upon a foundational layer of semantics and inference technologies. Despite all the rhetoric about the importance of semantics, and to a lesser extent, inference, it's not entirely clear what role semantics and inference play in today's web or tomorrow's semantic web. Different researchers see different roles for semantics and inference in the semantic web. Some questions to consider are:

- How do "intelligent" programs running on the web today function without an underlying layer of semantics and inference?
- To what extent does XML itself convey semantics?
- Does semantics play any role in the runtime operation of the semantic web or does it function primarily as a software engineering tool?
- Can the semantic web exist without inference?
- What needs to be done for these technologies to prosper commercially?

We are currently working to provide a foundational architecture to a new company focused on enterprise knowledge sharing and awareness using the semantic web. After spending years in a research setting we've taken a fresh look at the world from a more commercially-oriented perspective.

Semantics Today

An increasing percentage of web pages are dynamically generated from structured or semi-structured information sources (*e.g.,* databases, knowledge bases). Because the rendering of these pages occurs in HTML, the underlying structure and linkage to meta-data is lost. Many commercial applications attempt to recover the underlying "semantics" through screen scraping and technologies like wrapper induction. With all the problems and brittleness of this approach, it continues, demonstrating commercial demand for access to the structure and semantics of the underlying data.

Standards such as XML provide a convenient mechanism for bypassing the HTML layer and connecting producers more directly to consumers of semantic content. Nevertheless, XML by itself is not capable of conveying semantics. So why should using XML make the job of any easier? XML only conveys semantics to the extent that humans representing producers and consumers agree ahead of time on the meaning of the tags. Programmers, then embed their agreements into the programs that act as producers and consumers of XML content.

Standards such as XML Schema function as software engineering design specifications for programs and play a role in checking whether XML data is well formed or not. XML Schema does not convey the semantics of the data; the semantics must be agreed to at design time.

Semantics Tomorrow

Languages such as DAML+OIL and to a lesser extent, RDF Schema provide the means of describing semantics in a way that is machine interpretable. The fact that these languages are machine interpretable and have semantics does not imply that they will have any role in the runtime operation of the semantic web. These languages and their associated inferencing procedures could be used simply as a means to the same end as XML Schema. That is, they will play the role of sophisticated software engineering tools to help humans come to agreements about semantic exchange. Some researchers have argued that this might be the only role for semantics and inference in the semantic web. We disagree. While we believe this role is important, restricting semantics and inference to a software engineering role significantly limits the potential of the semantic web. It continues to leave developers with the problem of having to obtain *a priori* agreement between parties wishing to have a “semantic exchange”. While the semantic web can exist without runtime inference, the benefits of semantics and inference as purely software engineering tools aren’t enough to warrant significant investment by commercial enterprises.

The Challenge

We believe that runtime application of semantics and inference is an important factor in commercial adoption of semantic web technologies. An example of a technology desired by the commercial sector is automated generation of “glue functions” that can translate terms between related ontologies. Automated translation will free developers from having to embed hard-coded knowledge of the semantics of every meaningful tag into their applications. Applications will be able to reason about new terms provided they can relate them to something they already understand.

Having semantics and inference play a significant runtime role in the semantic web creates a different set of challenges than simply relying on them as software engineering tools. Runtime interpretation of semantics and inference introduce new problems that aren’t currently a major focus of the research community. These include such issues as:

- Performance and scalability
- Reliability
- Replication
- Incremental update of assertions and rules
- Security

We encourage the research community to join us in addressing these issues, thereby helping to bridge the gap between academia and widespread commercial adoption of these technologies.

Semantic Web Position Paper

With approximately 6,000 customers, 5,000 employees and $1 billion in revenue in fiscal year 2000, J.D. Edwards & Company (NASDAQ: JDEC) is the leading provider of agile, collaborative solutions for the Internet economy. For more than 22 years, J.D. Edwards has provided innovative, flexible business solutions essential to running complex and fast-moving multinational organizations ñ acting as a true business partner enabling companies of all sizes leverage their existing investments, take advantage of new technologies, maintain competitive advantage and deliver shareholder value.

J.D. Edwards provides a rich set of tools and a comprehensive set of business logic to our customers in a product called ìOneWorldÆî. OneWorldÆ tools use meta-data to describe the user interface and business logic in an operating system and platform neutral way. **Our solutions extensively use XML to represent the data.** OneWorldÆ is supported on the OS/400Æ, Solaris, AIXô, HPUX, Windows 2000 and Windows NT operating systems. We have a HTML, Java and Win32 client solutions.

We consider semantics to be the next logical step in web evolution. We believe execution of our vision of collaborative enterprise commerce will be greatly aided by Semantic Web.

Web services are gaining popularity among software developers and vendors. However, existing protocols, only address the discovery of the web services by humans and binding to the discovered web service as a design time activity. This is due to the reason that these protocols rely on syntactic interoperability, not a semantic interoperability. To achieve, true dynamic discovery and dynamic binding to web services by a web service client, we believe, web services shall provide machine understandable semantics. This will be achieved by open standards.

Our current interest is in researching the applications of Semantic Web for enterprise collaboration. We believe the Semantic web will impact applications in great enormity ranging from how applications are described and accessed to how software is distributed.

We believe effective search agents; intelligent commerce agents will aid in quicker adoption of Semantic Web in a collaborative commerce scenario.

We would like to participate in the Semantic Web Conference to find out about latest ideas, and also learn how we can contribute to the technology in the immediate future.

Position Paper for the International Semantic Web Workshop (2001): *Building a Semantic Web for the Intelligence Community*

Brian Kettler[1]
ISX Corporation
bkettler@isx.com

Intelink, the U.S. Intelligence's Community's classified intranet, shares many of the challenges of the World Wide Web. Users of Intelink are diverse (e.g., analysts, warfighters, policymakers) and distributed over the world. Information is still largely in English text (and HTML) documents on web servers with keyword-based search tools and browsers being the primary user applications. Security issues are paramount in a community that has traditionally lacked the culture and the infrastructure to share information widely beyond those who have a demonstrated "need to know". Despite the existence of community standards, Intelink producers are in the nascent stages of using HTML metadata tags and content-based markup languages such as XML. The latter presents an additional challenge of gaining agreement among user communities on sets of tags whose meaning is explicit and exploitable by human users *and* software agents.

The Horus Project is a joint effort by the Dept. of Defense Advanced Research Projects Agency (DARPA) and the Intelink Management Office (IMO) to bring semantic web technologies to Intelink and the Intelligence Community. In its second year, Horus is currently refining a toolkit to bring semantic web tools to user sites on Intelink. This will support users in building enhanced, web-based knowledge portals that provide access to both structured data in databases and unstructured data extracted from web sources.

The focus of Horus is to enable and exploit *semantic*-based markup of sources to promote information discovery and integration – ultimately by software agents as well as humans. Users and agents will access, manipulate, and create knowledge that is organized as Horus "knowledge objects". These (conceptual) objects represent real-world entities such as military units, terrorist organizations, and geopolitical events. Information in knowledge objects is linked to its source (i.e., a database or web page). This supports the maintenance of information pedigrees and drilldown to the original sources. User sites will build portals to provide access to these objects, resident in a Horus Knowledge Base (KB).

In addition to working with the site developers of user portals, we are also coordinating with the Joint Intelligence Virtual Architecture (JIVA) Project's Knowledge Map effort, IMO efforts for tagging standards (metadata, security, and content), and related Intelligence Community efforts for document markup (using XML, etc.). We view Horus' approach as unique and complementary to those efforts.

To provide access to information from web sources and databases, the Horus Project is applying the DARPA Agent Markup Language (DAML). Our previous work applied the Simple HTML Ontology Extensions (SHOE) language from the Univ. of Maryland. We building five classes of tools in the Horus Toolkit:

[1] Dr. Brian Kettler is the Technical Lead for the Contractor Team for the Horus Project. This team consists of BBN, ISX, and Logical Sterling Federal Systems. *The views in this position statement are those only of the author and do not necessarily reflect the views of the Horus Project's sponsors – DARPA and IMO.*

- **Ontology Tools** –facilitate the authoring, application, and maintenance of multiple, distributed, lightweight domain ontologies. Our ontologies are written in DAML and leverage DAML Program's emerging tools for ontology authoring and validation.
- **Markup Tool**s –facilitate the use of domain ontologies to mark up documents in DAML. Our markup tools support manual markup and automated markup. The latter currently uses Lockheed's AeroText product for rule-based parsing of natural language documents into structured entities.
- **Knowledge Base (KB) Tools** –support the storage of information from marked up documents, database extracts, and user-created assertions about knowledge objects. Currently we are using an Oracle DBMS and the Parka Knowledge Base Management System, a frame system developed originally by the Univ. of Maryland that supports inheritance-based inferencing.
- **Data Source Access Tools** –support access of online databases to extract information (as DAML) for the Horus KB and to drill down to details.
- **Portal Building Tools** –support the development of web-based, knowledge portals for users to access and manipulate knowledge objects in Horus KBs.

Horus has been an interesting "use case" for semantic web technologies thus far. We have encountered a number of issues in applying these technologies to a dynamic, real-world domain. These include:

- how usable are complex markup languages such as RDF and DAML+OIL, both for ontology definition and content representation. Right now the tools lag behind the language development.
- how much to mark up in web sources: i.e., is the goal to duplicate all of the document's content or merely provide a more semantically-rich index to that document
- how can the subjects of statements in document markup be matched with equivalent entities (e.g., a particular military unit or terrorist event) in information extracted from databases – especially given inconsistency in naming of entities, differences in levels of abstraction, differences in URLs between classified networks, etc.
- where does markup come from – e.g., how can existing HTML and XML markup be converted to DAML markup automatically
- where do ontologies come from – e.g., how can existing database schemas, XML DTDs, and other domain models/taxonomies be leveraged
- how to maintain historical information from data sources
- how to distribute Horus KBs over Intelink to agents, crawlers, etc.
- how to support the security/dissemination restrictions on the various versions of Intelink – e.g., (semi-)automated tools for restricted-word search in markup

We expect to encounter additional issues as we put in place the first "Horus-enabled" end user site on Intelink by Fall, 2001. We are coordinating with the integration contractor and principal investigators in the DAML Program to facilitate bi-directional technology transfer between Horus and DAML. We are applying tools from the DAML, XML, and RDF communities to Horus. To the DAML Program and related efforts, we are providing a source of real-world requirements and an "alpha test bed" for Semantic Web concepts and technologies in a dynamic and important domain.

VR Communities and the Semantic Web

Authors:
Mike Lindelsee
Gabe Wachob

From our perspective, the Semantic Web is going to provide an incredible layer of abstraction on top of the World Wide Web that will allow information to be discovered, correlated and viewed in completely new ways. The Semantic Web will allow interaction between producers and consumers of information without pre-arrangement -- using information gleaned from "casual sources" (i.e., arbitrary meetings between people, agents, data, services, etc. in cyberspace).

Our primary interest in the Semantic Web is in its application to various representations of the elements that make up the World Wide Web -- people, agents, services, data, etc. These representations could be visual, aural and/or tactile (other senses as I/O devices become available). With the constantly changing nature and structure of the data on the Web, it is extremely hard to find a stable way to visualize the information on the Web. The Semantic Web will lay the foundation for programmatic understanding of the relationships of the elements of the Web. This will lead to straightforward mechanisms for visualizing these elements. As new elements are added to the Web, removed from the Web, changed, moved, reorganized, etc., the relationships can be captured and visualized using the Semantic Web.

Traditional Virtual Reality (VR) attempts to visualize data in ways that are comfortable and "natural." Our interests go beyond this into completely new ways to visualize the elements of the Web while maintaining comfortable and intuitive interfaces. Our belief is that a more natural, intuitive interface on the Web (or Semantic Web) will foster the development of larger communities of interacting users on the Web and facilitate easier access to both data and services as they become available.

As we see it, the most important practical uses of the Semantic Web will be in areas with deep wells of interconnected information where meta information is very important to the consumers of the information. For instance, governments and the legal community meet this criterion. Issues such as "where did that document come from," "who asserted that statement" and "what other persons are affiliated with the author of the document" are important in these domains (governmental and legal) and will drive the adoption of the Semantic Web within these communities. Specific applications that allow rapid access to data that has been traditionally difficult to find and correlation of related data will immediately show the benefits of the Semantic Web.

Textual representations of highly "connected" data or large amounts of data are hard for people to comprehend. Hence our interest in finding alternate means of representing data, relationships, etc. The Semantic Web will enable programmatic access of this data, but will quite probably overload the people accessing the data (much as search engines do today). Our approach of alternate visualizations of the data and relationships is just one means of giving the end users of the Semantic Web one more tool for easily extracting meaningful data from the Web in a rapid fashion.

Our current research has led us to agent-related languages such as DAML+OIL and transport technologies such as APEX over BEEP. Although we aren't planning on providing services or data in the near term, we do plan on looking at how those technologies can be organized for efficient, large scale transfer between many points (i.e., agents and avatars in a virtual representation of a physical system).

In support of our work, research and tools in the following areas will be necessary.

- Large-scale storage
- Processing in real-time to common vocabularies
  - inferencing
  - filtering
- Protocols for transferring RDF from point to point
- Protocols and languages for describing who and for what purpose a party makes a statement, etc.

In summary, we see the Semantic Web as a foundation for not only better programmatic access to the elements of the Web, but also the basis for alternate ways to view and interact with the Web.

WiredObjects, Inc.
Mike Lindelsee
Gabe Wachob
Last modified: Wed May 30 22:29:50 PDT 2001

Expertise Matching using RDF

by Ping Liu, Jayne Curson, Peter Dew, and Richard Drew*

Information Research Institute,
University of Leeds, Leeds, LS2 9JT U.K.
pliu@comp.leeds.ac.uk, jaynec@geog.leeds.ac.uk and dew@comp.leeds.ac.uk

**Symularity Ltd*
Leeds Innovation Centre, 103 Clarendon Road, Leeds LS2 9DF U.K.
richarddrew@symularity.com

The Semantic Web aims to make data on the web machine understandable rather than only display the content. This results in all the resources being linked to each other and supports automated discovery services.

We are interested in investigating effective use of Semantic Web technology for Expertise Matching within an academic organization. This is part of a larger project to create an organizational memory for an academic institute. The importance of Expertise Matching has been noted by many researchers [Bannon96, Ackerman98, Bishop00, Cross00, Stewart97, Yimam99, Crandall98]. They realized that more emphasis should be put on the facilitation of collaboration among people through organizational memory, as the tacit knowledge and expertise held by individuals are more important than the explicit documentation. The University of Leeds Research Expertise and Publications Information System (REPIS) is a web-based research information management system that stores information about publications and research projects from a variety of different sources. A component within REPIS is an Expertise Matcher to help people locate experts in the University of Leeds and then facilitate collaboration and knowledge transferring between academia and industries. A new project, called KiMERA, aims to enhance out reach activities between the university and external organisations, and REPIS will be one of a number of existing databases to support this. In collaboration with Symularity Ltd, we are investigating the effectiveness of existing RDF tools to integrate several databases and the advantages that might accrue such as more automated brokering services.

Our belief is that the performance of expertise matching can be improved by making use of Semantic Web technologies. The obstacle of the existing broker system is that the same concept is often expressed in different terms by academia and industry. The Semantic Web will provide the technology to aid the understanding of the concept and help to break down the barriers of communication between different communities.

To understand the current state of the art we have used three research-related databases held by individual academic departments as well as by the central administration organization. We want to integrate all these databases and other semi-structured data and unstructured data (such as personal homepages and technical reports) in order to provide a single access point to the users. In our first experiment, we have integrated the REPIS database, a database of publications from the School of Computing (with sometimes conflicting data) and a database of technical reports. The tools we have used are Protégé, RDFSViz and RDFDB. Protégé is used to create and modify reusable ontologies, RDFSViz is used to visualize the ontologies represented in RDF Schema and RDFDB is used to store and query RDF data.

The experiment showed some advantages of using the RDF model:

- An RDF model greatly improves link capability. Since it is possible for separate data sources to be linked together, RDFDB provides an easy way to navigate the data warehouse. For example, all experts who have the same expertise are linked together, once you have found somebody who has a particular expertise you can easily find other people who also have expertise in the same area.
- An RDF model strongly supports extensibility. A relational database has a very static schema and it is very difficult to make any significant changes without impacting existing code. In comparison, RDF model is dynamic and it is easy to add new information.
- An RDF model provides the ability to integrate different database resources and there is no limitation on these data sources. A relational database can only store structured data. While this is not a restriction in RDFDB, structured data, semi-structured data and unstructured data can be integrated into RDFDB. In this case, REPIS database and webpages can be integrated.

However, some limitations do exist. When the user searches the RDFDB, some of the results are listed as the index of the resource and the user has to take further actions (click each of listed result items) to find more detailed information. There are also problems related to updating data. For example, when new data has been inserted in the original database, it cannot be reflected in the integrated system immediately.

For our RDF system there are two limitations existing that are not due to the RDF model but due to RDFDB itself. (1) RDFDB does not support the RDF Schema which, therefore, makes the ontology described in the Schema irrelevant and (2) RDFDB does not support part matching, so it is impossible to search in the following way: “find all the publications whose title include the word ‘visualization’”. This is limiting our ability to build practical systems.

As an increasing number of relational databases are put on the web, there is an urgent need to integrate all these operational databases. The most important practical use of the Semantic Web is to make the system understand the implicit relationships between these databases and integrate them through the semantic layer. As a result, the databases become semantically linked rather than syntactically linked which could be done in a relational database system.

More research needs to be done to support the use of the Semantic Web: firstly on the interoperability between the different schemas; secondly on how to make the links between the concepts and the terms. We also expect more powerful searching tools for RDF data and schema. The conference will enable us to learn about advances in the Semantic Web technologies and apply them to our expertise matcher and then transfer the learning to Symularity Ltd for commercial exploitation.

References

[Ackerman98]Ackerman, Mark S., Halverson, C., *Considering an Organization’s Memory* 1996; Conference on CSCW’98 pages 39-48, ACM Press, Seattle, WA, 1998

[Bannon96]Bannon, L., Kuuti, K. *Shifting Perspective on Organizational Memory From Storage to Active Remembering* in Proceeding of the HICSS’96, IEEE Computer Press, 1996, page 156-167

[Bishop00]Bishop, K. Heads or Tales: *Can Tacit Knowledge Really be Managed* Proceeding of ALIA 2000 Biennial Conference, 23-26 October, Canberra

[Crandall98]Crandall, N. F., and Wallace, M. J. *Work and Rewords in the virtual Workplace: A "New Deal" for Organizations and Employees* New York: AMACOM - American Management Association, 1998

[Cross00]Cross, R., Baird, L., *Technology Is Not Enough: Improving Performance by Building Organizational Memory* MIT Sloan Management Review Spring 2000 Vol 41, No.3 page 69-78

[Protégé] http://smi-web.stanford.edu/projects/protege/

[RDFDB] http://web1.guha.com/rdfdb/

[RDFSViz] http://www.dfki.uni-kl.de/frodo/RDFSViz/

[Stewart97]Stewart, T. *Tom Stewart on Intellectual Capital* Knowledge Inc., May 1997 available at http://www.brint.com/members/online/120205/stewart.html

[Yimam99]Yimam, D. *Expert Finding Systems for Organizations: Domain Analysis and the DEMOIR approach* ECSCW 99 Beyond Knowledge Management: Management Expertise Workshop, 1999

Project Renaissance: A Semantic Web Based Java Framework for Knowledge Management

Marcus Miguel Maciel
Project Renaissance
Rua Antônio das Chagas 1711
São Paulo, Brazil
+55 11 5183 5488
marcusm@project-renaissance.org

ABSTRACT

Knowledge Management services support universal mapping, discovery, creation and communication of knowledge. Web portals are well-known Knowledge Management services. The Resource Description Framework (RDF) adds Artificial Intelligence to web portals, transforming them into Semantic Webs. RDF enables inference queries to discover knowledge from Semantic Webs. Today, the most well-known form of web portal is the centralized, consumer-oriented hub. In this work, a distributed, collaborative form of a semantic web portal is described, the personal knowledge service. This paper presents Project Renaissance, an open-source Semantic Web based Java framework for Knowledge Management applications, which focus on the human being as a knowledge creator.

Keywords

Knowledge Management, Semantic Web, RDF, portal, Java.

1. INTRODUCTION

Project Renaissance is an open-source effort to build a Java framework for platform independent Knowledge Management applications, including portal, workflow and collaboration services.

The name "Renaissance" is inspired by the creative european movement between the XV and XVI centuries, a cultural rebirth in the arts and sciences powered by the invention of the press. Project Renaissance refers to a similar revolution happening again within the new Knowledge Society, powered by the Internet. Its anthropocentric style shifts the focus of the knowledge technologies from the machine to the human being as the source of innovation. Project Renaissance builds upon the vision that technology should be used to augment human intellectual skills and to support the freedom of expression in a free world.

