

# Enabling the Semantic Web for Scientific Research And Collaboration

## NSF Post Digital Library Futures Workshop Paper

[Eric Miller](mailto:em@w3.org), em@w3.org  
Semantic Web Activity Lead  
W3C World Wide Web Consortium

### Introduction

The January 2003 NSF report "Revolutionizing Science and Engineering Through Cyberinfrastructure" led off with the observation that "multiple accelerating trends are converging and crossing thresholds in ways that show extraordinary promise ... in how we create, disseminate, and preserve scientific and engineering knowledge. That study concluded that the "National Science Foundation should establish and lead a large-scale, interagency, and internationally coordinated Advanced Cyberinfrastructure Program (ACP) to create, deploy, and apply cyberinfrastructure in ways that radically empower all scientific and engineering research and allied education. It envisioned a program "to build more ubiquitous, comprehensive digital environments that become interactive and functionally complete for research communities in terms of people, data, information, tools, and instruments and that operate at unprecedented levels of computational, storage, and data transfer capacity.

The World Wide Web is an instrument for enhancing and extending human communication. With design choices tailored to the rapidly expanding Internet, the Web has achieved notable success in scientific communication as well as in commercial and business communication. The Web as it exists today has realized only a part of the potential that it is capable of providing. Today's Web is successful because humans have populated it with documents intended primarily for other humans to read; that is, the Web has provided a very social communications infrastructure.

The Semantic Web is an enhancement of the current Web to allow machine-processable data to span application boundaries in the same way that human-readable documents do currently. The Semantic Web extends the deployed infrastructure of HTTP (*HyperText Transfer Protocol*) [HTTP] and URIs (*Universal Resource Identifiers*) [URI]. In one of its several possible representations, data in the Semantic Web is exchanged using XML (*Extensible Markup Language*) [XML]. XML permits a single container to communicate both presentation markup for humans as well as semantic markup for data specific to software systems, including Semantic Web applications.

The goal of the Semantic Web initiative is as broad as that of the Web: to be a universal medium for data. It is envisaged to smoothly interconnect personal information management, enterprise application integration, and the global sharing of commercial,

scientific and cultural data. It is in this way that the Semantic Web may be viewed as an infrastructure for supporting the objectives outlined by the January 2003 NSF report.

## **Research Agenda for Digital Libraries**

Digital Libraries are an important component for shaping the infrastructure necessary for supporting the objectives outlined by the January 2003 NSF report. The following are a set of research areas for the Semantic Web which directly relate to the Digital Library community and is hoped will have a significant impact in addressing. This overlap will not be surprising to those familiar with Digital Libraries contribution the early design of the Semantic Web. These areas of overlapping research may be helpful in forming an agenda for Digital Libraries to facilitate the scientific research and collaboration in a global ubiquitous information space.

### **Enabling an Information Infrastructure**

The Semantic Web aims to build on the Web's contribution of *Uniform Resource Identifiers* [URI] and resilience to failures of closed world assumptions while bringing additional tools to bear on the exercise of communicating human knowledge. From simple descriptive metadata to advanced rules and formulae permitting higher-level reasoning, the Semantic Web is providing the additional infrastructure to express both the context and structure of information as well as relationships among information in a form that supports machine processing.

Making available open, domain and cross-domain taxonomies and ontologies useful for classifying and organizing content is important. Providing persistent URIs for these concepts and making these available via common Web architecture means provides an important backbone for the enabling the Semantic Web.

### **Creating a Policy Aware Infrastructure**

The computational machinery behind the Web is today somewhat like a small child in a shopping mall; it has no mechanism for distinguishing what sources of information to trust, no mechanism for identifying what is considered socially acceptable information to disclose, and only limited mechanism to act as a reliable transporter of information between third parties who have a more highly developed awareness of information provenance, reliability, and decay. The absence of these mechanisms will become a severe limitation on the long-term utility, and perhaps even viability, of Web search engines, including those that perform co-citation analyses to infer nodes of expertise. The introduction of trust awareness and social policy description mechanisms will provide the foundation for a set of services and applications we can only begin to imagine.

The development of a *Policy Aware Infrastructure* for the Web is required. The Semantic Web will only achieve its potential as an information space for the free flow of scientific and cultural information if its infrastructure supports a full range of fine-grained policy controls over the information contained in the Semantic Web. If we are going to entrust more of our knowledge to the Semantic Web, we must be assured that the Web will respect many more of the social agreements that we enforce in the physical world. For the

Semantic Web includes not only freely available information, but also personal information and information available to a person or agent only as a result of its membership in groups. A policy-aware infrastructure -- one that gives information creators and users the types of control over information we have all become accustomed to in the physical world such as the ability to assert and exercise privacy and intellectual property rights -- will make the Semantic Web into a vibrant and humane environment for sharing knowledge and collaborating on wide range of intellectual enterprises.

## **Evolution and Translation of Vocabularies**

An important goal of the Semantic Web is to address the problem that in the course of scientific (or any) endeavor, one changes the vocabularies one uses to organize, discover, and communicate. A given vocabulary may be refined, resulting in a need for migration from old to new. Communication between distinct groups using different vocabularies creates the need to create common vocabularies which optimally suit all involved. Semantic Web techniques should make this difficult process of creating new common vocabularies as easy as possible. The Semantic Web already removes confusion by giving each term a globally unique URI. OWL ontologies and rules languages allow relationships between old and new terms to be expressed. There is, however, little experience with the serious management of such evolution.