Project Renaissance emphasizes the production of knowledge (innovation) by offering support for personal productivity hubs. The Project Renaissance model abstracts from the centralized and distributed (peer-to-peer) models with an unified model. This model supports the concept of a personal knowledge server, which is a semantic web server working as a personal media server.

The objetive of Project Renaissance is to create a metamodel for Knowledge Management applications around knowledge maps, leveraging the Resource Description Framework (RDF) and other Semantic Web technologies. This metamodel shall integrate seamlessly with the Semantic Web models at the syntactic and semantic levels.

2. KNOWLEDGE SERVICES

Knowledge is useful information, but information can only be useful if it can be easily accessible. Web portals are considered a great solution to Knowledge Management, and Java offers a richer user experience as well as platform independent portal applications. With the Semantic Web, Java can leverage the web intrisic potential for Knowledge Management as hypermedia with a standard platform for knowledge servers.

A web portal is characterized by a combination of directory services, search engines, personalized content delivery and convenience services. Among the web portal features, one of the most interesting is its application of knowledge maps in directories and search engines. Semantic Web technologies enable smart portals, offering a true Knowledge Management infrastructure to deliver information of better quality.

A knowledge server supports the creation, distribution and universal access to knowledge. A true knowledge server must support the concept of a collective intelligence, which is the union of the knowledge of all the actors within a virtual community, both humans and robots. Knowledge as innovation, is considered the only long term sustainable resource in the Knowledge Society, for it´s unexhaustible and its value increases with the use, and declines with the time. Knowledge is created within communities, being corporations, education sites or other collaboration groups.

The Project Renaissance Knowledge Management Framework offer an unified metamodel for organizing knowledge services which supports and simplifies the development of Knowledge Management applications. This unified metamodel builds upon existing and emerging Semantic Web standards, like RDF, to represent knowledge maps.

3. AN ARCHITECTURE FOR KNOWLEDGE MANAGEMENT

Project Renaissance offers a Java API and an open-source Reference Implementation. Both are free, and the API is implementation independent. There can be other implementations of the API, both commercial and free software.

The Project Renaissance API is meant to support the development of the following kinds of both centralized and distributed Knowledge Management applications:

- digital libraries
- directories
- search engines
- personalized content delivery
- webmail
- discussion forums
- instant messaging
- intelligent whiteboards
- videoconferencing
- peer-to-peer collaboration
- brainstorming and creativity
- schedules
- project management
- web-based CASE tools
- process workflows
- case-based reasoning
- skills management
- e-learning
- knowledge trees

The Project Renaissance API must offer an universal model for organizing knowledge services which support all of these kinds of applications. This unified model builds upon the open standards Java and XML, and it must integrate seamlessly with other Java APIs like J2EE and J2ME, and peer-to-peer standards Jini and JXTA.

The framework is comprised of the following layers: UNION, Service, Channel, and Agent layers. The core layer is the UNION (Universal Object Network). It contains the knowledge metamodel classes, along with the actors hierarchy, which include creators (users) and servants (robots). There's also interest group classes to support the collective intelligence.

The knowledge metamodel is a resource-centric, content-addressable memory model. The concept of "resource" in the Project Renaissance API is not necessarily the same as RDF's. It corresponds roughly to an "object" in the Smalltalk pure object-oriented model. Anything is a resource, even an association. A resource is a composite, it aggregates other resources. This way, one can always ask, "

Resources are containned in contexts, or mindspaces, which can be seen as multidimentional resource pools or knowledge maps or models. Mindspaces are resources too, of course. Resources can reference resources in other mindspaces, either locally using memory references, or globally using URIs.

The UNION layer contains mindspaces, actors and groups. Project Renaissance focus on the human being as a knowledge creator. It represents users as creators (artists) which are supported by servants and participate in collaboration groups with other creators. Creators own mindspaces, servants don't. Mindspaces can be copied and shared by groups, but there's always only one owner for each mindspace instance.

The Service layer plugs service providers to handle communication at the protocol level. It offers an unified view of communication. There is a WebService for HTTP, MailService for SMTP-POP-IMAP, DirectoryService for LDAP-DNS-FileSystem (via JNDI), SearchService, RelationalService (via JDBC), RmiService (for CORBA, EJB), JiniService, JxtaService etc.

Next is the Channel layer. Channels are also known as portlets, or web portal components, but "channel" is a more widely known name and better reflects the concept. Channels implement content formatting and delivery, using XLST. Channels can be minimized, maximized and customized. Minimized channels are just links. Maximized channels are full pages. Customized channels are one or more channels integrated by a personalization channel. The personalization channel applies user layout preferences over the content for presentation.

Content management in Project Renaissance is performed by a controller, in an MVC-like fashion. The controller manages the workflow of the session, activating the channels chosen by the creator. The channels exchange messages with the world through services. Messages contain documents, which are, again, resources.

The Project Renaissance API supports intelligent agents as a kind of servant. Agents are built in the Agent layer. Support for mobile agents is still to be addressed in Project Renaissance. It's not the focus.

4. DISCUSSION

Most of the discussion relate to mindspaces. The greatest challenge in Project Renaissance is to build a metamodel for the Semantic Web. It started with RDF and a content-addressable memory model as a metamodel, the mindspace. Several design choices were considered for a mindspace, namely, a collection of resources, a collection of associations and a collection of resources and associations. Each one has its strength and weakness. A collection of associations was considered for it's similarity with the RDF model, but it's more specific and not resource-centric. A collection of resources and associations was considered for it's a familiar approach to graph modeling and diagram drawing. The choice was a collection of resources because it's the most abstract and resource-centric model.

There are questions about web services, and integration issues with SOAP and XML-RPC, WSDL, UDDI, and others. Also, questions regarding JMS, the Java Messaging Service and JAXM, the Java API for XML Messaging. The position is that Project Renaissance must be as much light-weight as possible. Project Renaissance is not meant to run on a smart card, or on a refrigerator, but it could be interesting to see it running on a PDA, and on a cellular phone, personal communication appliances. Although its modular structure scales for heavy configurations.

In this moment, Project Renaissance is been boostrapped. There is a special requirement that the tool is to be built with itself. The project web portal, www.project-renaissance.org, and the collaboration tools will be built with the evolving prototype. This portal is meant to be many things to many people. Beyond Project Renaissance, it's a portal of Knowledge Management, Semantic Web, and Renaissance Community projects. For the future, it's planned a Renaissance Foundation to support Project Renaissance and its applications in social fields, particularly in education.

Acknowledgements

I would like to thank Bill Joy from Sun Microsystems, Daniel Schwabe from Pontificia Universidade Catolica do Rio de Janeiro, Gina de Oliveira from Universidade Presbiteriana Mackenzie, Valdemar Setzer from Universidade de Sao Paulo, Fernando Giorno from Pontificia Universidade Catolica de Sao Paulo, Vera Vidal from Fundacao Instituto Oswaldo Cruz, and the Renaissance community for their support and encouragement, as well as the following open-source communities for their contribution: Java, Jini, JXTA, OpenOffice, NetBeans, Apache, Mozilla, Linux, BSD, GNU, RDF and Semantic Web.

References

- McClelland, J. L. & Rumelhart, D. E. 1989. Explorations in parallel distributed processing. Cambridge, MA: The MIT Press.
- Levy, Pierre. 1990. Les technologies de l'intelligence. Paris, France: Éditions La Découverte.
- Cox, Brad. 1987. Object-oriented programming. Reading, MA: Addison-Wesley.
- Bigus, J. P. & Bigus, J. 1998. Constructing intelligent agents with Java. New York, NY: John Wiley & Sons.
- Drucker, Peter. 1993. Post-capitalist society. New York, NY: Harper Collins.
- http://www.project-renaissance.org
- http://www.w3.org
- http://java.sun.com

Biography

Marcus Miguel Maciel is a Java consultant and instructor for Sun Microsystems in Brazil. He has 17 years of experience with software development, including 11 years doing Object-Oriented and Knowledge Management research and development, and five years working with Java. He has developed CASE tools, Artificial Intelligence applications, multiplatform GUI architectures, Object-Oriented environments and web portal tools. He has taught courses on Object-Oriented software development, C++, Java and UML, and gives lectures in universities and conferences. His areas of most interest include Knowledge Management and Cognitive Science.

A Machine Learning Perspective for the Semantic Web

SWWS Position Paper

Alexander Maedche

FZI Research Center on Information Technologies
at the University of Karlsruhe,
Haid-und-Neu-Str. 10-14, 76131 Karlsruhe, Germany
http://www.fzi.de/wim
maedche@fzi.de

According to [1] we consider the Semantic Web, as a **Meta-Web**, that is build on the existing WWW. We adopt the point of view that there are two main core **challenges for putting the vision of the Semantic Web in action**:

- *First*, one has to support the re-engineering task of "semantic enrichment" for building the Meta-Web. The success of the Semantic Web greatly depends on the proliferation of ontologies and relational metadata, which requires quick and easy engineering of them and the avoidance of knowledge acquisition bottlenecks.
  Additionally, we consider the task of merging and aligning ontologies for establishing semantic interoperability as an engineering artifact that may be supported by machine learning techniques (see below).

- *Second*, one has to provide means for maintaining and adopting the machine-processeable data that is the basic for the Semantic Web. Thus, we need mechanisms that support the dynamic nature of the Web.

The research area of Machine Learning has a long history as well on knowledge acquisition or extraction as knowledge revision or maintenance and provides a large number of techniques that may be applied to solve the challenges introduced above. Seminal work on combining machine learning with knowledge acquisition and maintenance has shown its practical usefulness (cf. [9]). The Semantic Web opens a wide range of new research challenges for the machine learning community.

We here give a short overview and research challenges with respect to the combination of machine learning research with Semantic Web research:

Extraction of ontologies from existing data on the Web. The task of extracting ontologies is a typical re-engineering task. In general one may roughly distinguish between existing ontologies (such as thesauri, lexical-semantic nets), schemata (such as relational database, web schemata), instances (in data- and knowledge bases), semi-structured data (e.g. in the form of XML documents), natural language documents. Each of these different kinds of data requires its specific import and processing techniques and learning algorithms. To derive ontologies from existing data on the Web a common picture and framework for re-engineering existing data as given in [8] is required. The integration of multiple resources seems to be a promising approach for the difficult task of extracting ontologies form the existing Web data.

Extraction of relational (meta-)data from existing data on the Web. There exists a number of approaches for (semi-)automatically generating relational (meta-)data from existing data on the Web. The active research field of machine learning for information extraction (cf. http://www.dcs.shef.ac.uk/~fabio/ecai-workshop.html) is concerned of the instantiation of so-called templates from natural language text of a restricted domain and structure. Along the same lines techniques for automatic generation of wrappers are researched.

Merging and mapping ontologies by analyzing extensions of concepts. Several approaches supporting the merging and mapping process simulating the behaviour of the ontology engineer have been presented. Machine learning techniques may be used by analyzing extensions of concepts to derive overlapping intensional descriptions. In the area of ontology mapping an approach using supervised classification has been described by [7]. A technique taken from conceptual knowledge discovery (based on the theory of formal concept analysis) for deriving a lattice of concepts as structural description of the overall merging process has been introduced by [10]. Recently, techniques well known from the area of data mining (namely association rules) have been used for the task of catalogue integration [1].

Maintaining ontologies by analyzing instance data. The idea of automatically generating a "T-Box" from given relational instance data is not a new one, e.g. an approach for deriving concept descriptions from instances has been presented in [6]. Recently, the work in this area is described under the name of "A-Box-Mining". A first approach on deriving a taxonomy from a given set of RDF statements has been described in [4]. Similar research has been done by the database community, where the critical problem of discovery of the structure implicit contained in semi-structured data and, subsequently, the recasting of the raw data in terms of this structure has been researched.

Improving Semantic Web applications by observing users. Web Mining applies data mining techniques on the existing web (e.g. Web usage mining analyzes the user behavior or web structure mining explores the hyperlink structure). Machine learning in general may be applied on Semantic Web data to improve existing applications by observing users. We refer the interested reader to an ECML/PKDD'2001 workshop on Semantic Web Mining (http://semwebmine2001.aifb.uni-karlsruhe.de/)

In this position paper we have presented a machine learning perspective for the Semantic Web. Based on the old idea of applying machine learning for knowledge acquisition and maintenance, we have presented several core challenges for the Semantic Web that may be approached by machine learning techniques. An important aspect is that there is a wide range of techniques available from the machine learning community that may directly applied for solving existing problems. Another important aspect is that the Semantic Web opens new research questions for the machine learning community (e.g. dealing with multi-relational data) and provides real-world data (e.g. for inductive logic programming algorithms, cf. [5])

References

[1] R. Agrawal and R. Srikant : On Integrating Catalogs. In: Proceedings of WWW-11, Hong-Kong, 2001. (http://www10.org/cdrom/papers/076/index.html)

[2] T. Berners-Lee. J. Hendler, O. Lassila: The Semantic Web. *Scientific American*, 2001.

[3] S. Decker, S. Melnik, F. van Harmelen, D. Fensel, M. Klein, J. Broekstra, M. Erdmann, I. Horrocks, The Semantic Web: The Roles of XML and RDF. *IEEE Internet Computing*, 4 (5): 63-74, September/October 2000.

[4] A. Delteil, C. Faron-Zucker and R. Dieng: Learning Ontologies from RDF annotations. In: Procedings of the IJCAI'01 Workshop on Ontology Learning OL-2001, Seattle, USA, 2001.

[5] P.A. Flach. Inductive Logic Databases: from extensional to intensional knowledge. In F. Bry, R. Ramakrishnan, and K. Ramamohanarao, editors, 5th International Conference on Deductive and Object-Oriented Databases (DOOD'97), Springer-Verlag, December 1997.

[6] J.U. Kietz and K. Morik: A polynomial approach to the constructive induction of structural knowledge. In: Machine Learning, 1994.

[7] M.S. Lacher and G. Groh: Facilitating the exchange of explicit knowledge through ontology mappings. In: Proceedings of FLAIRS'2001, AAAI Press, 2001.

[8] A. Maedche and S. Staab: Ontology Learning for the Semantic Web. In: *IEEE Intelligent Systems, Special Issue on the Semantic Web,* 16(2), 2001.

[9] G. Webb, J. Wells and Z. Zheng: An Experimental Evaluation of Integrating Machine Learning with Knowledge Acquisition. In: *Machine Learning*, 35(1), 1999

[10] G. Stumme and A. Maedche: FCA-Merge: Bottom-Up Merging of Ontologies. In: Proceedings of the International Joint Conference on Artificual Intelligence IJCAI'01, Seattle, USA, August, 2001

[11] Visser, P.R.S & Tamma, V.A.M., 1999: An Experience with Ontology-based Agent Clustering. In: Proceedings of the IJCAI'99 Workshop on Ontologies and Problem-Solving Methods: Lessons Learned and Future Trends, Stockholm, Sweden, August 2, 1999

Maritime Transportation and the Semantic Web

Raphael Malyankar
Dept. of Computer Science and Engineering
Arizona State University, Tempe, AZ 85287-5406, USA.
E-mail: rmm@acm.org

1 Introduction

My work involves the creation of a computational ontology for maritime information, and the use of this ontology in markup and integrated retrieval from sources of maritime information. Ontological information is acquired from multiple sources, including standards documents, database schemas, lexicons, collections of symbology definitions, and also from semi-structured documents. The computational ontology thus created is being used to create an XML-based markup language (Maritime Information Markup Language — MIML) for tagging information within this domain. This statement describes progress on this project and my views on its relation to the Semantic Web.

The ultimate purpose of this project is to help upgrade maritime information distribution for the Semantic Web. A number of disparate distribution channels are currently being used, ranging from paper documents published once every few years, to radio distribution of warnings and weather forecasts nearly in real-time. Early, accurate, and integrated information is very important to the waterborne transport industry for reasons of safety as well as cost, and there are a number of ongoing efforts concerning upgrading information distribution, but distributing and combining information from different sources is still difficult and resource-intensive. Semantic Web technology should help with these problems, while still allowing leverage of existing distribution channels. The intention is to get information into end-users' hands in a *usable* form, to make it easy for producers of information to 'maintain' (update, revise, or extend) the information they produce, and to make it possible for application developers to write applications that can process it.

The work is part of a larger interest in the representation and distribution of geospatial information, particularly that relating to navigation for all modes of transport. It is related to my view of practical uses for the Semantic Web; I foresee the most important practical use of the Semantic Web in the near future to be the exchange of 'knowledge-at-a-distance' for transportation and logistics, because a large part of this information is not available or cannot be confirmed until shortly before it is needed, and further it may not be in exactly the form desired by the user; but when it does become available, it needs to be integrated into existing plans and local knowledge bases very quickly.

2 Ontology and Markup Language

Ontological knowledge was extracted from various sources, ranging from standards created by standards bodies to semi-structured 'content' documents. Extraction was done by a mixture of semi-automated and human means, and is described elsewhere[1]. The ontology learning stage resulted in a collection of overlapping sub-ontologies, which were merged into one large taxonomy. This taxonomy is currently being used for two purposes: first, as a kind of hierarchical index into a multifarious, distributed, knowledge base consisting of knowledge sources of different kinds — digital charts, web sites, programs, and marked-up text documents; and second, as a basis for defining a markup language for our domain. Reasoning and inference will be investigated soon.

In its 'hierarchical index' role, it is being used to map concepts from the 'user domain' to the 'knowledge source domain', i.e., map concepts to (possibly multiple) information sources. A prototype information

[1] URL: www.isi.edu/dgrc/dgo2001/papers/session-1/malyankar.pdf

```
<Chart>
 <ChartNumber>18773</ChartNumber> <ChartNumber>18772</ChartNumber>
 <Description>San Diego Bay is where California’s maritime history...</Description>
 ...
  <PierArea> ...
   <Pier  name ="Tenth Avenue Marine Terminal">
    <Berth  name="Berths 1 and 2">
     Concrete bulkhead, 1,170 feet of berthing space; 27 feet alongside...
    </Berth>
    <Berth  name="Berths 3 and 6"> ... </Berth>
    <Berth  name="Berths 7 and 8"> ... </Berth>
    ...
</Chart>
```

Figure 1: Marked-up fragment of Chapter 4, Volume 7 of the *Coast Pilot*

retrieval application has been constructed; it retrieves information from different sources in response to a user query; for example, a request for weather information gets weather forecasts from Web sites as well as notes from the *Coast Pilot* (a text document containing information of interest to mariners sailing along the U. S. coastline). URIs are used to link knowledge sources to ontological concepts.

The Maritime Information Markup Language (MIML) is an XML-based language used for text markup in the prototype described. Figure 1 contains an example of its use. Some MIML tags are derived (via the ontology) from direct sources, such as an IHO (International Hydrographic Organization) standard, a lexicon published by NOAA (National Oceanic and Atmospheric Agency), feature classes in a sample Digital Navigation Chart, etc. A second class of tags was also needed to denote information elements within the document, for example the <Chart> tags, which denote sections that pertain to a specific (identified) nautical chart in the NOAA chart numbering system, and the <Description> tag, which is used to denote general text information that cannot be placed into a more specific category. These were invented as necessary. The DTDs (Document Type Definitions) for the Coast Pilot were prepared ‘by hand’.

3 Technological and Research Issues

Some general issues important to my work are the use of Semantic Web technologies for distributing and processing location-dependent information, reasoning about such spatially dependent information, and integrating this information into local knowledge bases, especially for information that needs to be distributed in near-real-time, such as weather warnings. The use of spatial reasoning for navigation of all types also needs exploration, especially reasoning for navigation in a dynamic domain where navigation planning is affected by dynamic processes as well as mobile entities and obstacles, and ‘feature-based’ reasoning, where a navigator must use landmarks and required routes. Further issues are the retrieval and indexing of all kinds of information, ranging from text to pictorial and diagrammatic representations. Last but not least, it is extremely important to create common standards for information markup, and gain the support of governing bodies or organizations; without these, widespread acceptance of Semantic Web technologies will not be possible in these application domains.

Acknowledgements

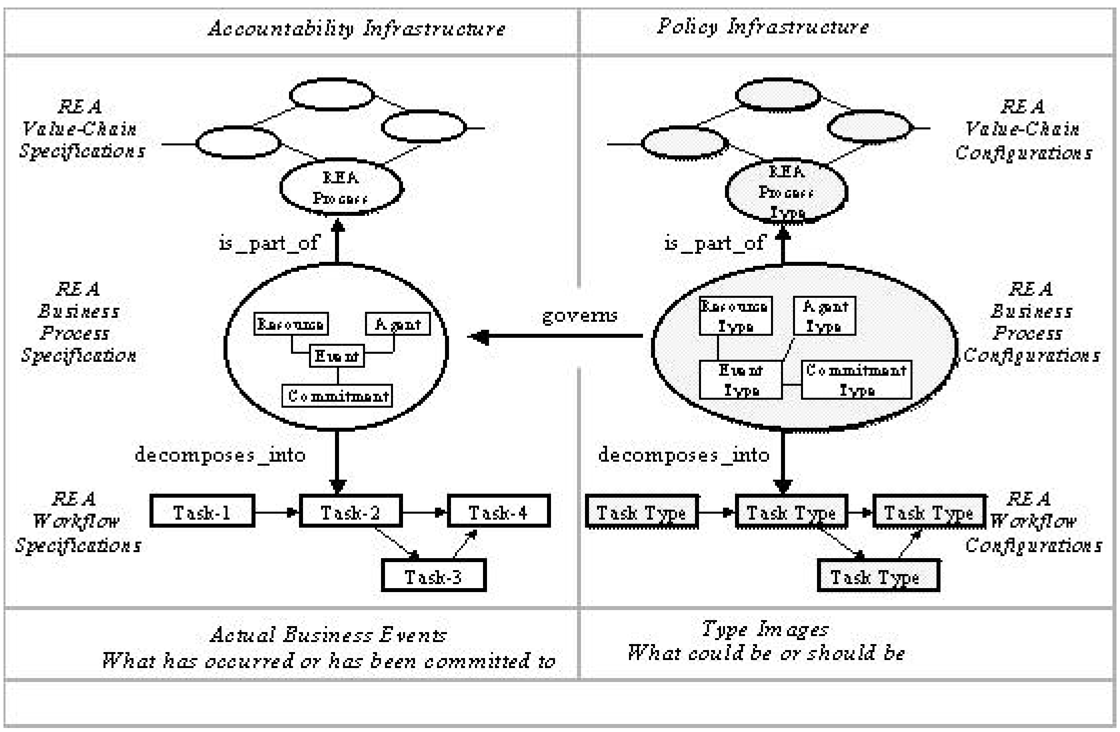
This work was partially supported by the National Science Foundation under grant EIA-9983267, NOAA, Maptech, and the U.S. Coast Guard. The opinions, findings, conclusions and recommendations expressed in this material are those of the author and do not necessarily reflect the views of these entities.

REA Ontology Use in ebXML and the UN/CEFACT Modeling Methodology (N90)

Guido Geerts, University of Delaware, geertsg@be.udel.edu
William McCarthy, Michigan State University, mccarth4@msu.edu

Two ongoing standardization efforts in the area of e-commerce that rely on the notions of semantic content and domain ontologies are (1) the UN/CEFACT TMWG Unified Modeling Methodology (known as UMM or the N090 document) and (2) ebXML (electronic business XML). The UMM is a proposed methodology for object-oriented development of e-commerce solutions that uses multiple layers of abstraction. However, its primary repository of microeconomic semantics is the BRV (Business Requirements View) level that utilizes elements of the REA (Resource-Event-Agent) ontology for its primitive concepts. ebXML is a set of e-commerce standard specifications that is closely intertwined with the UMM and that concomitantly uses REA elements as part of its analysis and design methodology.

An overview of the REA ontology is given in the figure below.



The REA ontology originated in the field of accounting, but its primitives have been extended out to other business enterprise phenomena, and we now estimate that over 60 % of the strongly-typed data in a typical company (its internal ERP systems, its external EDI segments, etc.) can be interpreted (somewhat passively) within REA.

At the Semantic Web Workshop, we want to explore with other theorists and practitioners the possibilities for extending this parsimonious enterprise ontology into a more active environment wherein its components can be used for automated comprehension and intensional reasoning. This seems to be where the semantic web in general is headed, and we want to avail ourselves of the workshopís assembled expertise to chart our course. Because of ebXML infrastructure time constraints, REA is embedded somewhat passively as an analysis guideline and worksheet template at present, and we need to see how semantic web ideas can aid us in getting it embedded as a run time component for reasoning about the economic, accounting, and legal contracting aspects of business exchanges and transformations, both within and between business enterprise partners.

Description Logics Emerge from Ivory Towers

Deborah L. McGuinness
Stanford University, Stanford, CA, 94305
dlm@ksl.stanford.edu

Abstract: *Description logic (DL) has existed as a field for a few decades yet somewhat recently have appeared to transform from an area of academic interest to an area of broad interest. This paper provides a brief historical perspective of description logic developments that have impacted their usability beyond just in universities and research labs and provides one perspective on the topic.*

Description logics (previously called terminological logics and KL-ONE-like systems) started with a motivation of providing a formal foundation for semantic networks. The first implemented DL system – KL-ONE – grew out of Brachman's thesis [Brachman, 1977]. This work was influenced by the work on frame systems but was focused on providing a foundation for building term meanings in a semantically meaningful and unambiguous manner. It rejected the notion of maintaining an ever growing (seemingly adhoc) vocabulary of link and node names seen in semantic networks and instead embraced the notion of a fixed set of domain-independent "epistemological primitives" that could be used to construct complex, structured object descriptions. It included constructs such as "defines-an-attribute-of" as a built-in construct and expected terms like "has-employee" to be higher-level terms built up from the epistemological primitives. Higher level terms such as "has-employee" and "has-part-time-employee" could be related automatically based on term definitions instead of requiring a user to place links between them. In its original incarnation, this led to maintaining the motivation of semantic networks of providing broad expressive capabilities (since people wanted to be able to represent natural language applications) coupled with the motivation of providing a foundation of building blocks that could be used in a principled and well-defined manner. KL-ONE provided an important first step in description logic history and since then, many systems have been designed and implemented taking differing positions on the requirements of expressive power, completeness of reasoning, and tractability of reasoning. One early paper describing some description logic systems up to 1990 is provided in [MacGregor, 1991] and includes some of the earlier implemented and used systems such as BACK [Peltason, 1991], CLASSIC [Brachman et al, 1989], and LOOM [MacGregor, 1991]. K-REP [Mays et al, 1991], although not included in that article, also appeared in the same time frame.

Some of these early systems are interesting from the perspective of DLs emerging from ivory towers since one (BACK) made an attempt to be the basis of a company, another (K-REP) was the basis of a domain-specific commercial application in medical information systems which spun out of IBM, another (CLASSIC) was the basis of a family of some commercially fielded applications in the areas of data mining (IMACS [Selfridge-et-al, 1993]), knowledge-enhanced search (FindUR [McGuinness-et-al, 1998; McGuinness,2000]), and a family of configurators fielded at AT&T and Lucent that were deployed for over a decade(PROSE/QUESTAR[McGuinness-Wright, 2000]). Another (LOOM) was also was used extensively in a number of government research and application programs. Some of these (and other) early systems have had success moving from their roots in universities or industrial research labs into use in fielded (e.g. [Brachman et al, 1999], [Rychtyckyj, 1996]) applications and provide good examples of use in practice for the description logic-based applications of today.

These early systems however typically sacrificed something (usually expressive power, but sometimes completeness) in order to maintain some forms of usability (typically efficiency but sometimes understandability). The more recent set of implemented description logics are expressive (at least with respect to concept reasoning) and also maintain complete reasoners with computationally efficient implementations. A few examples of implemented description logics in this class today are DLP [Patel-Schneider,1999], FACT [Horrocks,1998], and RACE [Haarslev-Moeller, 1999]. These systems are interesting since they do not need to limit the number of "epistemological primitives" as much as earlier usable description logics did in order to maintain a handle on computational efficiency of reasoning. Thus they can support certain applications that need more expressive power along with guaranteed deductive closure of reasoning with efficiency. While work such as [Horrocks-Patel-Schneider, 1999] that discusses efficiency of description logic reasoning has facilitated a broader range of possible applications using today's DLs, arguably, this was not enough to really draw description logics out into the mainstream. Similarly, while work providing environments to support DL-usage also arose, such as Ontosaurus (http://sevak.isi.edu:8950/ploom/shuttle.html), and usability learnings were compiled, such as [McGuinness-Patel-Schneider, 1999], and supporting materials such as tutorials became available, such as (http://www.bell-labs.com/project/classic/papers/ClassTut/ClassTut.html), these were useful but arguably also not enough to draw description logics into mainstream usage. Similarly, although description logics saw maintained interest in a few application communities such as configuration with PROSE and Ford's system, databases with continuous KRDB (http://sunsite.informatik.rwth-aachen.de/Societies/KRDB/) workshops since 1994, and medicine (e.g., [Rector et al, 1996], [Mays et al, 1997]), arguably this also was not enough to make the order of magnitude increase in interest in description logics and really pull them out of academic settings into the mainstream commercial world.

One progression that may be of most interest to those viewing description logic's movement into more mainstream use is its progression into web usage. Arguably, this is the single use that has drawn description logics out of ivory towers more than anything else. Some communities recognized that description logics, with its long researched area of formal foundations for structured knowledge representation formalism, might be just the thing that web languages, such as XML and RDF(S) [Lassila-Swick, 1999][Brickley-Guha, 2000], could benefit from. The merging of the goals from frame-based systems of usability, from web languages of broad web usage, and from description logics of formal foundations for extensible, semantically understood systems led to efforts such as OIL [Fensel-et-al, 2001]. OIL may epitomize the effort to take DLs to the web. Most recently the OIL work was used when the same combination of goals emerged for the web language [McGuinness et al, 2001] and (http://www.daml.org/2000/10/daml-ont.html) for the Darpa Agent Markup Language program [Hendler-McGuinness, 2000] and (http://www.daml.org/about.html).

This program has a goal of facilitating the next generation web. The resulting DAML+OIL language now provides a foundation on which web applications can be built that is compatible with the emerging web standards of XML and RDF(S) and provides the formal foundations for unambiguous specification of term meanings.

There appear to be many forces that may be supporting the transition of description logics into more mainstream usage. The World Wide Web Consortium (W3C www.w3c.org) is arguably the strongest force in web standards and it now supports a semantic web activity (http://www.w3.org/2000/01/sw/Overview.html). The language for the DAML program – thus a description logic-inspired language – is expected to be the initial proposal for the web ontology language to be worked on through W3C. Additionally, many corporations are acknowledging that ontologies are central to their knowledge-oriented applications [McGuinness, 2001]. Essentially every e-commerce application, whether from a somewhat recently formed company such as VerticalNet or from a more established bricks and mortar company, such as Dell, has some ontological information stored behind its applications. Some of us who consult to companies on knowledge representation and reasoning applications, such as representation for e-commerce, are finding that CEOs and marketing directors are the people who are calling to explore the types of ontology-based applications that might be included in commercial products. Also, some venture capitalists are becoming knowledgeable and interested in the field. This is a stark contrast to the recent past when calls, if they came, typically came from technologists. Additionally applications of today and projected applications for the future appear to require more inferential power than past ontology-based applications such as simple taxonomy-based applications like Yahoo. Many people are looking for the "smarter" applications of tomorrow that will make some deductions for the user. This may provide exactly the requirements that not only allow description logics to shine, but also provide challenges to simpler "knowledge management" approaches.

In summary, description logic's history of emphasis on formal foundations may have been the thing that kept it (and its literature) from emerging into the mainstream in the past because a plethora of formal papers may have appeared daunting to prospective readers/users. Today however, the needs of emerging applications, such as those appearing on the web have motivated people to look for foundations on which long-lived and extensible applications may be built. Thus the fact that description logics are strong in formal foundations may now be the thing that is supporting its emergence into the broader world of web applications and other application areas. This, along with the tangible results of early DL applications, work such as reasoning efficiency that is now embodied in today's implemented systems, learnings of usability efforts, and finally, and potentially most importantly, efforts such as OIL and DAML+OIL may be putting description logics in a place where they can find commercial need, acceptance, and demand.

Acknowledgments

While any errors in presentation are my own, my description logic perspective has been greatly enhanced by my long collaboration with my CLASSIC co-authors [Borgida, 1989] Borgida, Brachman, Patel-Schneider, and Resnick, more recent collaborations with my OIL co-authors [Bechhofer et al, 2000], [Fensel et al, 2001], Bechhofer, Broekstra, Decker, Erdmann, Fensel, Goble, van Harmelen, Horrocks, Klein, Motta, Staab, and Studer, and my many DAML+OIL co-authors from the joint committee (http://www.daml.org/committee/) and DAML colleagues (not previously mentioned) including Berners-Lee, Connolly, Dean, Decker, Fikes, Hayes, Hendler, Lassila, McIlraith, and Stein. Also, influencing this position are a number of applications of DL and ontology technology including many done at AT&T, (e.g., [McGuinness-Wright, 2000], [McGuinness, 1999], [Brachman-et-al, 1993]), Stanford [McGuinness et al, 2000], Cisco, and VerticalNet [Das et al, 2001].

References

Franz Baader and B. Hollunder. `KRIS: Knowledge representation and inference system', *SIGART Bulletin*, 2(3), pp 8-14, 1991.

Sean Bechhofer, Jeen Broekstra, Stefan Decker, Michael Erdmann, Dieter Fensel, Carole Goble, Frank van Harmelen, Ian Horrocks, Michel Klein, Deborah L. McGuinness, Enrico Motta, Peter F. Patel-Schneider, Steffen Staab, and Rudi Studer. ``An informal description of OIL-Core and Standard OIL: a layered proposal for DAML-O''. Posted at http://www.ontoknowledge.org/oil/downl/dialects.pdf .

B rachman, R., *A Structural Paradigm for Representing Knowledge,* Ph.D. thesis, Harvard University, USA, 1977.

Ronald J. Brachman, Alex Borgida, Deborah L. McGuinness, and Peter F. Patel-Schneider. "Reducing" CLASSIC to Practice: Knowledge Representation Theory Meets Reality. In Artificial Intelligence 114(1-2) pages 203-237, October, 1999. http://www.ksl.stanford.edu/people/dlm/papers/aij99-abstract.html.

Ronald J. Brachman, Deborah L. McGuinness, Peter Patel-Schneider, Lori Alperin Resnick, Alex Borgida. `Living with CLASSIC: When and how to use a KL-ONE-like language', *Principles of Semantic Networks* ed, Sowa, Morgan Kaufmann, 1991.

Ronald J. Brachman, Peter G. Selfridge, Loren G. Terveen, Boris Altman, Alex Borgida, Fern Halper, Thomas Kirk, Alan Lazar, Deborah L. McGuinness, Lori Alperin Resnick. ``Integrated Support for Data Archaeology.'' In *International Journal of Intelligent and Cooperative Information Systems*, **2**:2 1993, pages 159--185.

Dan Brickley & R.V.Guha, "Resource Description Framework (RDF) Schema Specification 1.0", W3C Candidate Recommendation 27 March 2000, World Wide Web Consortium, Cambridge (MA); available on-line as http://www.w3.org/TR/rdf-schema/.

Alex Borgida, Ronald J. Brachman, Deborah L. McGuinness, and Lori Alperin Resnick. ``CLASSIC: A Structural Data Model for Objects'', *Proceedings of the 1989 ACM SIGMOD International Conference on Management of Data*, Portland, Oregon, June, 1989, pp. 59--67.

Aseem Das, Wei Wu, and Deborah L. McGuinness. ``Industrial Strength Ontology Management''. Stanford Knowledge Systems Laboratory Technical Report KSL-01-07 2001. In the Proceedings of the International Semantic Web Working Symposium, Stanford, Ca, July, 2001. http://www.ksl.stanford.edu/people/dlm/papers/ontologyBuilderVerticalNet-abstract.html.

Dieter Fensel, Ian Horrocks, Frank van Harmelen, Deborah L. McGuinness, and Peter F. Patel-Schneider. ``OIL: An Ontology Infrastructure for the Semantic Web ''. In IEEE Intelligent Systems, Vol. 16, No. 2, March/April 2001. http://www.computer.org/intelligent/ex2001/x2038abs.htm

Richard Fikes and Deborah L. McGuinness. ``An Axiomatic Semantics for RDF, RDF Schema, and DAML+OIL'' KSL Technical Report KSL-01-01, 2001.

Enrico Franconi. `CRACK', *Proceedings of the International Workshop on Description Logics - DL-98*, pp 58-59 , Trento, Italy, 1998.

V. Haarslev, Ralf Moeller. `RACE System Description', *Proceedings of the International Workshop on Description Logics - DL-99*, pp 130-132, 1999.

Jim Hendler and Deborah L. McGuinness. The DARPA Agent Markup Language. *IEEE Intelligent Systems* 16(6): 67-73. 2000. http://www.ksl.stanford.edu/people/dlm/papers/ieee-daml01-abstract.html.

Ian Horrocks. `FaCT and iFaCT', *Proceedings of the International Workshop on Description Logics - DL-99*, pp 133-135, 1999.

Ian Horrocks and Peter Patel-Schneider. `Optimizing description logic subsumption', *Journal of Logic and Computation*, Vol 9(3), pp 267-293, 1999.

Ian Horrocks, Alan Rector, and Carole Goble. A Description Logic Based Schema for the Classification of Medical Data, *Knowledge Representation Meets Databases - KRDB-96,* 1996.

Ora Lassila & Ralph Swick, "Resource Description Framework (RDF) Model and Syntax Specification", W3C Recommendation 22 February 1999, World Wide Web Consortium, Cambridge (MA); available on-line as http://www.w3.org/TR/REC-rdf-syntax/.

Robert MacGregor. *The Evolving Technology of Classification-based Knowledge Representation Systems, Principles of Semantic Networks,* ed, Sowa, pp 385-400. Morgan Kaufmann, Los Altos, CA, 1991.

R obert MacGregor, Inside the LOOM Description Classifier', *SIGART Bulletin*, 2(3), pp 88-92, 1991.

Eric Mays, Robert Dionne, and R Weida. *K-Rep System Overview*, *SIGART Bulletin*, Vol 2(3), pp 93-97, June 1991.

Eric Mays, R. Weida, Robert Dionne, M. Laker, B. White, C. Liang, and F. J. and Oles. Scalable and expressive medical terminologies. Proceedings/AMIA Annual Fall Symposium 1996; 259-263.

Deborah L. McGuinness. *Ontologies Come of Age*. To appear in Dieter Fensel, Jim Hendler, Henry Lieberman, and Wolfgang Wahlster, editors. The Semantic Web: Why, What, and How, MIT Press, 2001. http://www.ksl.stanford.edu/people/dlm/papers/ontologies-come-of-age-abstract.html.

Deborah L. McGuinness. Ontology-enhanced Search for Primary Care Medical Literature. In the Proceedings of the International Medical Informatics Association Working Group 6- Medical Concept Representation and Natural Language Processing Conference, Phoenix, Arizona, December 16--19, 1999. http://www.ksl.stanford.edu/people/dlm/papers/imia99-abstract.html .

Deborah McGuinness, Richard Fikes, James Rice, and Steve Wilder. An Environment for Merging and Testing Large Ontologies. *Principles of Knowledge Representation and Reasoning: Proceedings of the Seventh International Conference (KR2000).* A. G. Cohn, F. Giunchiglia and B. Selman, editors. San Francisco, CA, Morgan Kaufmann Publishers. http://www.ksl.stanford.edu/people/dlm/papers/aaai00-abstract.html.

Deborah L. McGuinness, Richard Fikes, Lynn Andrea Stein, and James Hendler. ``DAML-ONT: An Ontology Language for the Semantic Web ''. To appear in Dieter Fensel, Jim Hendler, Henry Lieberman, and Wolfgang Wahlster, editors. The Semantic Web: Why, What, and How, MIT Press, 2001. http://www.ksl.stanford.edu/people/dlm/papers/daml-ont-abstract.html.

Deborah L. McGuinness and Peter F. Patel-Schneider. ``Usability Issues in Knowledge Representation Systems''. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, Madison, Wisconsin, July, 1998. This is an updated version of ``Usability Issues in Description Logic Systems'' published in *Proceedings of International Workshop on Description Logics*, Gif sur Yvette, (Paris), France, September, 1997.

Deborah L. McGuinness and Jon Wright. ``An Industrial Strength Description Logic-based Configurator Platform''. *IEEE Intelligent Systems*, Vol. 13, No. 4, July/August 1998, pp. 69-77.

Peter Patel-Schneider. `DLP System Description', *Proceedings of the International Workshop on Description Logics - DL-98*, pp 87-89, Trento, Italy, 1998.

Chris Peltason. `The BACK System - An Overview', *SIGART Bulletin*, Vol 2(3), pp 114-119, 1991.

Alan Rector, S. Bechhofer, Carole Goble, Ian Horrocks, W. A. Nowlan, and W. D. Solomon. The GRAIL concept modelling language for medical terminology. Artificial Intelligence in Medicine 1997; 9:139-171.