The Semantic Web needs to incorporate versioning and provenance within its foundation. Human understanding changes and statements that we once thought were accurate are later described to be inaccurate. However, the original statement should not be deleted from our corpus of human knowledge. The Semantic Web should not be required to forget that a statement was once believed to be a true statement. Versioning is such a common approach to representing discrete states of understanding that it warrants explicit treatment in the Semantic Web.

## **Web of Trust**

Trust in the human social context is based on constantly evolving and adapting information. Two parties may trust each other based on a history of mutual interaction, based on formal contracts that in turn rely on other established systems (e.g. legal and legislative), and based on risk analysis of a failure of any party to perform as agreed. A trust language for the Semantic Web that is capable of representing these complex and evolving relationships will be crucial to our future ability to build software that behaves more in the manner of an intelligent assistant than a rote rules processor.

Data dissemination and usage rights description languages are especially important if intellectual property of any value is to be placed in the Web. Rights description must be aligned with legislative decision-making yet must reflect differences between autonomous communities. Similarly, data dissemination and reuse agreements must be able to express the detailed nuances of acceptable practice that social communities adopt and evolve.

Data access rules should be able to express trust based on peer relationships without having to enumerate peers. The objective is a Web in which anyone can publish

information and assert distribution policies over each object at various levels of granularity. The distribution policies can be permitted to change over time, making more of the data available to others if desired. The Web should recognize and allow the expression of privacy concerns and ad-hoc collaboration arrangements. The goal is that more content is recorded once people have confidence that the system will protect their property rights.

While the early policy-capable tools that exist on the Web today were designed primarily for the HTTP/HTML interactions between individual users and Web servers (often in a commercial context), policy awareness on the Semantic Web raises a new set of challenges. These challenges include identifying the privacy interests of those who contribute intellectual content (as opposed to just browsing it), developing a model to represent access control needs of heterogeneous communities of research collaborators, and modeling the special needs of the digital repositories being deployed by academic libraries.

### **Information Flow, Synchronization, and Collaborative Life**

Many tools used with collaborating groups today instrument the flow of data, information, and knowledge. For instance; data in a typical software development environment has dependencies recorded in Makefiles, while an annotated history of the development of code, data, and documentation files is kept in source code control systems. Collaboration involves a series of meetings which have related documents, agendas and minutes. Not only is scientific data important, but the record ideally includes the steps, social and automatic, by which the associated information evolved. These steps may include for example sign-off, delegation, consensus decision and peer review. Semantic Web tools will permit more such information to be captured and will simplify the combination of information for different sources. The Semantic Web will leverage such information, enabling the extraction of existing relationships and the discovery of new relationships which, when available in common Semantic Web form, will allow more powerful analysis of the past, and will support the development of more powerful tools in the future.

One of the challenges we will meet is to strike a balance between requiring authors to do more at the outset to make information machine processable, insisting that everything the machine could use to answer a question be recognized and identified by the (human) questioner, and leaving large quantities of information inaccessible to the machine.

## **Conclusion**

A global ubiquitous information infrastructure is the goal of the Semantic Web. The Semantic Web is the natural evolution of Web technology. It brings richer, more descriptive data to the current Web, which can be used more effectively by software and applications - making people more productive. The successful realization of such an infrastructure will be supported by a science of information management that will yield new generations of knowledge environments. Digital Libraries are an important component for enabling such an infrastructure. The identification of a Digital Library Research Agenda whose goal is to facilitate enabling this infrastructure is important step.

I look forward to working with the workshop members in helping establish such an agenda.

## References

Atkins, Daniel E., et. al., "Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue Ribbon Advisory Panel on Cyberinfrastructure," January 2003, Available at [http://www.communitytechnology.org/nsf\\_ci\\_report/](http://www.communitytechnology.org/nsf_ci_report/).

[HTTP] "RFC 2616: Hypertext Transfer Protocol -- HTTP/1.1", J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, T. Berners-Lee, June 1999. Available at <http://www.ietf.org/rfc/rfc2616.txt>.

[URI] "RFC 2396: Uniform Resource Identifiers (URI): Generic Syntax", T. Berners-Lee, R. Fielding, L. Masinter, August 1998. Available at <http://www.ietf.org/rfc/rfc2396.txt>.

[XML] "Extensible Markup Language (XML) 1.0 (Second Edition)", T. Bray, J. Paoli, C.M. Sperberg-McQueen, E. Maler, 6 October 2000. This W3C Recommendation is available at <http://www.w3.org/TR/2000/REC-xml-20001006>.

[ACKERMAN] Ackerman, M., Darrell, T., & Weitzner, D. J. (2001). [Privacy in context](#). Human-Computer Interaction, 16, pp. 167-176. (privacy design considerations in pervasive computing environments)

[WEITZNER] Weitzner, D (2000) [United States Senate Commerce Committee Hearing on Online Privacy](#) (on P3P)

[BERMAN] Berman, J. & Weitzner, D., Abundance and User Control: Renewing the Democratic Heart of the First Amendment in the Age of Interactive Media, [104 Yale L.J. 1619 \(1995\)](#) (introducing the idea of giving users control over policy-related aspects of the user experience)