Nestor Rychtyckyj. `DLMS: An Evaluation of KL-ONE in the Automobile Industry', Proceedings of the International Knowledge Representation and Reasoning Conference, pp 588-596, 1996.

SWWS Position Paper

Bryan Pelz, CEO, Fetch Technologies (pelz@fetch.com)
Dr. Steven Minton, CTO, Fetch Technologies (minton@fetch.com)

Today, the Internet is made up of countless distributed and autonomous sources that contain valuable information, but present it for human consumption. Humans get by with only informal conventions to facilitate communication, and normally do not require detailed, pre-specified standards for information exchange. In contrast, machines must communicate in a fashion that conforms to syntactic and semantic standards that have been carefully worked out in advance.

Fetch Technologies has developed a set of tools for accurately and reliably extracting data from websites and transforming it into a structured data format, such as XML. In the process, the data can be normalized and aligned to an arbitrary ontology, making it available to software agents. These tools provide a bridge between the Human Web and the Semantic Web. Today, this bridge is useful in providing a critical mass of information sources to the Semantic Web. In the future, these tools will allow agents access to a richer, fuller information context.

Key to these tools is the use of machine learning techniques to access, integrate, and transact with Human Web sources. This provides the scalability necessary to learn highly accurate extraction rules, to verify wrapper functioning, to automatically adapt to website changes, and to integrate disparate extracted data.Ý Furthermore, because our machine learning technology enables software agents to be trained to automatically recognize and extract semi-structured content, it provides a practical means for bootstrapping the Semantic Web. We believe that this type of bootstrapping is necessary to establish the critical mass required for broad adoption of the Semantic Web.

Toward a knowledge portal in environmental health: taking advantage of the semantic web

Bernard Moulin, Dr. Professor
Computer Science Department and Geomatics Research Centre,
Laval University, Pouliot Building
Ste-Foy, Québec G1K 7P4 , Canada,
Phone:1- (418) 656-5580,
Email: Moulin@ift.ulaval.ca

To reduce and guard against health risks of environmental origin requires easy and rapid access to high quality statistics and information which must be analysed in a useful manner to support decisions and interventions. The problems that environmental and public health officials (EPHO) must confront are extremely diverse. Sources of contaminants may be local, regional, or planetary and may be measured in humans or in several other vectors. The knowledge needed by EPHOs is also very diverse (medecine, environment, chemistry, etc.) and the sources of information are varied (data bases, books, proceedings, journals, radio recordings, video recordings, etc.). When confronting a new case (ex: during the winter season several meningitis cases identified in Quebec city area), EPHOs must act rapidly (ex: decide whether to launch a vaccination campaign or to wait for the detection of other cases, inform parents, etc.). To this end s/he must get the best available information that will help make a decision. The Internet is a very abundant source of information that health professionals use everyday. However, as most professionals trying to search information on the Internet, they are confronted to the problem of getting back from search engines huge lists of URLs, most of them being irrelevant to their requests.

Considering the specialized domain of environmental health, we developed an initial version of a domain ontology with the help of senior EPHOs and domain specialists. We developed a tool (a kind of meta-search engine) that uses this ontology to help users create precise queries in terms of conjunctions of domain expressions. These queries are submitted to search engines such as Google. The search results are filtered based on an analysis of the content of the descriptions of URLs (for instance the small summaries that Google associates to URLs) in order to verify how well they match with the content of the initial requests: do they contain the expressions used in the user's query, in which order? How close are they of each other?

In order to improve the filtering process, it would be very useful to compare the semantic representations of the query and of the URL descriptions. Such semantic representations are not available yet and we work on techniques that exploit lexico-semantic knowledge to analyze the words the user's expressions found in the URL description as well as their relationships. In some ways we try to use lexical knowledge to approximate semantic knowledge. But, this is a temporary solution. In the future we will need more elaborate semantic descriptions of document summaries.

Another problem is related to multimedia documents (audio, video documents). How to index them and how to retrieve them on the basis of their semantic content. The MPEG7 proposal which provides guidelines to index multi-media documents is still uncomplete

and has not yet been largely adopted. It would provide some ways to specify the semantic content of multi-media documents. We need search engines which take advantage of these semantic representations of documents.
These various indexing and search facilities will be integrated in a Knowledge Portal devoted to environmental health. Our plan is to create a repository of URLs relevant to EPHOs organized into categories corresponding to the domain ontology. This repository will be created and continuously updated by a population of agents that will query domain specific search engines, gather the results and filter them according to the URL repository content. Other agents (crawlers) will explore the web seraching for new categories that might be relevant to EPHOs. Such an URL repository will guarantee users precise results, excellent response time, notification of new entries that they have not consulted yet, as well facilities to generate summaries of relevant documents.

When semantic content will be available for documents accessible on the web, it will be interesting to measure the gains of efficiency of searches and indexing over classical methods.

The Conceptual Web – our research vision.

Ambjörn Naeve <amb@nada.kth.se>
Mikael Nilsson <mini@nada.kth.se>
Matthias Palmér <matthias@nada.kth.se>
Centre for user oriented IT Design (CID),[*]
Royal Institute of Technology, Stockholm

Abstract

We present a vision of an extension of the emerging semantic web into what we call a *Conceptual Web*, where the semantics is not only machine-understandable, but also available for the user in an appealing form, which creates substantial benefits in terms of overview and clarity. We are using visual modeling in UML and a technique called *conceptual browsing* to present the conceptual web to the user. This construction lives on top of the ordinary semantic web and thus shares the advantages of RDF, such as distributivity and scalability.

1 Introduction

The stated goal of the semantic web is to enable machine understanding of web resources: "One of the major obstacles [...] has been the fact that most information on the Web is designed for human consumption [...] the structure of the data is not evident to a robot browsing the web"[1]. The rationale behind this goal has been that deriving meaning from contemporary HTML or other web resources is nearly impossible due to the lack of a common meta-data framework for describing resources. In fact, most resource descriptions today are in the form of HTML text in a containing document. While such semantic descriptions are meaningful only to the human reader, the semantic web will provide such descriptions in machine readable format.

2 Conceptual (un)clarity on the Web

However, it is not at all evident that such machine readable semantic information will be clear and effective for human interpretation. The hyper-linked structure of the web presents the user with a totally fluid and dynamic relationship between context and content, which makes it hard to get an overview of the conceptual context within which the information is presented. As soon as you click on a hyperlink, you are transferred, helplessly, to a new and often unfamiliar context. This results in the all too well-known "surfing-sickness" on the web, that could be summarized as "Within what context am I viewing this, and how did I get here?" The conclusion we draw is that *extracting usable meaning from web pages is often as difficult for a human reader as it is for a machine.* This strongly suggests that there is a need for a human-understandable semantics for web resources as well.

This form of semantics becomes even more important within the emerging field of e-learning. In a learning context, the conceptual structure of the content is an essential part of the learning material. Losing the contextual information of the content means more than just "surfing-sickness". It means that you will not be able to contextually integrate the concepts that you are trying to learn, which is vitally important in order to achieve an understanding of any specific subject area.

The semantic web initiative, as it looks today, does not provide such a semantics. It provides descriptions of web resources, but no way to present them to the user in a contextually clear way. There are initiatives, such as topic navigation and visual history browsers, that try to address this problem, but they fail miserably in giving the necessary overview of the conceptual context.

3 The Conceptual Web

In order to solve this problem, we are working on ideas to extend the semantic web in order to provide not only semantic information for the machine, but also conceptual information for the human user. This form of extended semantic web, which we call the *Conceptual We*b, is a long-term vision with many parts:

- **RDF and RDF Schema** provide the underlying model and representation. We also use standard RDF vocabularies such as Dublin Core and IMS/IEEE LOM (the RDF binding available in IMS metadata 1.2[2] was constructed by us). The addition of ontology layers such as OIL[3] is of course also a fundamental part of resource description on the web.

[*] `http://cid.nada.kth.se/il`
[1] `http://www.w3.org/DesignIssues/Semantic.html`
[2] `http://www.imsproject.org/metadata`
[3] See `http://www.ontoknowledge.org/oil`

- We will participate in the upcoming **Edutella** effort[4], which aims to produce a distributed meta-data network to serve as an underlying meta-data infrastructure for (primarily educational) applications. Edutella will provide standardized access to distributed meta-data query and search facilities with reasoning capabilities, an important part of a fully-functioning semantic web[5].

- The fundamental building block of the conceptual web is **conceptual modeling**, which provides a human-understandable semantics for both abstract ideas and concrete resources. We use the Unified Modeling Language (UML) for conceptual modeling, which provides a well-proven and standardized vocabulary for this purpose. Unfortunately, the relationship between RDF and UML is still rather unclear[6]. We strongly support the forces that try to refactor UML in order to achieve a more precise meta-model[7], as well as the efforts to merge/combine RDF and UML[8]. We regard these strategic efforts as necessary prerequisites for building the Conceptual Web.

- Using the above technologies, we are designing the Conceptual Web as a ***knowledge manifold***[9]. A knowledge manifold is an educational architecture, developed at CID, that provides an overall strategy for the construction, management and use of well-defined contexts for distributed content.

- One of the fundamental tools of the conceptual web is a new type of knowledge management tool which we call a ***concept browser***[10].This tool allows the user to browse conceptual contexts in the form of concept maps (typically UML diagrams) with rich annotations. Thus the full power of visual modeling is combined with the distributivity and universal annotation property of RDF into a hyper-linked web of conceptually clear material. This combination gives the user a clear overview of the subject area (= context), while at the same time allowing the exploration of its various forms of content. Incorporating web resources as content is done similarly to ISO Topic Maps[11], or the Conceptual Open Hypermedia system[12], in that content is linked to concepts in the conceptual web, with the important added benefit of a clear and browsable visual overview of the context. Combined with our form of visually configurable query/search/filter engines this results in a new and revolutionary web experience.

  Our first incarnation of a concept browser is called **CONZILLA**[13], and has been developed as an open source project at CID over the last three years. It is proving to be a very valuable tool for providing an overview of complex web-based material. Using Conzilla, several instances of knowledge manifolds are presently under construction at CID, e.g. within the fields of mathematics, e-adminstration, IT-standardization and interoperability between different systems for e-commerce[14]. Conzilla also has the potential to become a very useful and visually pleasing presentation tool for any kind of RDF data.

- An added benefit of using the semantic web as a basis for the conceptual web is **application-independence**. Just as the semantic web gives the machine (software agents and applications alike) a sort of “sixth sense” about the meaning of web resources, the conceptual web gives the human user a sixth sense about the conceptual context and the underlying meaning of the current situation, which is independent of the currently used application. We are therefore studying ways to introduce the conceptual web into other environments. Apart from their usage on the ordinary web, we are investigating the fascinating possibility of introducing conceptual structures in **3D environments**. A 3D environment filled with semantics and conceptual structures would present a fundamentally different experience, enabling for the first time a virtual reality full of meaning, and not only packed with dead 3D objects whose meaning is defined by the graphics engine. This semantics could even be accessed from outside such an environment, making the 3D environment fully transparent.

4 Conclusions

The Conceptual Web is a powerful idea which has yet to become a reality. However, several of the important tools and technologies already exist. The remaining obstacles include a fully working semantic web infrastructure with mature vocabularies and tools as well as the availability of conceptual modeling constructs on top of the semantic web, which would be enabled by the UML/ RDF vocabulary.

---

[4] a collaboration between Stanford Infolab, CID and KBS in Hannover, within the Wallenberg Global Learning Network.
[5] A sort of Semantic Web Bus, see `http://www.technetcast.com/tnc_play_stream.html?stream_id=459`
[6] See `http://www.w3.org/TR/NOTE-rdf-uml/`
[7] See `http://www.cs.york.ac.uk/puml/`
[8] See `http://www-db.stanford.edu/~melnik/rdf/uml/`, `http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Cranefield/`.
[9] See `http://cid.nada.kth.se/pdf/cid_52.pdf` and `http://cid.nada.kth.se/pdf/cid_17.pdf`
[10] Described in `http://cid.nada.kth.se/pdf/cid_52.pdf`
[11] See `http://www.topicmaps.net/`
[12] See `http://inanna.ecs.soton.ac.uk/cohse/`
[13] See `http://www.conzilla.org/`
[14] See `http://www.cenorm.be/isss/Workshop/ec/Projects.htm#ECIMF`

Management of Data, Knowledge, and Metadata on the Semantic Web: Experience with a Pharmacogenetics Knowledge Base

Diane E. Oliver, Micheal Hewett, Daniel L. Rubin,
Joshua M. Stuart, Teri E. Klein, and Russ B. Altman
Stanford Medical Informatics, Stanford University, Stanford, CA
oliver@smi.stanford.edu, hewett@smi.stanford, rubin@smi.stanford.edu,
stuart@smi.stanford.edu, klein@smi.stanford.edu, altman@smi.stanford.edu

Biomedical researchers are decoding the human genome with astonishing speed, but the clinical significance of the massive volumes of data collected remains largely undiscovered. Progress requires communication and data sharing among scientists. These data may be in the form of (1) raw data, derived data, and inferences that result from computational analyses, or (2) text documents published by experts who present their conclusions in natural language. The World Wide Web provides a valuable infrastructure for enabling researchers to share the rapidly growing knowledge about biology and medicine, and a fully functional Semantic Web is necessary to support data submission and retrieval, the sharing of knowledge, and interoperation of related resources.

We are working with scientists from multiple medical research centers to build a pharmacogenetics knowledge base that is publicly accessible on the Web, called PharmGKB (http://www.pharmgkb.org). The collaborative nature of our work demands building consensus data models for data submission and data retrieval. We need techniques that make it possible for both humans and machines to make sense of the data, methods for storing the data in a carefully designed knowledge base that contains up-to-date domain knowledge, and approaches for using metadata standards to express requirements for submitted data and inferred data.

The distinction between data and knowledge is often blurred, but for our purposes, **data** are the elements of information collected by researchers for experimental studies in a laboratory or clinical setting. Scientists submit these data to PharmGKB. PharmGKB is a knowledge base that is implemented in Protégé, a frame-based knowledge-management system (http://protégé.stanford.edu). PharmGKB contains an ontology comprised of clinical, pharmacologic, and biological knowledge. The domain concepts are arranged in a hierarchy of classes. Each class has slots that denote relations to other classes. Experimental data are represented by instances in the knowledge base. Thus, for our purposes, **knowledge** is conceptual information stored in the PharmGKB class-slot structure that depicts the domain. Finally, **metadata** is information about experimental data and derived data that may be imported to and exported from PharmGKB.

The architecture we have developed for PharmGKB involves seven components:

(1) **Data-entry forms** that allow people to visualize subsets of the data model and enter data
(2) An **XML processing system** that allows data to be submitted in a semantically-tagged machine-processable form
(3) A **knowledge base** that stores submitted data linked to the domain ontology
(4) An **ontology-editing environment** that developers use to extend the knowledge-base into new subdomains
(5) A **knowledge-base browsing and query tool** that permits users to view the contents of the knowledge base and to perform queries across sets of data

(6) **Standard interfaces** that enable applications to communicate with PharmGKB and that facilitate interoperation with other Web resources
(7) A **production system** and a **development system** that coexist and that must be merged on a periodic basis

The knowledge base is the central resource of the PharmGKB system that is accessible to users and applications over the Web. Some users prefer to submit data through Web forms because the forms guide them in entering certain data by hand, and for large datasets, they can upload tab-delimited files. After a user enters data through Web forms, the PharmGKB system automatically generates XML files from the data entered. However, other users, who know how to create XML files themselves, would rather avoid data entry through Web forms and prefer to submit XML files directly. In either case, the data requirements must be clear, and we have developed an XML schema that expresses metadata for data input.

The XML schema for data input must have a well-defined mapping to the domain ontology in the knowledge base, and it also must be consistent with the design of the Web forms. Thus, there are three levels of data representation that must be compatible: (1) the HTML form elements used in the user interface, (2) the XML schema, and (3) the ontology. We are working toward developing automated ways for keeping these three representations synchronized as the system evolves over time.

Evolution of the PharmGKB domain ontology is expected. When a standard relational database has a relatively stable database schema, evolution is primarily a matter of adding, deleting, and updating instance-level data, or field entries in the tables. In contrast, in PharmGKB, not only does instance-level data change as users submit data, but the PharmGKB ontology also evolves due to the addition of new subdomains and the growth of scientific knowledge. Thus, we maintain a production version to which users submit data, and a development version to which developers add ontological features. The ontology of the production version does not change continuously, but in steps or releases, while the ontology of the development version is more fluid. At each new release of the system, the development version and the production version must be merged. We are developing methods and tools for performing what can be a complicated merge process.

One of the goals of the PharmGKB project is to provide to the community data and knowledge that do not currently exist in other Web resources, and to facilitate interoperation with other databases in beneficial ways. For example, dbSNP is a public database that permits scientists to enter newly discovered data about single nucleotide polymorphisms (SNPs) in genes. PharmGKB users can submit SNP frequency data to PharmGKB, and if the SNP is not known to dbSNP, an application associated with PharmGKB can automatically make a SNP submission to dbSNP. PharmGKB also coordinates with Medline by storing citations for relevant published articles, and with GenBank, LocusLink, and Online Mendelian Inheritance in Man (OMIM) by storing relevant accession numbers. These links will make it possible to offer expanded services that require interaction with these other databases in the future.

Any knowledge-based system that serves as an evolving resource on the Web to which users from a particular community contribute, that interoperates with other Web resources, and whose underlying data and knowledge model is constantly changing, requires careful management of data, knowledge, and metadata. In our work on PharmGKB, we are confronting these challenges, and our experiences will be applicable to other similar resources that could be available and accessible on the Semantic Web.

Advertising and Matching DAML-S Service Descriptions

Terry R. Payne, Massimo Paolucci, Katia Sycara

Robotics Institute, CMU
5000 Forbes Avenue
Pittsburgh, PA 15213 USA
+1 (412) 268 8780
[*terryp*|*paolucci*|*katia*]*@cs.cmu.edu*

ABSTRACT

Software agents and Web-based services extend beyond traditional Web activities (i.e. locating and browsing information) by allowing one to effect some action or change in the world, such as the sale of a product or the control of a physical device. Agents should be able to locate, select, employ, compose, and monitor Web-based services automatically. To do this, a computer-interpretable description of the service is required, as well as a means by which it can be located. This paper presents work in progress on the ATLAS matchmaker; a location service that facilitates the discovery of agents and services described using DAML-S. We outline how services are advertised, and describe how the ATLAS matchmaker identifies matching profiles for a given service request.

INTRODUCTION

As the number and variety of services on the Semantic Web increases, frameworks that allow services and agents to interoperate, seek and cooperate with each other will be required. Infrastructures developed within the MAS community provide such a framework, as they facilitate automatic discovery of agents and web-services on the basis of their capabilities. In this paper we leverage previous experience in developing yellow-pages registry services (e.g. LARKS [6]) to develop a mechanism for automated service discovery within the Semantic Web. As part of the DARPA Agent Markup Language program, an ontology of services[1], called DAML-S [1], has been developed, which provides a set of basic concepts and relations for declaring and describing services, by utilizing the ontology structuring mechanisms provided by DAML [3].

A DAML-S service is characterized by three types of knowledge; a *service-profile*, *process-model*, and *service-grounding*. The *service-profile* describes *what the service does*; it provides the type of information needed by a service requester to determine whether the service has the desired capabilities. The *process-model* describe *how the service works*; i.e. how it is composed, and what happens when the service is executed. A *service-grounding* specifies the details of *how an agent can access a service*. Generally speaking, the *profile* provides the information needed for an agent to discover a service. Taken together, the *process-model* and *grounding* describe how agents and services access, and interoperate with each other.

Service profiles consist of three types of information: a human readable *description* of the service; a specification of the *functionalities* that are provided by the service; and a list of *functional attributes* which provide additional information and requirements about the service that assist when reasoning about several services with similar capabilities. *Service functionalities* are represented as a transformation from the inputs required by the agent to the outputs produced. For example, a news reporting service would advertise itself as a service that, given a date, will return the news reported on that date. *Functional attributes* specify additional information about the service, such as what guarantees of response time or accuracy it provides, or the cost of the service. Table 1 lists the properties defined by the service profile (note that the properties *input* and *output* are defined as sub-properties of *parameter*).

A service provider can register, or advertise its profile with one or more *Middle Agents* [4]. There are several different ways in which middle agents interact with service providers and service requesters [5], depending on factors such as reliability, privacy, efficiency etc. A sub-class of middle agents are generally known as *matchmakers*, *yellow pages* or *directory agent* systems [2][4][5]. These only have knowledge about the capabilities of service providers. Thus, if an agent has some preferences, it can query the middle agent, which then returns a list of agents whose capabilities match the preference query. In contrast, service requesters have *preferences* for particular parameters associated with desired capabilities. Providers advertise their capabilities and service parameters with one or more middle agents, and requesters request agents with particular capabilities and select a provider according to their preferences.

SERVICE MATCHING

A variety of different approaches have been used to match agent advertisements and service requests. The LARKS matchmaker [6] consists of a number of filters, each of which performs partial matching on subsets of the descriptors. Several of the filters perform semantic matches, by determining a semantic distance between co-existent terms within shared ontologies. The ATLAS matchmaker utilizes two separate filters: one compares *Functional Attributes* to determine the applicability of advertisements (i.e. do they

---

[1] More details about the DAML-S Semantic Markup can be found at http://www.daml.org/services/

support the required type of service or deliver sufficient quality of service); the second compares *Service Functionalities* to determine if the advertised service matches the requested service. A DAML-based subsumption inferencing engine is used to determine if one concept subsumes another.

Description	**Service Functionalities**	**Functional Attributes**
serviceName	*parameter*	*geographicRadius*
intendedPurpose	*input*	*degreeOfQuality*
textDescription	*output*	*serviceType*
role	*conditionaloutput*	*serviceCategory*
requestedBy	*precondition*	*serviceParameter*
providedBy	*accesscondition*	*communicationThru*
	effect	*qualityRating*
	domainResource	*qualityGuarantees*

Table 1: Properties of the Service Profile

Matching Functional Attributes

Matching is achieved by performing conjunctive pair-wise comparisons for the properties listed in Column 3 of Table 1. As each of the functional attributes may refer to different ontologies, different types of inference are used to test each pair. For example, *geographicRadius* comparisons may use a different mechanism to determine if the request falls within the geographic scope of the advertisement, to that used to compare *serviceTypes*. As property pairs are assumed to match comparisons cannot be performed (e.g. due to missing property values), the inclusion of property values within the request will provide a way of constraining the set of candidate advertisements that match the request.

Matching Service Functionalities

Different service providers may describe similar services in different ways, by labeling properties as inputs instead of preconditions, effects instead of outputs, etc. For this reason, service functionalities are grouped into two sets; an input set (consisting of the union of *input*, *precondition*, and *accesscondition* properties) and an output set (consisting of the union of *output*, *conditionaloutput*, and *effect* properties). The input and output sets are determined for the request (R_I and R_O) and for each advertisement (A_I and A_O). The input set R_I is compared with an advertisement input set, A_I, and a match is determined if $A_I \subseteq R_I$, i.e.

$$match(R_I, A_I) \Leftarrow (\forall j, \exists i: (i \in R_I) \wedge (j \in A_I) \wedge subs(i, j)) \wedge R_I = \emptyset$$

where $subs(i, j)$ is true when i subsumes j.

Thus, for each concept in the advertisement input set, there is some concept in the request input set that subsumes it. Thus, there is a match when a request can provide all necessary inputs required by an advertisement. However, an agent may not have a model of what inputs may be required, and may need to obtain this information from the advertisement. To support this, input sets also match when the request set is empty.

Output sets match when all the elements in R_O are subsumed by elements in A_O, i.e.

$$match(R_O, A_O) \Leftarrow \forall i, \exists j: (i \in R_O) \wedge (j \in A_O) \wedge subs(j, i)$$

To avoid the possibility of one element matching several others, a greedy approach will be used, whereby each matching pair of elements are removed from the input and output sets. A mapping of corresponding inputs and outputs will also be generated, and returned to the requesting agent if a match is determined.

CONCLUSION

The ATLAS matchmaker provides registration and lookup capabilities of web-based services described using the DAML-S service profiles. Service requests are compared with service advertisements through a subsumption based inferencing mechanism using DAML ontologies. Two separate filters are being developed: one for comparing functional attributes that characterize the service along descriptive dimensions (such as business types, rating mechanisms etc); the second compares the functional capabilities of the service in terms of inputs and outputs.

Several evaluation studies are planned for this system, to determine the effectiveness and speed of each of the filters, to evaluate the utility of Boolean subsumption inferencing for matching DAML concepts; and to evaluate the design of the DAML-S profile descriptions.

ACKNOWLEDGMENTS

This research was supported by a DARPA grant F30601-00-2-0592. Special thanks go to the DAML Services Coalition for their assistance in the development of DAML-S.

REFERENCES

1. Ankolekar, A., Burstein, M., Hobbs, J.R., Lassila, O., Martin, D.L., McIlraith, S.A., Narayanan, S., Paolucci, M., Payne, T.R., Sycara, K., and Zeng, H. DAML-S: Semantic Markup for Web Services. Submitted to International Semantic Web Workshop (SWWS), 2001
2. Genesereth, M., and Katchpel, S. Software Agents. *Communications of the ACM*, 37(7):48-53, 1994.
3. Hendler, J. and McGuinness, D. DARPA Agent Markup Language. *IEEE Intelligent Systems*, 15(6):72-73, 2001
4. Sycara K., Decker, K., Pannu A., Williamson M., and Zeng D. Distributed Intelligent Agents. *IEEE Expert: Intelligent Systems and their Applications*, Vol. 11, No. 6 (December 1996).
5. Sycara K., Decker, K., and Williamson M. Middle-agents for the Internet. In *Proceedings of IJCAI-97*, January 1997.
6. Sycara, K., Klusch, M., Widoff, S. and Lu, J. Dynamic Service Matchmaking Among Agents in Open Information Environments. *SIGMOD Record (ACM SIG on Management of Data)* Vol. 28 No. 1, pp47-53, 1999.

Digital Library Portal using Semantic Tools in WWWPal

John R. Punin and Mukkai S. Krishnamoorthy
Department of Computer Science,
Rensselaer Polytechnic Institute, Troy, NY 12180, USA
{puninj,moorthy}@cs.rpi.edu

Abstract

The WWWPal system and associated languages and tools (such as LOGML, XGMML, webbot and the graph browser) have been developed to perform syntactic analysis of web sites. In this paper, using WWWPal and semantic analysis tools (such as RGML, clustering and the graph browser) we construct digital library portals. We describe a method of obtaining the portal.

Introduction

Semantic web was introduced [BERNERS] to make the tangled information in the web more accessible to search engines and other applications. The semantic web is not a separate web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation [BERNERS01]. Specifying semantic information to the web content will make this task easier. On the other hand, it may be harder for the content developer to provide the semantic information that the user agent may want. Modifying the existing web pages with semantic information may result in additional errors. W3 Consortium arrived at a solution of specifying the meaning of a web resource using Resource Description Framework (RDF)[RDF]. RDF encodes the metadata in sets of triples, each triple being rather like the subject, verb and an object of an elementary sentence. A number of papers [SEMWEB01] have been published about RDF, vocabularies based on RDF and RDF applications.

In this paper we develop a general purpose framework, a collection of semantic tools using RDF, which can be applied for digital library web portals. Designing a portal for a digital library, a collection of information that can be browsed and searched by search engines and humans, is a simpler task with this framework.

WWWPAL Support for Digital Library Portals

WWWPal provides support for the Digital Library Research by collecting, filtering, and classifying the available metadata of a web site. In a digital library, there are two standard models of delivering information. This information is either static (i.e., supplied by a librarian with a keyword classification) or dynamic (i.e., the system tries to obtain a keyword using heuristics). WWWPal provides both the static support (by providing an RDF editor) and the dynamic support (by providing clustering and keyword classification). The web robot of WWWPal navigates a web site, saves the structure of the web site and collects the metadata information of the web site. The structure of the web site is saved in an XGMML (XML vocabulary for graphs) document [LOGML]. The metadata of the web pages and hyper links is appended to the nodes (web pages) and edges (hyperlinks) using the RDF/XML serialization. The XGMML document is transformed into an RGML (RDF vocabulary to describe graphs) document [RGML]. This RDF document is read by an RDF parser to produce a set of triples. Further, we can group several web documents by finding the clusters of the webgraph. These clusters are represented as subgraphs of the webgraph, and the metadata of each node of the subgraph is merged to form the metadata of the cluster. All of these subgraphs are saved in an RGML file. We have also developed a simple web portal so users can browse and search the information gathered in this repository.

Metadata Collection

The metadata collection is achieved using the web robot of WWWPal [WWWPAL]. The web robot navigates a web site using a breadth first search algorithm. Each visited web page is parsed to find the following metadata information:

- Title of the web page in the <TITLE> tag

- Metadata information in the <META> tag
- Metadata linked to the web page using the <LINK> tag
- Anchor text information in the <A> tag
- Headers of the page in the <H1> to <H6> tags

This metadata information may not be sufficient to describe with a web page. Our experiments suggested that we get most of the metadata information from the above five cases. All the metadata information of the web page is attached to the node of the graph that represents the web page. We use RDF to represent the metadata information and XGMML to save the whole webgraph structure. The following example is an XGMML document using RDF vocabulary to represent metadata. The collected metadata is title, date, format and keywords. We have used the Dublin Core vocabulary [DC] to represent these RDF properties of the web page. Figure 1 shows the structure of this webgraph using the WWWPal Graph Browser.

```
<?xml version="1.0"?>
<graph xmlns = "http://www.cs.rpi.edu/XGMML"
       directed="1" >
<node id="3" label="http://www.cs.rpi.edu/courses/" weight="6968">
<att>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.0/">
  <rdf:Description about="http://www.cs.rpi.edu/courses/"
    dc:title="Courses at Rensselaer Computer Science Department"
    dc:subject="www@cs.rpi.edu; M.S. requirements; CSCI-1190 Beginning C Programming for Engineers; Courses; People;
    Graduate Program; CSCI-4020 Computer  Algorithms; CSCI-2220-01  Programming in Java; Research; Course Selection Guide;
    CSCI-4961-01,  CSCI-6961-01 Advanced Robotics; Programming in Java; CSCI-2400 Models  of Computation"
    dc:date="2000-01-31"
    dc:type="Text"
    >
    <dc:format>
      <rdf:Bag
        rdf:_1="text/html"
        rdf:_2="6968 bytes"
      />
    </dc:format>
  </rdf:Description>
</rdf:RDF>
</att>
</node>
<node id="7" label="http://www.cs.rpi.edu/research/" weight="13732">
<att>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.0/">
  <rdf:Description about="http://www.cs.rpi.edu/research/"
    dc:title="Research at Rensselaer Computer Science
    Department"
    dc:subject="www@cs.rpi.edu; Computing  Twin Primes and Events; TEMPEST; Courses; People; Graduate Program;
    High-Performance  Object-Oriented Programming in Fortran 90;
    High Performance Problem-Solving Environment for Optimization and Control of  Chemical and Biological Processes;
    Computer Vision; Theory and Algorithms; technical report library; RPInfo; Undergraduate  Program;
    Research; Research; Design Conference Room; Rensselaer; Bryan Rudge; Engineering Databases; anonymous ftp; I.SEE;
    info@cs.rpi.edu; Scientific Computing; OpenMath; Proactive Network Problem Avoidance; Computing Facilities; Computer Science Departm
    dc:date="1999-11-19"
    dc:type="Text"
    >
    <dc:format>
      <rdf:Bag
        rdf:_1="text/html"
        rdf:_2="13732 bytes"
      />
    </dc:format>
  </rdf:Description>
</rdf:RDF>
</att>
</node>
<node id="8" label="http://www.cs.rpi.edu/undergrad/" weight="7672">
<att>
<rdf:RDF
```

```
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.0/">
  <rdf:Description about="http://www.cs.rpi.edu/undergrad/"
    dc:title="Undergraduate Program at Rensselaer Computer Science Department"
    dc:subject="www@cs.rpi.edu; Association for Computing Machinery; People; Graduate Program; Minor; Prospective Students FAQ;
    Admissions Office; Research; BS-MS Degree; Rensselaer Catalog; Dual Majors; Admissions;Computing Facilities; Undergraduate Program;
    Rensselaer; Course Descriptions; Computer Science Department"
    dc:date="2000-01-26"
    dc:type="Text"
    >
    <dc:format>
      <rdf:Bag
        rdf:_1="text/html"
        rdf:_2="7672 bytes"
      />
    </dc:format>
  </rdf:Description>
</rdf:RDF>
</att>
</node>
<node id="1" label="http://www.cs.rpi.edu/" weight="3402">
<att>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.0/">
  <rdf:Description about="http://www.cs.rpi.edu/"
    dc:title="Rensselaer Computer Science Department"
    dc:subject="www@cs.rpi.edu; faculty positions; Rensselaer; Graduate Program; info@cs.rpi.edu; Current Events;
    RPInfo; Research; Undergraduate  Program"
    dc:date="2000-01-26"
    dc:type="Text"
    >
    <dc:format>
      <rdf:Bag
        rdf:_1="text/html"
        rdf:_2="3402 bytes"
      />
    </dc:format>
  </rdf:Description>
</rdf:RDF>
</att>
</node>
<edge source="1" target="3" weight="0" label="SRC IMG gfx/courses2.jpg" />
<edge source="7" target="3" weight="0" label="SRC IMG ../gfx/courses2.jpg" />
<edge source="8" target="3" weight="0" label="SRC IMG ../gfx/courses2.jpg" />
<edge source="3" target="7" weight="0" label="SRC IMG ../../gfx/research2.jpg" /
<edge source="1" target="7" weight="0" label="SRC IMG gfx/research2.jpg" />
<edge source="8" target="7" weight="0" label="SRC IMG ../gfx/research2.jpg" />
<edge source="3" target="8" weight="0" label="SRC IMG ../../gfx/ugrad2.jpg" />
<edge source="7" target="8" weight="0" label="SRC IMG ../gfx/ugrad2.jpg" />
<edge source="1" target="8" weight="0" label="SRC IMG gfx/ugrad2.jpg" />
<edge source="3" target="1" weight="0" label="SRC IMG ../../gfx/corner2.jpg" />
<edge source="7" target="1" weight="0" label="SRC IMG ../gfx/corner2.jpg" />
<edge source="8" target="1" weight="0" label="SRC IMG ../gfx/corner2.jpg" />

</graph>
```

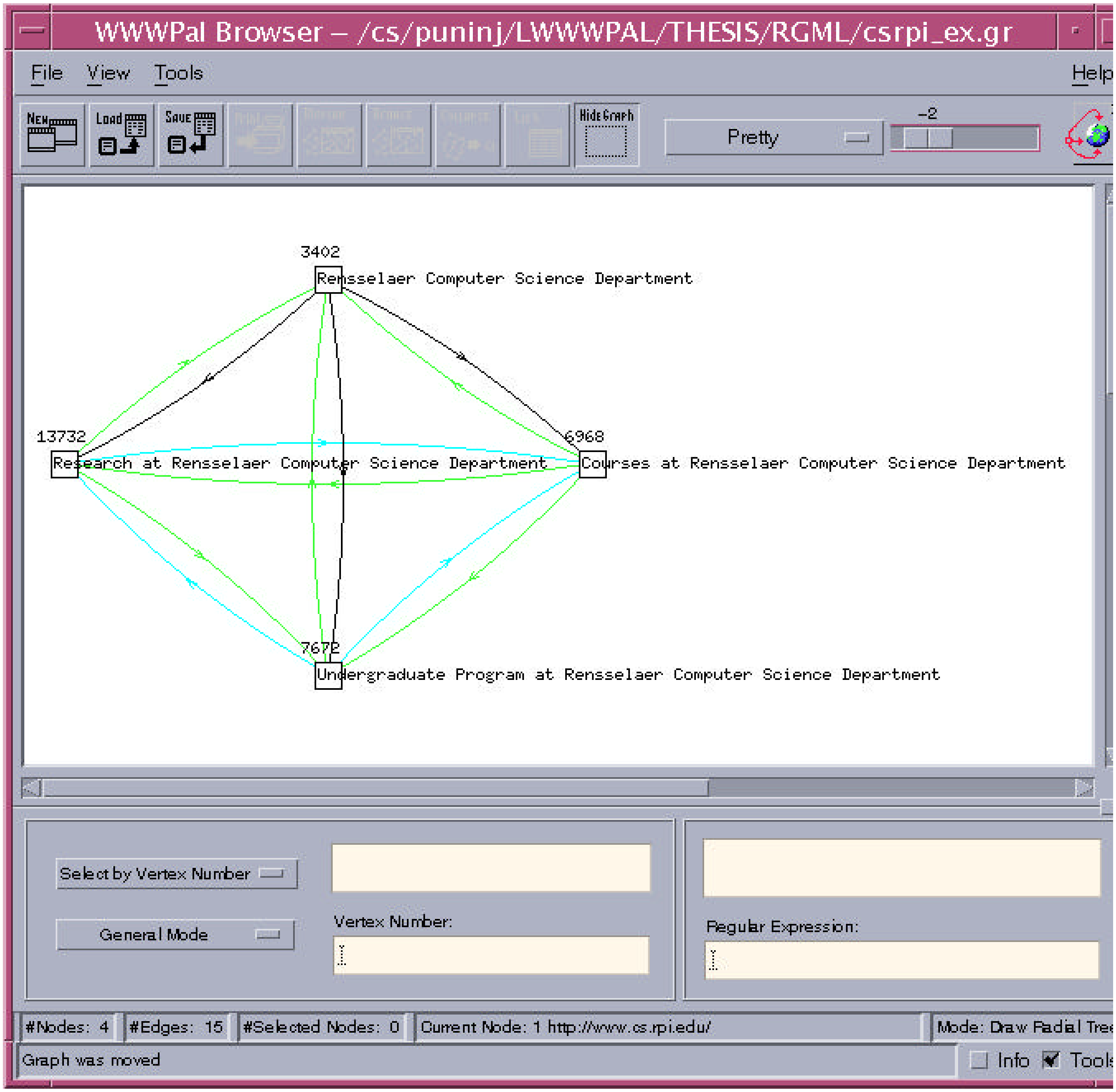


Figure 1: Webgraph of the main web pages of the RPI Department of Computer Science.

Metadata Filtering

RDF has been conceived to represent semantic information. We use RGML to save the metadata of a given web site . One of the WWWPal modules transforms the XGMML document into an RGML document. The RGML vocabulary follows the RDF syntax and makes it simple to combine different RDF vocabularies such as Dublin Core [DC] and Vcard [VCARD]. The generated RGML file can be read by any RDF parser. The parser generates a set of triples which are the RDF statements: subject, predicate and object. Before parsing the RGML file, it is important to cluster the webgraph. The generated clusters provide metadata information such as keywords. Clusters are subgraphs of the webgraph and hence RGML can represent them by using the *graphs* property. WWWPal has implemented several clustering methods and they are fully explained in [WWWPAL].

Digital Library Portals

Popular web portals such as Yahoo and Netscape classify and present important information such as news, weather and entertainment in just one web page so the user does not have to spend too much time in finding valuable information. An educational web portal has been described in [JASIG]. We know that most of the educational web sites do not offer portals and hence it is difficult to find information. WWWPal provides a web interface to browse and search a repository of RGML documents. Each RGML document contains the information of an educational web site such as Computer Science Department sites. The user can visit this web interface and quickly find the specific information that they are looking for. We consider this web interface as a simple example of a Digital Library Portal where the RGML documents provide semantic information. Future development will add an inference engine and several rules to construct a powerful knowledge base so that the Digital Library supports semantic search and presentation of knowledge.

References

[BERNERS] T. Berners-Lee, *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*, San Francisco, CA, HarperCollins 1999

[BERNERS01] T. Berners-Lee, J. Hendler, and O. Lassila, *The Semantic Web*, Scientific American, May 2001.

[DC] S. Weibel, J.Kunze, C. Lagoze, and M. Wolf. *Dublin Core Metadata for Resource Discovery*, Internet RFC 2413. 1998.

[JASIG] JA-SIG Portal Framework Project, www.mis2.udel.edu/ ja-sig/portal.html; Bernard W. Gleason, Boston College University-Wide Information Portal: Concepts and Recommend ed Course of Action, Jan. 26, 2000.

[LOGML] J. Punin, M. Krishnamoorthy, M. J. Zaki, *Web Usage Mining: Languages and Algorithms*, to appear in Studies in Classification, Data Analysis, and Knowledge Organization, Springer-Verlag, 2001.

[RGML] J. Punin and M. Krishnamoorthy, *Describing Structure and Semantics of Graphs Using a RDF Vocabulary*, to appear at Extreme Markup Languages 2001.

[RDF] O. Lassila, and R. Swick, *RDF Model and Syntax Specification W3C Recommendation* , February 1999.

[SEMWEB01] *The Second International Workshop on the Semantic Web*, April 2001.

[VCARD] R. Iannella, *Representing vCard Objects in RDF/XML,* W3C Note February 2001.

[WWWPAL] J. Punin, M. Krishnamoorthy. *WWWPal System - A System for Analysis and Synthesis of Web Pages*. In Proceedings of the WebNet 98 Conference, Orlando, November, 1998.

Linguistic Techniques and Standards for Semantic Mark-up

James Pustejovsky
LingoMotors, 585 Mass. Ave.
Cambridge, MA 02139
jamesp@lingomotors.com
and

Computer Science Department
Brandeis University, Waltham, MA 02254
jamesp@cs.brandeis.edu

1 Semantic Technologies for Mark-up Standards

In this position paper, we present several points regarding current efforts towards the establishment of standards for web content, particularly from the contribution of natural language technologies. The position argued here is that the modeling of natural language meaning is critically important for the robust establishment of content markup languages for web protocols and communication. It is the view here that the largest, and indeed, the most successful, semantic webs are existing linguistic communities. Although this might appear trivial or irrelevant to the concerns of web content mark-up and communication, we shall argue that the efforts recently underway in the natural language technologies community, such as with LingoMotors and other companies, are relevant to the establishment and dissemination of semantic web standards.

The research and development effort at LingoMotors is focused on the automatic identification of semantic content, in the form of digital assets such as web text, for subsequent use by a consuming application. Such applications include information and database retrieval systems, CRM, content and knowledge management systems, as well as categorization and clustering algorithms. From the general perspective of the present workshop, such technologies can be seen as enabling several opportunities relating to the automatic semantic markup, interpretation, and verification of web content and interactions. More specifically, LingoMotors technology is helping to realize a user interface for the web that is mediated through ordinary language interactions.

2 Encoding and Recognizing Content

The semantic technologies being developed at LingoMotors contribute directly to the realization of a robust *Language User Interface* (LUI) for the web, helping to drive the interpretation of business exchanges, commerce, and other communicative transactions. A Language User Interface provides a platform from which users interact with the computer in ordinary language, for most everyday applications. The underlying technology responsible for this is what we call the *Concept Machine* (c-machine). Such an interface effectively hides the complexity of the knowledge and actions behind the words and text.

Web interactions outfitted with semantic technologies can offer distinct capabilities that are impossible to achieve with current search and navigation technologies, for identifying the rich and pertinent structural information about objects in an application area. For example, a "knowledge object" corresponding to the notion of an ecommerce product has `performance`, `form factor`, `price`, and `customer attributes`, each of which can take specific values. The concept machine not only contains these structural descriptions but also the knowledge for recognizing and attaching attributes and their values to the objects. Some of the major areas enabled by semantic technologies include the following:

1. Relation Searching and filtering: Many searches or filtering conditions are interested not in things, but in events and actions; that is, a category or type of company buying or acquiring another company, the announcement of a new alliance or product introduction or price drop, etc. This is possible only with robust recognition of the semantic types of the relations in the texts.

2. Category-based searching and filtering: A significant shortcoming of today's searching and filtering techniques is that the user has to know the name of what he or she is looking for in order to find or exclude it. Semantic technologies allow the user specify search or filtering conditions based upon categories, such as product type, company type, personal title, without having to know the literal names.

3. Categorization and Clustering: Semantic technologies provide recognition of entities and relations between entities. As a result, they enable richer, more meaninful categorization possibilities. Similarly,

clustering algorithms are driven by concepts rather than word tokens, resulting in more informed classifications.

4. Similarity matching: Current similarity searching methods primarily rely upon statistical comparisons of literal strings. Semantic mark-up enables similarity searches and matches to combine the rich description inherent in the concept machine with statistical information metrics to provide more meaningful matches.

We believe that the success and deployment of semantic technologies is essential for the realization of the goals and vision of the semantic web community. The LingoMotors technology is focused directly at addressing these challenges.

A Prototype DAML+OIL Ontology IDE

Abir Qasem
Visiting Assistant Professor
Department of Computing and Mathematics
Embry Riddle Aeronautical University
aqtg@netzero.net

Introduction

The emergence of the Semantic Web opens up boundless possibilities by enabling agents to reason about its content and provide rational responses to unanticipated situations. However [Lassila & McGunniess, accessed 2001] observes that acceptance by the mass (people outside the KR community) is critical to the success of the Semantic Web. Looking back at the Web revolution of the nineties, we see that there were at least two key enablers of its mass acceptance. One was the standardization and global acceptance of HTML and the related protocols and the other was the free availability of universal and easy to use Web processing tools like Netscape. Similarly, in the case of the Semantic Web, considerable development and standardization effort is resulting in the evolution of DAML+OIL as the de-facto ontology language. However, at present, most of the native DAML+OIL processing tools are built by KR specialists for KR specialists. As a result they do not attempt to hide the details of an ontology building task from the user. The situation can be compared to the "edit code - compile it - link it - load it and then run it" days of programming. It is conceivable that an integrated development environment (IDE) to handle DAML+OIL ontology building details (such as creating, locating, reusing, merging, validating) will add to the appeal of the Semantic Web. In this short paper, we describe a DAML+OIL ontology development project at Embry Riddle Aeronautical University (ERAU) where a prototype DAML+OIL IDE to support knowledge workers is presently being developed. We describe the project, identify its specific requirements - as related to ontology building, consider adaptation of a few available tools, and then present a high level description of the IDE under development.

The WOSE project

The Web-enabled Ontology of Software Engineering (WOSE) project at Embry Riddle Aeronautical University was conceived in an attempt to address the need for alternative approaches to collection, categorization and dissemination of software engineering body of knowledge (SwE BOK). The Joint IEEE and ACM Software Engineering Coordinating Committee has identified that achieving consensus by the profession on a core body of knowledge is crucial for the evolution of software development practices into a professional engineering discipline. There are many ongoing efforts to achieve this consensus [Hilburn, accessed 2000] but the successes have been very limited.

We propose an ontology driven approach to the SwE BOK development. It is interesting to note that although it seems that the use of ontologies should be an intuitive choice for representing an evolving body of knowledge, we did not find many applications of ontologies in this area. The knowledge acquisition community has used KA^2, a collaborative environment in developing a "research topic ontology" for the community [AIFB, accessed 2000].

The details of the WOSE project may be found at http://java-emporium.com/WOSE. One of the major tasks of the project is to enable geographically dispersed SwE BOK authorities to collaborate on building an initial DAML+OIL ontology of SwE BOK. This involves development of the DAML+OIL IDE. Details of the project tasks may be found at http://java-emporium.com/WOSE/Tasks.html.

WOSE specific requirements

The knowledge workers in the WOSE project are faculty members and practitioners with experience in software engineering, but not necessarily with much interest in KR intricacies. Given this background, we have identified the following requirements for the IDE:

- **Reuse**: We would want the tool to explicitly encourage the user to start from one or more existing ontologies. We will provide search and visualization support to facilitate this.
- **Diagnostic**: We would want the tool to identify and report the logical inconsistencies of ontology merges. It should also provide validation support of a new ontology.

- **Remote ontology access**: We would want the tool to access URIs of DAML+OIL files
- **Visual Browsing:** We would want the tool to have a drag and drop merging facility and graph visualization support for the DAML+OIL ontologies.

Survey of the Existing tools in light of the requirements

Since there is no native DAML+OIL IDE [DAML , accessed 2001] we have looked at some other ontology building tools that may be adapted to suit our purpose. All of the tools that we looked at support DAML+OIL through some export/import or wrapper function. The following table shows the tools in terms of our requirements.

Requirements	*Ontoedit*	*OilEdit*	*Chimaera*	*Protégé*
Reuse	Some	Some	Some	Some
Diagnostics	Some	Some	Excellent	Some
Remote ontology access	None	None	Manual	None
Visual browsing	Some	Some	Some	Some

Since we did not have a clear winner, we have decided to develop a DAML+OIL IDE for the SwE BOK knowledge workers of the WOSE project.

Description of the tool

Functionally the tool integrates four existing DAML+OIL/ontology building resources namely the catalog of the DAML+OIL ontology library, the publicly accessible DAML+OIL ontologies on the Internet, ***Chimaera's*** diagnostic services and the DAML+OIL validator. The XML format of the DAML+OIL ontology library is used to present the user with a tree type rendering of information. The interface is a typical JTree rendering, with the left pane showing the tree and the right pane showing the content of each selected node. The selected node (a DAML+OIL ontology URI) will be displayed as a graph using techniques borrowed from RDF visualization. The user will be able to initiate a merge by dragging and dropping tree nodes. The ontologies will then be submitted to ***Chimaera*** (implemented using the OKBC API) and the results returned and automatically submitted for addition to the DAML+OIL library. The tool will also have an interface with the recently developed DAML+OIL validator. Please see http://students.db.erau.edu/~qasema/May15Arch.jpg for a functional diagram of the tool.

Since we are implementing an emerging technology we need a highly adaptable architecture. The software has frequent use of adapter objects to support this open architecture. The development is hosted on the sourceforge.net

Conclusion and future work

We expect the alpha version of the tool to be completed by mid-July. At that point we would want to validate the tool by:

- Giving the same ontology building task to two test subjects; one uses the IDE and the other uses a combination of the other tools
- Identifying a reuse metric (one that shows that Ontology X has better reuse than Ontology Y)
- Using this reuse metric to measure the improved reuse (if any)

Reference

[AIFB accessed 2000] http://www.aifb.uni-karlsruhe.de/WBS/broker/KA2.html Web site for Institut für Angewandte Informatik und Formale Beschreibungsverfahren accessed 2000

[DAML accessed 2001]
http://www.daml.org/tools/wishlist.html accessed 2001

[Hilburn, accessed 2000]
http://faculty.db.erau.edu/hilburn/se-educ/, Software Engineering Education accessed 2000

[Lassila & McGunniess accessed 2001]
The Role of Frame-Based Representation on the Semantic Web,
http://www.ida.liu.se/ext/epa/cis/2001/005/tcover.html, Linköping Electronic Articles in Computer and Information Science accessed 2001

Towards a User-Focused Approach to the Semantic Web

A Position Paper for the International Semantic Web Workshop

Scott Tsao and Christina Portillo
Boeing Commercial Airplanes
July 30-31, 2001

Introduction

The announcement of the World Wide Web Consortium (W3C) Semantic Web (SW) Activity has drawn tremendous interest around the world since the beginning of this year. Most of the discussions on this topic tend to be "technology focused," i.e., interest or working groups formed among technicians familiar with a certain set of standards or technologies with the intent to explore and justify their role in the (vaguely-defined) "Semantic Web." Ironically, coming from the user community of the (yet-to-be-developed) SW technologies, we often find it difficult to be engaged in any of those groups because we lack prerequisite knowledge of the local "dialect." For the SW initiative to be successful, The Boeing Company believes that a "user-focused" approach is required. As such, we would like to share our view of the SW primarily from a user's perspective.

Since the fall of the "Tower of Babel" we as human beings have been obsessed with the desire to communicate with each other. Now that the World Wide Web (WWW) is a reality in our daily lives, the same obsession motivates us to find better ways to communicate via machines that are Web-aware. For the purpose of this paper we categorize our potential use of the Semantic Web in the following four (not necessarily mutually exclusive) areas:

- Data Interchange
- Application Integration
- Information Sharing
- Knowledge Discovery

Data Interchange

XML-based data interchange normally requires the use of an XML-based information model definition mechanism (i.e., DTD or Schemas). However, as the number of DTD's and Schemas increases, along with the variety of underlying information model definition languages (e.g., W3C XML DTD, W3C XML Schema, and RELAX), we see an emerging need for standards and technologies enabling "semantic convergence" among those information model definition mechanisms. One approach is to segregate the DTD's and Schemas into domain-specific repositories that can be looked up through federated registries, and to define common structures, services, and behaviors. For example, the CommerceNet eCo Framework Project defines a seven-layer architecture of collaborating registries from "Networks" all the way down to "Information Items." We believe that SW-enabled registry services facilitating automated index and intelligent query through controlled vocabularies are needed to bring the WWW from "semantic dispersion" into "semantic convergence." And we expect that coordination between the W3C SW

Activity and XML Protocol Activity will leverage and harmonize similar initiatives such as UDDI and ebXML.

Application Integration

XML-based application integration normally employs a mapping mechanism between an application's native meta-model definition languages (e.g., UML, EXPRESS, SQL) and an XML-based meta-model definition language (i.e., DTD or Schemas). One approach is to specify the mapping rules between two different languages (e.g., UML and DTD), as exemplified by the OMG XMI specification for interoperability between CASE tools. The ISO 10303 STEP Part 28 Project (XML Representation of EXPRESS Schemas and Data) is taking a similar approach to enable the interoperability between CAD/CAM/CAE tools for product design and engineering and other applications used within the product life cycle. For example, the PLCS Initiative "seeks to provide global agreement on the definition and communication of the information needed by users to plan and execute support for complex, long life assets such as aircraft, ships, and large industrial plants." We believe that SW-enabled mapping services facilitating the discovery, merger, and transformation of meta-models underlying various applications used in the product life cycle are needed to accomplish the PLCS vision.

Information Sharing

While Data Interchange and Application Integration are required primarily for structured information, a significant portion of valuable information assets is natively stored in semi-structured or unstructured format. Any addressable piece of data on the Web could be a piece of the puzzle needed to complete the whole picture of the sharable information asset. Ever since publication has been available in the civilized world, human beings have been employing various ontological representations (e.g., taxonomy, thesaurus, index, and context rules) to facilitate the process of "piecing together the puzzle." The basic approach we are taking even in the Internet age today is not much different. For example, the portal interface to the Web allows people within a given community to organize information in terms of customized portlets that can be navigated using a pre-defined taxonomy, and more advanced search and navigation can be accomplished through a built-in thesaurus or index. When information sharing is required across multiple communities, semantic mapping is required between multiple ontological representations. We believe that SW-related standards and technologies that enable the construction, merger, and exchange of such ontological representations, independent of the Web information resources, are needed to facilitate advanced search and navigation across community boundaries. In addition to XML-based linking technologies such as XLink (along with linking features available in DTD, RDF and XML Schema), we expect the W3C SW activity will leverage accomplished works such as ISO 13250 Topic Maps to accelerate the availability of standards and technologies for Information Sharing.

Knowledge Discovery

The ultimate goal of the practice of knowledge management is to make tacit knowledge of human beings explicit and reusable. One approach is to perform text mining and data mining of any “raw” (in whatever shape and form) information available using various (often sophisticated) algorithms, and to perform “sense making” using various visualization, clustering, and affinity analysis techniques. This is an interesting research area that has not seen significant employment of XML-based standards and technologies. We encourage active cross-pollination between those related communities to explore, identify, and develop SW technologies to facilitate Knowledge Discovery.

Glossary of Acronyms

CAD	**C**omputer-**A**ided **D**esign
CAE	**C**omputer-**A**ided **E**ngineering
CAM	**C**omputer-**A**ided **M**anufacturing
CASE	**C**omputer-**A**ided **S**oftware **E**ngineering
DTD	**D**ocument **T**ype **D**efinition
ebXML	**e**lectronic **b**usiness **XML**
ISO	**I**nternational **S**tandards **O**rganization
OMG	**O**bject **M**anagement **G**roup
PLCS	**P**roduct **L**ife **C**ycle **S**upport
RDF	**R**esource **D**escription **F**ramework
RELAX	**RE**gular **LA**nguage description for **X**ML
SQL	**S**tructured **Q**uery **L**anguage
STEP	**ST**andard for the **E**xchange of **P**roduct Model Data
SW	**S**emantic **W**eb
UDDI	**U**niversal **D**escription, **D**iscovery, **I**ntegration
UML	**U**nified **M**odeling **L**anguage
W3C	**W**orld **W**ide **W**eb **C**onsortium
WWW	**W**orld **W**ide **W**eb
XLink	**X**ML **Link**ing Language
XMI	**X**ML **M**etadata **I**nterchange
XML	e**X**tensible **M**arkup **L**anguage

International Semantic Web Working Symposium (Stanford, July 30th - Aug. 1st, 2001)

Semantically Marked Up… Now What?

Simon Buckingham Shum, John Domingue and Enrico Motta

Knowledge Media Institute, The Open University
Milton Keynes, MK7 6AA, UK
{S.Buckingham.Shum, J.B.Domingue, E.Motta}@open.ac.uk

Broadening the vision

"Wired…now what?", was the challenge from Apple's Mark Miller to educationalists as the Internet took off. A new infrastructure was being laid, but even now we still barely know how to use it effectively for learning. We can expect a similar scenario to unfold for the semantic web. Currently, the semantic web community is necessarily preoccupied with infrastructure design (e.g., knowledge representation languages for the web, scaling issues, standardising terminologies and agent interoperability). But once this infrastructure is in place, more persistent questions will remain concerning the technology's relationship to the people expected to use it. Specifically, we ask: How can the semantic web be used to support *people* in *knowledge-intensive work*?

"Semantically marked up…now what?", is therefore a worthwhile question to balance prevailing discussion. In response, many offer machine-centric rather than human-centric visions: semantically aware search engines, integrated e-commerce, and self-stocking fridges. These are worthy and challenging goals, but there is little discussion of the *people* required to make these systems function. Is the semantic web only about interoperable machines, or is there a vision of "augmenting human intellect", to pursue Doug Engelbart's forty year mission? In other words, how will the semantic web support learning, collaboration and knowledge sharing? (We note in passing that this was also Berners-Lee's original goal in designing the web.)

Towards ubiquitous knowledge acquisition

When we talk about the semantic web becoming pervasive, the implication is that knowledge acquisition (KA) will become ubiquitous. It is increasingly the case that if one's work/profile is not on the web, it is invisible. In 1997 one of us co-authored a paper which stressed the need for new forms of literacy in order to maintain one's presence in the emerging semantic web space *[5]*. One obvious concern must be with the user interfaces to enable individuals to participate in the world of the semantic web, and the expertise required to use them. New methods for KA are required to support this codification process. However, we must go beyond user interfaces and training. A closer examination of the "interface" between such technologies and the workplace (academic or business) quickly leads one to issues of trust and ownership, since the tools now embody, more explicitly and richly than ever, ontologies codifying perspectives on how the world works. We are talking about the cognitive, social and political interface that a semantic web infrastructure unavoidably has with communities of practice.

In our experiences with designing and implementing semantic web technologies for the workplace, we are grappling with a different order of questions to those raised at the infrastructure level. When semantic web technologies embrace elements of *human* understanding and interpretation, with possibly suspicious, non-technical stakeholders, we encounter questions such as: "Who gets to define the ontology? (and can we trust them?)"; "How can we be involved in the design process, when we don't understand these formalisms?"; "We're all too busy. Who's going to encode material? (and can we trust them?)"; "An ontology requires consensus. This domain is in flux. Do you still have anything to offer?".

The semantic web for "augmenting human intellect"

These are big issues, which we are beginning to tackle in a number of projects, all of which are based on an underlying mission: *How do we harness the semantic web to become an enabler for "augmenting human intellect"?* We sketch below a number of strategies that seem to hold promise.

Support conflicting interpretations and perspectives. What does the semantic web have to offer in domains where there is little consensus? In the Scholarly Ontologies project [www.kmi.open.ac.uk/projects/scholonto], we are developing a semantic digital library server that seeks to provide services for researchers, whose business is, of course, the construction and debating of world views [1]. Our approach is to provide a discourse ontology for making, extending and challenging claims—an ontology for disagreeing as much as for stable knowledge. Although currently being applied to research literatures, the underlying approach could be extended to any domain where it is as important to capture principled *disagreement* as it is to capture consensus.

Personalise information sources. There is plenty of information out there: the challenge is to find the meaning. Filtering based on semantics is one obvious way to assist intellectual work. In the Advanced Knowledge Technologies consortium [www.aktors.org] we have developed ontology-based support for personalizing an electronic newsletter. This integrates web, language and knowledge modelling technologies to enable readers of an electronic newsletter to easily define an interest profile according to several different dimensions, formalised in a shared ontology [3].

Provide active support for semantic markup to users with different levels of expertise. We are developing an annotation tool which provides different mechanisms for semantic markup by expert and naïve users. Expert users are given powerful editing/browsing facilities. Naïve users are supported by information extraction (IE) methods [6], which can be used directly from a library, or easily customised through the use of a simple interface for developing new IE methods. Finally, expert-level support for defining new IE methods are also provided.

Make semantic markup a by-product of work. The success of the web demonstrates that users are prepared to learn markup formalisms when the benefits are immediate and clear. It is as yet unclear whether the perceived benefits of semantic markup are so evident, and in any case manual markup will not scale. These issues are crucial in organizational contexts, where time and workload rule the day. However, if markup can occur as a by-product of ongoing work, sustained adoption will be more likely. Therefore our current work focuses on developing knowledge management solutions predicated on the above premise (i.e., transparent semantic markup). For example, we are currently developing technologies for our lab, which will automatically update a knowledge base (KB) when new reports are published, a new project is launched, new staff join, or other significant events take place. This KB will provide the main source for a query answering facility to allow visitors and staff to easily find information about the lab's work. The KB will integrate existing databases and will rely on information extraction methods to find information in relevant web pages. The process is ontology-driven, in the sense that the KB is based on an ontology which identifies the key concepts and relations required to describe academic life.

Another example of transparent markup is the *Compendium* approach to knowledge capture. The approach enables real time capture of knowledge arising in meetings, expressed by members of diverse communities of practice, integrating hybrid material into a reusable group memory. The shared, visual interface serves as a participatory front-end for eliciting information from domain experts who are not literate in knowledge representation. The resulting database can then be exported to knowledge-based tools, or to generate documents in requested formats [4].

Deliver customisable reasoning services to non-programmers. The semantic web is going to be about delivering sophisticated services to users. One project investigating this area is the *Internet Reasoning Service* [www.kmi.open.ac.uk/projects/irs] which provides a Web-based front-end enabling non-programmers to prototype knowledge-based applications quickly, using reusable components from distributed libraries. This approach will make artificial intelligence (AI) technology widely available, thus allowing non-AI experts to create knowledge-intensive services, either for experimentation purposes, for learning or to solve a problem. In addition, we also envisage that specialised configurations of the IRS will be developed for particular communities - e.g., we are working in the palaeontology domain and developing specialised ontologies and problem solving methods to support site interpretation.

Co-design ontologies with communities of practice. In this final example, we broaden the notion of "trust" on the semantic web beyond digital certification. Since significance resides in the interpretations placed on symbols by a given agent, semantic markup is always embedded in a perspective. Consequently, a semantic web application has integrity within a work setting only to the extent that its ontology is trusted. It is well established that systems are trusted when co-designed with the stakeholders expected to use them. A recent case study shows that when a community's perspective on an issue is stable, trusted knowledge services based on an ontology (embodying that perspective) can be co-constructed by knowledge engineers and stakeholders [2].

Further Reading

1. Buckingham Shum, S., Motta, E. and Domingue, J. ScholOnto: An Ontology-Based Digital Library Server for Research Documents and Discourse. *International Journal on Digital Libraries*, 3, 3, 2000, pp. 237-248 [www.kmi.open.ac.uk/projects/scholonto]
2. Domingue, J., Motta, E., Buckingham Shum, S., Vargas-Vera, M. and Kalfoglou, Y. Supporting Ontology Driven Document Enrichment within Communities of Practice. *Submission to K-CAP 2001: First Int. Conf. on Knowledge Capture*, Victoria, BC, Canada, 2001
3. Kalfoglou, Y., Domingue, J., Motta, E., Vargas-Vera, M. and Buckingham Shum, S. myPlanet: an ontology-driven Web-based personalised news service. *IJCAI'01 Workshop on Ontologies and Information Sharing*, Seattle, WA, 2001 [kmi.open.ac.uk/tr/papers/kmi-tr-102.pdf]
4. Selvin, A.M. and Buckingham Shum, S.J. Rapid Knowledge Construction: A Case Study in Corporate Contingency Planning Using Collaborative Hypermedia. *Journal of Knowledge and Process Management*, 2001, in press [kmi.open.ac.uk/tr/papers/kmi-tr-92.pdf].
5. Stutt, A. and Motta, E. Knowledge Modelling: An Organic Technology for the Knowledge Age. In: *The Knowledge Web: Learning and Collaborating on the Net,* Eisenstadt, M. and Vincent, T., (Ed.), Kogan Page: London, 1998, pp. 211-224
6. Vargas-Vera, M., Domingue, J., Kalfoglou, Y., Motta, E. and Buckingham Shum, S. Template-Driven Information Extraction for Populating Ontologies. *IJCAI'01 Workshop on Ontology Learning*, Seattle, WA, 2001 [kmi.open.ac.uk/tr/papers/kmi-tr-105.pdf]

International Semantic Web Workshop, Stanford University, July 30–31, 2001
Carolyn Sleeth, Research Analyst
SRI Consulting Business Intelligence
csleeth@bic.sri.com

The Semantic Web, Trust, and the Business Environment

Recent discussion in the business community has focused on the network economy and the change brought by the Internet-enabled enterprise. No longer can a company focus solely on itself—an independent unit—as its primary concern. Rather, it must focus on the partnerships and relationships it maintains with other companies. These relationships and their communication processes have become the keys to successful business.

In theory, the Internet makes these partnerships and relationships easier to build and maintain. The W3C's vision of the Semantic Web charts one of the many possible paths with which business uses can extend theory to reality. The Semantic Web will augment current Web content, which is primarily available in human-readable format only, to include machine-readable content. Machines can then use these data for automation and integration in a variety of applications, bypassing time-consuming human intervention. Technologies and protocols such as intelligent software agents, XML (Extensible Markup Language), RDF (resource description framework) schema, DAML (DARPA Agent Markup Language), and OIL (Ontology Inference Layer) enable this machine-readable Web. Making information on the Web machine readable ultimately means making the Web a more meaningful place for human users.

A major limitation inherent in a straightforward implementation of these technologies, however, is that they provide only for the communication of information. How do machines determine whether the source of the information is trustworthy? Within a single enterprise, sources of information are strictly under the control of the company's information technology department. For example, on a traditional factory floor, a control system knows exactly which fieldbus it is communicating with. It operates with the knowledge that the control system's information is qualified and trustworthy. Even in a closed, interbusiness supply chain, partners have considerable control in making certain that sources are reliable. But the Semantic Web will enable an expansion beyond closed communities, and devices will be able to communicate and perform transactions with other devices that they have never met. The genie that opens once-closed corporate networks to external communications also opens a Pandora's box of undifferentiated, unqualified sources. One of the major problems that most users, both business and consumer, face today is the proliferation of low-quality data and misinformation.

How can devices know that particular information is high quality—that it reliably and correctly serves the task at hand? For an intelligent device to operate usefully, its data inputs must be certified as accurate and of high quality. For many applications involving transactions between parties with no foreknowledge of each other, a trusted third party must assure that the correlation of the information to the real world is reliable. Only then does a device or system have the knowledge to proceed, reliably, in its task.

What can provide this assurance? At the XML 2000 conference, Tim Berners-Lee pointed out that the incorporation of digital signatures in the information architecture of the Web could resolve the "how" of this problem. Digital signatures, in combination with

RDF schemas and ontologies can provide the means not only to authenticate information but also to automate much of the filtering process, automatically eliminating vast quantities of data from consideration, according to the quality expectations or needs of the user. Such a system has the potential to bring an element of trust to the Web, with implications for both businesses and consumers.

The integration of digital signatures into the Web information architecture would bring a much-needed granularity to the digital-signature and public-key-infrastructure (PKI) cryptography systems on the Web. The current system operates in a binary manner, allowing only trusted or not-trusted designations for specific sources or transactions. The Semantic Web will allow users to designate trust for a certain source in only certain domains or in only specific areas of a particular ontology, or in certain types of transactions. Various combinations of trusted sources from a diverse set of information providers in different domains of expertise can result and will enrich the user's Web information and electronic-commerce environment.

Within limited, closed communities, such as intercompany supply chains, users themselves are able to define who they trust as information providers in particular domains of expertise. As communication expands beyond closed communities, however, users looking to corroborate their information will need trusted third parties to help determine whether the information is reliable or not.

Trusted third parties are already emerging in a number of industries. Businesses that have already earned trust in a corporate setting see an easy entry into the role of a trusted third party for online interactions. This role could easily extend into an authority that devices in different industries automatically and electronically turn to in order to authenticate the unfamiliar source. Organizations that are currently investigating opportunities as trusted third parties include banks, online auctioneers, dispute-resolution services, health-information banks, security-system vendors, credit-card companies, postal services, and government agencies. PKI vendors, such as Verisign, also pitch themselves as trusted authorities that can confirm identities of individuals or corporations.

Within their own industries, these enterprises have positioned themselves as authorities in certain domains. The digital-signature capabilities of the Semantic Web will allow them to communicate their authentication of information—does it have a reliable correlation to the real world—to the devices receiving it. This authentication of information—the award of a digital certificate—gives the information the backing of the trusted third party. If a device accessing the Web finds information certified by an appropriate digital certificate (as specified by the device's user), the device can proceed in interactions with a level of authority. The input from the trusted third party enables machines to interact with each other in a vastly more useful, reliable, and trusted way.

Berners-Lee's vision of incorporating digital signatures at several architectural levels of the Semantic Web will be an essential element for the continued success of the Web as unqualified, undifferentiated data proliferate more rapidly than qualified data on the Web. Digital-signature capabilities will be particularly important in the business environment as corporations attempt to break out of their private networks to build ad hoc networks and alliances and to expand their reach to unfamiliar partners and customers.

Position Paper:
How the Semantic Web can work with the Web Services?

Naohiko Uramoto (uramoto@jp.ibm.com)
IBM Research, Tokyo Research Laboratory and
National Institute of Informatics

The goal of the Semantic Web is to make the Web machine processable by using:

- Sets of assertions with RDF
- Inference
- Ontology

The Semantic Web is a very emerging technology for modeling web of the Web. However, the another trend of modeling the Web is Web Services based on:

- Sets of operations (functions) with SOAP
- Service description and binding
- Directory (UDDI)

Many companies including Microsoft, IBM, HP, and so on are very keen on supporting the Web Services in their products. The goal of the two is the same, but they have very different approaches.

My interest is that what is the essential difference between them, and how to integrate the two technologies, since they should not grow separately.

We can imagine how to integrate them.

- The Semantic Web on the Web Services. We can define a Web services that provide assertions and ontology information. By using discovery framework provided by UDDI, we can find appropriate assertions without knowing where they are.
- The Web Services on the Semantic Web. By processing assertions and ontology information, we can provide intelligent services. It is very useful in the real world, since many business middleware will support Web services soon. At this moment, many useful tools are available for free.

There are some common constructs for both technologies. For example, digital signature is one of very important for them. Semantic Web requires it for establishing a "trust model" to qualify assertions. Web services needs it since security is an important building block for developing real B2B or B2C systems.

Another my interest is how to make trust system on the Web space. As many people pointed out, keeping consistency of assertions in distributed environment is very difficult. Digital signature is a good tool, but it is not enough.

We are now developing a set of core security services including digital signature, encryption, and access control. These services can be applied to Semantic Web applications. Digital signature is used to qualify assertions and protect integrity of them. Encryption is required to keep confidentiality of valuable assertions. Access control for any elements of XML documents are very important technology to maintain assertions. The services are being developed based on standard technologies from W3C, IETF, and other standard bodies. We are using open-source tools such as Apache XML tools.

Semantic Web at this moment is a vision, but we should create something that can run in the real world (like XML and HTML). To develop real tools, I think we need a shared "playground" for developing tools and evaluating new ideas. The playground provides a set of assertions written in RDF and other language, database, and ontology. Only small environment is enough for early stage, but we must improve scale of it since scalability is very important for the Semantic Web research.

ONTOSERVER — Infrastructure for the Semantic Web

Raphael Volz
Institute AIFB, University of Karlsruhe, 76128 Karlsruhe, Germany
volz@aifb.uni-karlsruhe.de
http://www.aifb.uni-karlsruhe.de/WBS

1 Motivation

Ontologies build the conceptual backbone of the Semantic Web. By providing a shared and common understanding of a domain they can be communicated across people and application systems for facilitating knowledge sharing and exchange. In the context of the Semantic Web ontologies describe domain theories for the explicit representation of the semantics of relational (meta-)data.

As ontologies aim at consensual domain knowledge their development requires a cooperative process which has to be supported by central repositories that enable persistent storage and concurrent access. We believe that many Semantic Web applications are also charaterized by concurrent access (i.e. building a Web portal to support the Semantic Web Research Community[1]). Part of the Semantic Web infrastructure currently implemented at our institute is ONTOSERVER, a multi-user capable metadata and ontology repository that uses external inference engines to provide model checking and querying.

In our design we identified the following general requirements for a central repository:

- **Persistence** A repository must prevent data loss by providing persistent storage of data.
- **Update semantics** Data must be updateable, users must be able to provide new data (concering ontology schema data as well as ontology instance data)
- **Concurrency** A repository must be capable of providing concurrent access to its data.
- **Security** In a networking environment the feature of security is of very high importance.

In the Semantic Web the technology of choice for exchanging and storing ontologies is the Resource Description Framework (RDF). Its abstract data model provides a fundamental standard for the Semantic Web and especially appeals because of its generality and extensibility.

RDF can be used to represent all data occurring in ontological applications (cf. figure 1):

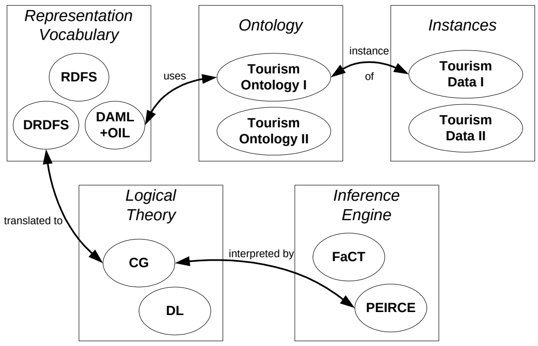


Figure 1: Representation Vocabularies, Ontologies and Instances

1. **Representation Vocabularies** that describe ontology languages (like DAML+OIL[2], DRDFS [DFZD01] or the RDF Schema proposal [BG99]) and usually have semantics based on a logical theory that can be interpreted by a inference engine.
2. **Ontologies** that use a specific representation vocabulary and represent a domain theory.
3. **Relational Metadata** that uses the vocabulary specified by an ontology and provides instance data for the domain model.

Recently, RDF repositories like RDFdb[3], RDFSuite[4] or Sesame[5] among others have been developed. These systems can be used as a starting point to implement the above mentioned requirements. In fact, due to RDF's generality RDF frameworks in general can be reused across a wide range of Semantic Web applications to provide generic data access and storage components.

2 Why a RDF repository is not enough!

Although it is possible to store all data (representation vocabularies, ontologies and relational (meta-)data) related to a Semantic Web application in one "bag", the distinction of this data is of very high importance in applications, especially in ontology engineering scenarios. A typical task in these scenarios is performing consistency checks of ontologies. This involves structural and semantic checking of the conformance

[1] You are welcome to visit http://www.ontoweb.org/
[2] cf. http://www.daml.org/2001/03/daml+oil-index
[3] cf. http://www.guha.com/rdfdb/
[4] cf. http://www.ics.forth.gr/proj/isst/RDF/
[5] cf. http://sesame.aidministrator.nl/

of a given set of relational (meta-)data towards an ontology as well as conformance of an ontology towards a given representation vocabulary. This tasks poses special challenges to RDF repositories as a repository would have to understand the formal semantics of a given representation vocabulary to perform this task. Other tasks that require understanding of semantics are inferencing and querying of RDF data. Interestingly these common tasks are not available in current RDF repositories.

When provided these services should also be scalable and performant. Usually this implies usage of highly optimized (thus specific) structures as found in today's inference engines. This is also true for database technology, where a second important property can be identified: Usually any ontology modification (comparable with schema modification in databases) can only happen when the part of the database affected by the change is not available to users. This is due to the fact that ontology (= schema) modification normally implies data modification. Also, optimization techniques usually depend on schema and access structure (for instance choosing hash tables or b-trees for indexing).

As the number of possible representation vocabularies whose semantics would have to be understood by a repository is infinite we conclude that it is not possible for a single repository to provide this for all possible representation vocabularies.

Therefore, we propose to build a comprehensive infrastructure around a basic repository. We believe that it is beneficial (especially for less tractable representation vocabularies) to use external inference engines (like SilRI [DBSA98] or FaCT [Hor99]) to deploy a given RDF model whose ontology (schema) is fixed[6]. Then, a specific inference engine could be used for querying and checking of given RDF models. In general, the integration of such external inference engines should be hidden from the users of such infrastructure, as this setup clearly is an administrator task. To implement this, we are dynamically registering inference engines to the XML namespace of given representation vocabularies and allow to dynamically locate service descriptions that wrap engine specific interfaces to ONTOSERVER access interfaces.

3 ONTOSERVER – Infrastructure for the Semantic Web

Due to lack of space, we can only briefly mention the rationale that lead our architecture of ONTOSERVER. The general design reflects the reusability of ONTOSERVER's components for stand-alone Semantic Web applications. Therefore we introduced three levels:

1. The application framework level contains all components that can be reused in non-server applications, specifically data access and storage components as well as external service connectors.
2. The server framework level contains additional components specific to server applications, this includes user/rights/security management and transaction facilites.
3. The Application level contains custom applications and provides one prototypical implementation: ONTOSERVER itself.

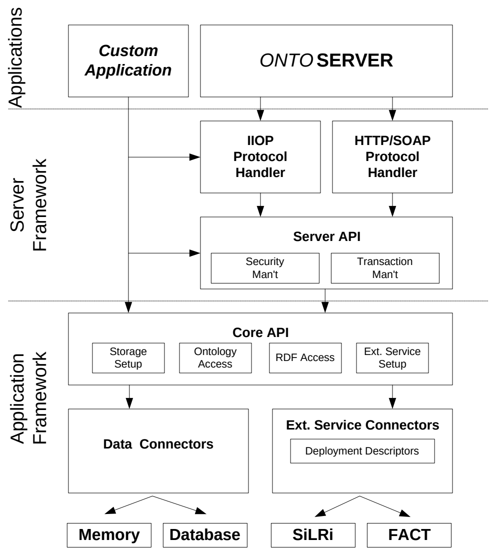


Figure 2: ONTOSERVER: General architecture

4 Related Work

Protégé is an ontology engineering system that supports the flexible definition and instantiation of different representation vocabularies has been introduced in [NSD+01]. However the internal data model does not build on RDF, thus it's components can not be reused to build Semantic Web applications. Interestingly, the formal semantics of representation primitives can only be provided outside the tool, to do this users must write plug-ins. In the future Protégé's plug-in mechanism opens the possibility to use Protégé as an editing client for ONTOSERVER.

5 Future Work

A general challenge is the interoperability between different representation vocabularies and concrete ontologies along with translation of relational metadata instances. As [DvHB+00] mentioned, RDF's "Subject-Predicate Structure leads to independance of objects". It is therefore technically possible to provide mappings between two RDF models, a feature would that would ultimately leverage the Semantic Web.

Another challenge is posed by the fact that ontologies are rarely stable following initial creation especially in dynamic settings like the Web, where content changes with light-speed. Therefore versioning and controlled evolution of metadata is a requirement for the Semantic Web.

Last but not least many of our applications require to have views on metadata that present virtual subsets of one model to special audiences.

[6] The reader may note, that this could also be the quiescent state (version) of a submodel (view) that has been tailored to meet the requirements of a certain application or audience.

References

[BG99] D. Brickley and R.V. Guha. Resource description framework (RDF) schema speci£cation. Technical report, W3C, 1999. W3C Proposed Recommendation. http://www.w3.org/TR/PR-rdf-schema/.

[DBSA98] S. Decker, D. Brickley, J. Saarela, and J. Angele. A Query and Inference Service for RDF. In *Proceedings of the W3C Query Language Workshop (QL-98), http://www.w3.org/TandS/QL/QL98/, Boston, MA, December 3-4*, 1998.

[DFZD01] Alexandre Delteil, Catherine Faron-Zucker, and Rose Dieng. Extension of rdfs based on the cgs formalism. In *Conceptual Structures - Broadening the Base. Proc. ICCS 2001. Lecture Notes in Arti£cial Intelligence 2120.*, pages 272–285, 2001.

[DvHB+00] S. Decker, F. van Harmelen, J. Broekstra, M. Erdmann, D. Fensel, I. Horrocks, M. Klein, and S. Melnik. The semantic web: The roles of XML and RDF. *IEEE Internet Computing*, 4(5), 2000.

[Hor99] I. Horrocks. FaCT and iFaCT. In *International Workshop on Description Logics (DL'99)*, pages 133–135, 1999.

[NSD+01] N. F. Noy, M. Sintek, S. Decker, M. Crubezy, R. W. Fergerson, and M. A. Musen. Creating semantic web contents with protege-2000. *IEEE Intelligent Systems*, 16(2), 2001.

Semantics for Multimedia on the Web

B.J. Wielinga A. Th. Schreiber J. Wielemaker

University of Amsterdam, Social Science Informatics
Rotersstraat 15, NL 1018 WB Amsterdam, The Netherlands
E-mail: {wielinga,schreiber,jan}@swi.psy.uva.nl

Position paper for the Semantic Web Workshop SWWS
Stanford 30-31 July 2001

The vision of the Semantic Web entails that large amounts of multimedia data should be annotated with semantic meta data. Current technology for content-based image interpretation is too limited for automated annotation of visual material. Techniques used by image search engines are also very poor and are unlikely to be improved in the near future. So, human annotations are required to make large annotated image corpora available on the Semantic Web. Currently, image archives use human annotators to select a set of keywords to describe an image. There are several problems with this approach: consistency of vocabulary is hard to enforce, keywords cannot describe relational properties of the images and the keyword approach does not scale up. Searching an image archive such as GETTYONE[1] which contains some 350.000 images, with the keyword `ape` delivers 575 hits. The Lycos multimedia engine finds over 5000 images for the query `GORILLA`. More specialized queries tend to give unstable results due to inconsistent indexing methods. Searching the full Semantic Web with tenths of millions of images with a simple keyword approach is not realistic anymore. A more sensitive method for creating semantic annotations for multimedia material is needed.

We are developing a paradigm for indexing and retrieval of images and other multimedia materials for the Semantic Web. The paradigm is based on the following key notions:

- Annotations are built from structured sets of descriptive elements, which are extensions of the Dublin Core [1] and VRA 3.0 [2] metadata element sets.
- Relational information is represented using complex instantiated structures, resembling sentence structures [3].
- Annotations can be hierarchically structured. The subject slot of an image description can contain references to other descriptions of objects in the image.
- Existing thesauri such as WordNet and AAT [4] are used as sources for closed vocabularies.
- The thesauri are augmented with additional knowledge to create proper ontologies. These ontologies are used to infer additional information from a partial description.
- Annotation tools and retrieval tools (search engines and browsers) are automatically configured on the basis of the ontology.

[1] UEL: http://www.gettyone.com.

- Representation of the ontologies and the indexed material, and implementation of the tools is compliant with current W3C standards, notably RDF, RDFS, and the usability of emerging standards such as RULE-ML are being investigated.

In an earlier project on stolen art objects [5] we have used this approach (without the RDF representations) successfully for the structured description of stolen art and antique objects. More recently we are applying the paradigm in a number of domains (photographs of animals, landscapes, buildings, paintings) [6]. In these experiments a number of tools were used and developed. We use Protg 2000 [7] for construction of the ontologies and for generating RDFS data. The Protg WordNet plugin was used to select certain parts of the WordNet ontology that were relevant to a particular domein. An RDFS version and browser for the Art and Architecture Thesaurus was developed and integrated with the description ontologies. An image annotation toolkit that automatically generates a user interface from the RDFS specification of an ontology was used in a number of the experiments. A preliminary conclusion from the experiments is that the highly structured annotation of images allows a much more fine-grained retrieval of images than the standard keyword approach. The use of RDFS technology turned out to be staifactory, with the exception of the representation of inference rules [8]. In structured annotations it is often possible to aid the annotation process by limiting the set of possible values based on constraints. In many situations it is also feasible to provide sensible defaults based on a partial annotation to speedup the annotation process. This requires knowledge about constraints between properties as well as default knowledge, neither of which can be expressed in RDFS.

References

[1] Dublin Core Metadata Initiative. *Dublin Core Metadata Element Set Version 1.1: Reference Description*, July 1999. Url: http://dublincore.org/documents/1999/07/02/dces/.

[2] Visual Resources Association Standards Committee. VRA core categories, version 3.0. Technical report, Visual Resources Association, July 2000. URL: http://www.gsd.harvard.edu/ staffaw3/vra/vracore3.htm.

[3] A. M. Tam and C. H. C. Leung. Structured natural-language description for semantic content retrieval. *Journal of the American Society for Information Science*, to appear.

[4] T. Peterson. *Introduction to the Art and Architecture Thesaurus*. Oxford University Press, 1994. See also: http://shiva.pub.getty.edu.

[5] B. J. Wielinga, J. A. C. Sandberg, and A. Th. Schreiber. Methods and techniques for knowledge management: What has knowledge engineering to offer? *Expert Systems With Applications*, 13(1):73–84, 1997.

[6] A. Th. Schreiber, B. Dubbeldam, J. Wielemaker, and B. J. Wielinga. Ontology-based photo annotation. *IEEE Intelligent Systems*, 2001. To appear. URL: http://www.swi.psy.uva.nl/usr/Schreiber/Schreiber01a.pdf.

[7] N. Fridman Noy, R. W. Fergerson, and M. A. Musen. The knowledge model of Protégé-2000: combining interoperability and flexibility. In *Knowledge Engineering and Knowledge Management: 12th International Conference EKAW2000, Juan-les-Pins*, volume 1937 of *Lecture Notes in Artificial Intelligence*, pages 17–32, Berlin/Heidelberg, 2000. Springer-Verlag. Also as: Technical Report Stanford University, School of Medicine, SMI-2000-0830.

[8] B. J. Wielinga, A. Th. Schreiber, J. Wielemaker, and J. A. C. Sandberg. From thesaurus to ontology. Report ICES-MIA project, University of Amsterdam, Social Science Informatics, 2001. Submitted for publication. URL: http://www.swi.psy.uva.nl/usr/Schreiber/Wielinga01a.pdf.

Ontologies:

Dynamic Networks of Formally Represented Meaning

Dieter Fensel

Vrije Universiteit Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands,
phone/fax: +31-(0)84-872 27 22, dieter@cs.vu.nl, http://www.cs.vu.nl/~dieter.

Abstract. The computer was invented as a device for computation. Meanwhile the "computer" becomes a portal to cyberspace. It has become an entry point to a world-wide network of information exchange and business transactions. Therefore, technology that supports access to unstructured, heterogeneous and distributed information and knowledge sources will become as essential as programming languages were in the 60's and 70's. In this essay, we examine some of the essential requirements for such a technology.

1 Ontologies: Formal and Real, based on Consensus

The World-Wide Web (WWW) has drastically changed the availability of electronically available information. Currently there are around one billion documents in the WWW which are used by more than 300 million users internationally. In addition, this number is growing fast. However, this success and exponential grow makes it increasingly difficult to find, to access, to present, and to maintain the information of use to a wide variety of users. Currently, pages on the web must use representation means rooted in format languages such as HTML or SGML and make use of protocols that allow browser to present information to human readers. The information content, however, is mainly presented by natural language. Thus, there is a wide gap between the information available for tools that try to address the problems above and the information kept in human readable form. The current state of Web technology generates serious obstacles to its further growth. The technology's simplicity already caused bottlenecks that hinder searching, extracting, maintaining, and generating information (cf. [Fensel et al., 2000]). Computers are only used as devices that post and render information, but they do not have access to the actual content. Thus, they can only offer limited support in accessing and processing this information.[1] So, the main burden not only of accessing and processing information but also of extracting and interpreting it is on the human user.

Tim Berners-Lee envisioned a *Semantic Web* (cf. [Berners-Lee et al., 2001], [Fensel et al., to appear (b)]) that provides automated information access based on machine-processable semantics of data and heuristics that use these meta data. The explicit representation of the semantics of data, accompanied with domain theories (that is, ontologies), will enable a Web that provides a qualitatively new level of service. It will weave together an incredibly large network of human knowledge and will complement it with machine processability. Various automated services will help the user achieve goals by accessing and providing information in a machine-understandable form. This process might ultimately create an extremely knowledgeable systems with various specialized reasoning services-systems that can support us in nearly all aspects of our life and that will become as necessary to us as access to electric power.

Ontologies (cf. [Fensel, 2001]) are key enabling technology for the semantic web. They need to interweave human understanding of symbols with their machine processability. Therefore, it seems highly justified to take a closer look on the nature of Ontologies and on whether and how they can actually provide such a service. Ontologies were developed in Artificial Intelligence to facilitate knowledge sharing and reuse. Since the beginning of the nineties Ontologies have become a popular research topic investigated by several Artificial Intelligence research communities, including Knowledge Engineering, natural-language processing and knowledge representation. More recently, the notion of Ontology is also becoming widespread in fields such as intelligent information integration, cooperative information systems, information retrieval, electronic commerce, and knowledge management. The reason ontologies are becoming so popular is in large part due to what they promise: *a shared and common understanding of a domain that can be communicated between people and application systems*.

Because Ontologies aim at consensual domain knowledge their development requires a cooperative process. Ontologies are introduced to facilitate knowledge sharing and reuse between various agent, no matter whether they are of human or artificial nature. They should provide this service by providing a consensual and formal conceptualizations of a certain area. Spoken in a nutshell, *Ontologies are formal and consensual specifications of conceptualizations providing a shared and common understanding of a domain that can be communicated across people and application systems*. Therefore, Ontologies glue together *two essential aspects* that help to bring the web to its full potential:

- Ontologies define a *formal* semantics for information allowing information processing by a computer.
- Ontologies define a *real-world semantics* allowing to link machine processable content with meaning for humans based on *consensual* terminologies.

These two orthogonal aspects will be discussed during the following, however, our main focus is devoted to the second aspect.

---

1. It is like using a telephone mainly for decorating a living room.

2 Ontologies define formal semantics

Ontologies provide formal semantics enabling machine-processable semantics of information. This aspect is already well-understood and several language proposals have been made (see [Fensel, 2001] and [Fensel et al., 2001]). Formal semantics is achieved by a layered language architecture. At the lowest level, XML[2] provides a serialized *syntax* for tree structures. RDF[3] defines a basic *data model* on top of XML consisting of (object, property, value)-triples. RDF schema (RDFS)[4] defines basic ontology primitives in RDF: classes with is-a and instance-of relationships, and properties with is-a relationships and domain and range restrictions. OIL[5] extends RDFS to provide a full-fledged web-based ontology language. One of the central design ideas of OIL is its onion model (see Figure 1). There will never be one language meeting all man purposes. OILs onion model reflects this need. Languages of different complexity are provided allowing applications to select the degree of complexity they require. One of its dialects called **DAML+OIL**[6] reflects a broad European and (US) American consensus on modeling primitives for the semantic web and is departure point for standardization by the W3C[7].

3 Ontologies define real-world semantics

This aspect is still far from being studied properly. In this essay, we will focus on it, i.e., on how can Ontologies be used to communicate real-world semantics between human

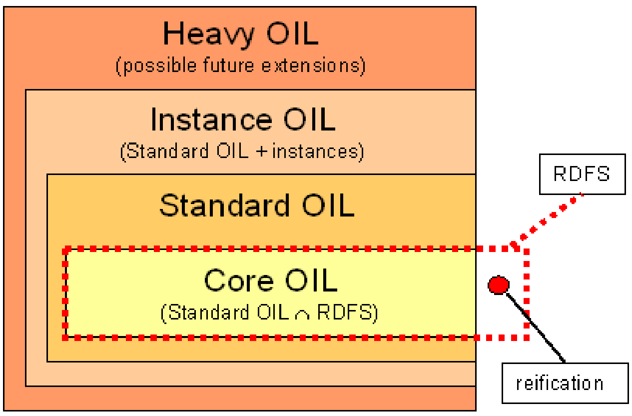


Fig 1. The onion model to control complexity.

2. http://www.w3.org/XML/
3. http://www.w3.org/RDF/
4. http://www.w3.org/TR/2000/CR-rdf-schema-20000327/
5. http://www.ontoknowledge.org/oil
6. http://www.daml.org
7. http://www.w3c.org

and artificial agents. For understanding this potential, we have to bring in an important point on how to look at them. This point of view is required to bring Ontology technology to its full potential and also brings into mind that most of the work on Ontologies is partially miss-focussed, i.e., ignores the main problems in building and using them.

Every first years philosophy student may have heart about the evils circle in trying to explain our ability of communication as a way to exchange meaning and to create understanding between human beings. On the one hand, people can only communicate and exchange meaning based on a common understanding of symbols and intensions. Therefore, a joined set of symbols and a consensual interpretation is the pre-request for communication. On the other hand, such a joined set of symbols and a consensual interpretation can only be established as a result of communication.[8] Therefore, what is a result of successful communication is at the same time a pre-requisite for it. In consequence, its existence is required for explaining its existence. Our first years philosophy student may also have learned how to overcome such a paradoxical situation. There must be an underlying *process* that takes both sides as intermediate and repeatedly taken sub-steps relying on something that mediates between its extremes.[9] Then successful communication and a joined set of understandings are just two sites of the same coin. The reader may found our arguments “too” philosophical. However, we want to undermine the principal difference between viewing ontologies as *“true” models of the real world* or steps in a *process of organizing evolving consensus*. Therefore, a brief argument on the cyclic nature of understanding and communication seems appropriate.

Viewed from an abstract philosophical point of view it looks like a miracle that two humans are able to understand each other. Taken in the extreme we cannot even be sure about our mutual existence. Since Descartes we take the fact that we are aware of our own thinking as proof of our own existence.[10] However, we make notice of the existence of other agents via our perception and it is their existence in our perception and not their actual existence that follows from it.[11] Again we have to make the doubtful deduction that their existence in our perception reflects their actual behavior and existence. Even taking this assumption to be grounded we are still far away from explaining on how meaning can be exchanged between such brittle agents. Meaning and intention cannot be exchanged *directly*. Neither can it be expressed directly nor can we access the actual meaning that is perceived and understood by our counter part. We can only express our intension by some action that influences the perception of our counter part. And we can only guess what this is supposed to mean to him by analyzing his behavior as much as it is reflected in our perception.[12] In consequence, establishing meaning and communication (to exchange

8. At least as long as this interpretation is not hard-coded via instincts.

9. Cf. G. W. F. Hegel: *Wissenschaft der Logik*.

10. Already this conclusion could be viewed as being doubtful, however, its discussion would leave the scope of this paper.

11. See for example I. Kant: *Critik der reinen Vernunft*.

12. In principle, it is not even important whether another agent actually thinks. He “understands” our communicative acts properly if it is properly contained in the way he cooperates with us.

meaning) is per definition a *process*. People can only establish joined meaning and communicate it to each other in a process where they co-ordinate some of their actions to achieve common goals. Therefore, from the early beginning[13] it can only be a social process that creates a joined understanding that is the basis for exchanging meaning with communicative symbols.

Following this argument it is also rather clear that there will be neither such a thing as THE Ontology where everybody subscribes to. Instead, ontologies arise as pre-requisite and result of cooperation in certain areas reflecting task, domain, and sociological boundaries. In the same way as the web weaves billions of people together to support them in their information needs, Ontologies can only be thought as a network of interweaved Ontologies. This network of Ontologies may have overlapping and excluding pieces, and it must be as dynamic in nature as the dynamics of the process it underlies. This view on *Ontologies as dynamic networks of formally represented meaning* is what we want to stress in the essay. Most work on Ontologies view Ontologies as a isolated theory containing possible large number of concepts, relationships, and constraints that further detach formal semantics to them. Here we take a much broader view on Ontologies. Basically, there are two main dimensions in which these mediators of communication differ from current work on Ontologies: *Ontologies must have a network architecture* and *Ontologies must be dynamic*.

3.1 Heterogeneity in Space: Ontology as Networks of Meaning

Island of meaning must be interwoven to form more complex structures enabling exchange of information beyond domain, task, and sociological boundaries. This implies two efforts. Tool support must be provided to define local domain models that express a commitment of a group of agents that share a certain domain and task and that can agree on a joined world view for this purpose. Here much work has already been spent and significant methodological support is available (see [Fensel et al., to appear (a)] for a survey). Second, these local models must be interwoven with other models like the social practice of the agents that use Ontologies to facilitate their communicational needs. Here not much work has been spent. We do no longer talk about a single Ontology but rather about a network of Ontologies. Links must be defined between these Ontologies and this network must allow overlapping Ontologies with conflicting and even contradictionary conceptualizations. From the early beginning heterogeneity is an essential requirement for this Ontology network. Means to deal with conflicting definitions and strong support in interweaving local theories are essential requirements for making this technology workable and scalable.

Take a Peer-to-Peer (P2P) network like Gnutella as an example (cf. [Oram, 2001]). Agents can dynamically enter and leave the network. Agents can communicate with a

13. Both, in a *historical* and in a *logical* sense.

local environment of other agents. This network is dynamically set up and collapsed according to the joined needs of a group of agents. Current work on Ontologies that focuses either on local domain theories or on principles, structures, and content of the right upper-layer Ontology are far way from supporting such a vision. What is needed is focus on:

- *linking local conceptualizations* dealing with heterogen definitions and personalized views,
- support in easy *configuration and re-configuration of such networks* according to the communication needs of agent coalitions, and
- methods and tools that help agents in *organizing consensus* allowing them to exchange meaning.

Ontologies ensure communication between various agents. They are "right" if they fulfill this purpose.

3.2 Development in time: Living Ontologies

Originally, an Ontology should reflect the "truth" of a certain aspect of reality. It was the holy task of a philosopher to find such truth. Nowadays Ontologies are used as means to exchange meaning between different agents. They can only provide this if they reflect an inter-subjectual consensus. Per definition they can only be the result of a social process. This gives ontologies a dual status for the exchange of meaning.

- Ontologies as *pre-requisite* for consensus: Agents can only exchange meaning when they have already agreed on a joined body of meaning reflecting a consensual point of view on the world.
- Ontologies as a *result* of consensus: Ontologies as consensual models of meaning can only arise as result of a process where agents agree on a certain world model and its interpretation.

In consequence, ontologies are as much a pre-requisite of consensus and information sharing as they are its results. Therefore, ontologies cannot be understood as a static model. An ontology is as much required for the exchange of meaning as the exchange of meaning may influence and modify an ontology. In consequence, *evolving* ontologies rather describe a process than a static model. Having protocols for the process of evolving ontologies is the real challenge. Evolving over time is an essential requirement for useful ontologies. As the daily practice constantly changes, Ontologies that mediate the information needs of these processes must have strong support in *versioning* and must be accompanied by *process models* that help to organize consensus.

Centralized process models have standardization bodies as central clearing unit. This central unit may soon become a bottleneck for the scalability of the entire process. Often such standardization works slow and lead to mongrelized results. Decentralized process models for consensus achievement can be based on the natural consensus of working networks. Therefore, they can reflect true, proven useful, and broadly used consensus. In this context, one may want to take a look at P2P, where networks arise and are configured dynamically according to joined interests of loosely coupled groups.

4 Conclusions

Ontologies help to establish *consensual terminologies* that make sense to both sites. *Computers* are able to process information based on their machine-processable semantics. *Humans* are able to make sense of this information based on their connection to real-world semantics. Building up such ontologies that are pre-requisite and result of joined understanding of large user groups is far from being trivial. A model or "protocol" for driving the network that maintains the process of *evolving Ontologies* is the real challenge for making the *semantic* web reality.

Most work on Ontologies view ontologies as a isolated theory containing possible large number of concepts, relationships, and constraints that further detach formal semantics to them. In the paper we took a much broader view on ontologies. We view Ontologies as highly interwoven *networks* allowing to deal with heterogenic needs of the communication processes that should mediated by them. Second, these ontologies must shift over time as the processes they mediate based on consensual representation of meaning. It is the network and dynamic character of *Ontologies* that make further research work on them so exiting. The *glue*, that link together Ontology networks in space and time, is the actual challenge on current work on ontologies. It is the glue, stupid!

References

[Berners-Lee et al., 2001]

T. Berners-Lee, J. Hendler, and O. Lassila: The Semantic Web, *Scientific American*, May 2001.

[Fensel, 2001]

D. Fensel: *Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce,* Springer-Verlag, Berlin, 2001.

[Fensel et al., 2000]

D. Fensel, F. van Harmelen, H. Akkermans, M. Klein, J. Broekstra, C. Fluyt, J. van der Meer, H.-P. Schnurr, R. Studer, J. Davies, J. Hughes, U. Krohn, R. Engels, B. Bremdahl, F. Ygge, U. Reimer, and I. Horrocks: OnToKnowledge: Ontology-based Tools for Knowledge Management. In *Proceedings of the eBusiness and eWork 2000 (EMMSEC 2000) Conference,* Madrid, Spain, October 2000.

[Fensel et al., 2001]
D. Fensel, I. Horrocks, F. van Harmelen, D. McGuiness, and P. Patel-Schneider: OIL: Ontology Infrastructure to Enable the Semantic Web, *IEEE Intelligent Systems*, March/April 2001.

[Fensel et al., to appear (a)]
D. Fensel, J. Hendler, H. Lieberman, and W. Wahlster: The Semantic Web: Why, What, and How. In [Fensel et al., to appear (b)].

[Fensel et al., to appear (b)]
D. Fensel, J. Hendler, H. Lieberman, and W. Wahlster (eds.): *Semantic Web Technology*, MIT Press, Boston, to appear.

[Oram, 2001]
A. Oram (eds.): *Peer-to-Peer: Harnessing the Benefits of a Disruptive Technology*, O´Reilly, Sebastobol, 2001.

Index