

Metadata: A Fundamental Component of the Semantic Web

by Jane Greenberg, Stuart Sutton and D. Grant Campbell

Jane Greenberg, assistant professor, School of Information and Library Science, University of North Carolina at Chapel Hill, can be reached by e-mail at janeg@ils.unc.edu. Stuart Sutton, associate professor, The Information School, University of Washington, can be reached at sasutton@u.washington.edu. D. Grant Campbell, assistant professor, Faculty of Information and Media Studies, University of Western Ontario, is at gcampbel@uwo.ca.

In their widely discussed May 2001 article on the Semantic Web in *Scientific American* (www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21) Tim Berners-Lee, James Hendler and Ora Lassila present a scenario in which a person named Pete is listening to the Beatles through his home entertainment center. Lucy, Pete's sister, phones from the doctor's office to explain that their mother needs a series of bi-weekly physical therapy sessions. The first few paragraphs of this article tell how both Pete's and Lucy's Semantic Web agents (hereafter referred to as *agents*) communicate with each other and transverse the Semantic Web to schedule their mother's physical therapy session, how Pete is not pleased with the initial plan and how later that evening Pete sends his agent back onto the Semantic Web to find an alternative plan. Pete's Web agent completes this second task and reschedules several of his personal and less important work appointments.

Realizing this scenario is dependent not only on the ability of Pete's and Lucy's agents to communicate with each other, but their ability to transverse a system of structured semantic knowledge that is forming the Semantic Web. This system of semantics is *metadata*. With efforts to build the Semantic Web, we are beginning to see the metadata infrastructure that agents need to carry out tasks that the initial Web has not been able to support. To this end, implementing and harvesting metadata is fundamental to the success of the Semantic Web. This article provides an overview of metadata as a key component of the Semantic Web – its vision and architecture; metadata vocabularies; enabling technologies; and authoring and annotation.

Semantic Web: Vision & Architecture

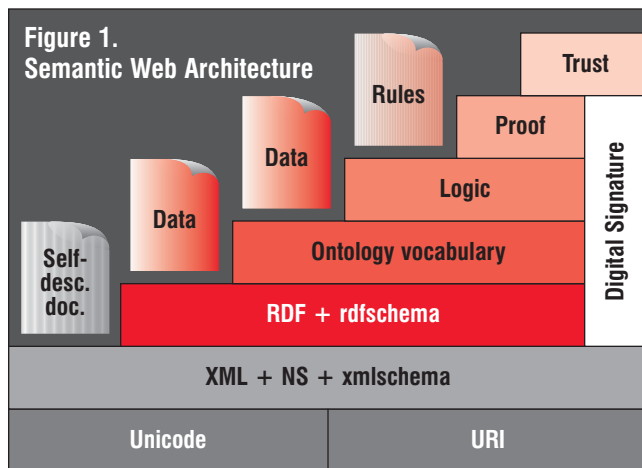
Clearly articulated visions and architectural plans, drawn by great thinkers and experts of the time, form the underpinnings of many of the world's most significant structures. Consider the

Panama Canal. Crossing the mass of land called the Americas at its narrowest point by means of a waterway was a vision shared by many throughout history, including indigenous people of the Americas, Christopher Columbus and merchants worldwide. Execution of architectural plans, first by leading French engineers in 1879 and then by a U.S. Commission, led to completion of the Panama Canal in 1914. In like fashion, an evolving and shared vision, supported by an architectural plan, underlies the development of the Semantic Web.

The Vision. The Semantic Web was envisioned by Tim Berners-Lee, inventor of the World Wide Web (Web), and is now being further defined by researchers and visionaries, but it was inspired by a host of creative thinkers who have, throughout history, looked to technological innovation as a way not only to control, but also to transform the world's mass of information into intelligence. Vannevar Bush was one early pioneer in this area with his vision of the Memex (www.theatlantic.com/unbound/flashbks/computer/bushf.htm) – a mechanism using associative indexing to link the world's vast body of scientific recorded knowledge to discover new knowledge. Another significant idea is Alan Turing's conceptualization of the Turing Machine and its use of logic to transform numbers into intelligence. (See www.turing.org.uk/turing/scrapbook/machine.html.) The vision supporting the Semantic Web draws upon these ideas and new ideas inspired by technological developments to create intelligence.

The Architecture. The Semantic Web's architecture, captured by Berners-Lee, is represented in Figure 1. Each layer supports or has a connection to metadata.

- **URIs and Unicode.** URIs (uniform resource identifiers) are unique identifiers for *resources* of all types—from schemas to people. A major



(Berners-Lee, T., 2000) www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html

component of the base layer, URIs are metadata and function like ISBNs (international standard book numbers) or Social Security numbers in the context of the Web.

- **XML + NS + XMLschema.** Extensible Markup Language (XML) and more recently XML schemas facilitate the creation, use, and syntactic interoperability of metadata vocabularies. NS (namespaces), which are identified via URIs, secure semantic interoperability among metadata vocabularies.
- **RDF and RDFschema.** The RDF family further supports interoperability at the semantic level. RDF developments comprise the base Web language, so that agents, like Pete's and Lucy's discussed above, can make logical inferences, based on metadata, to perform tasks.
- **Ontology vocabulary.** Ontologies are metadata systems (referred to as metadata vocabularies in this article). The ontology layer represents the Semantic Web's central metadata artery, where simple descriptive to complex classificatory schemas are to be created and registered so that agents can intelligently interpret data, make inferences, and perform tasks. Jacob's article in this issue discusses ontologies in detail.
- **Logic.** We make logical inferences in our performance of daily tasks. For example: If *N* denotes new unread email in an email inbox, then if an *N* appears by a particular message, the message is new unread email. This inference is based on evidence provided by the letter *N*. The logic layer of the Semantic Web works on this basic principle through First Order Predicate Logic. An agent can derive a logical conclusion (or reason) in the process of completing a task based on what are essentially "facts" rendered from semantically encoded metadata. Other types of logic may also be applicable in the Semantic Web.
- **Proof and Trust.** The last two horizontal layers build off of the logic layer, *proof* being a validation of the "evidence" stemming from the inferential logic activity and

trust relating to the integrity of the proof, which may be traced back down through the other layers in Berners-Lee's diagram. The functionality of these two layers is highly dependent on creation of accurate and trustworthy metadata.

- **Digital signature.** Digital signatures run horizontal to the RDF family up through the proof layer and support the notion of trust. Developments in the area of digital signatures are progressing, and could eventually help validate the integrity of metadata that an agent will use for reasoning and task completion.

The synopsis provided here shows that metadata permeates each layer of the Semantic Web's architecture, although it is not the only piece, as agents, enabling technologies, authoring and annotation programs, and metadata vocabularies need to be further developed to realize the full potential of the Semantic Web.

Metadata Vocabularies

Metadata vocabularies are synonymous with ontologies, as discussed above. A vocabulary, in a general sense, is a shared system of semantics for concepts or objects. Think about the vocabulary that comprises Portuguese. It is a system of agreed upon meanings permitting intelligible communication among Portuguese people and other persons who speak this language.

The metadata world hosts a range of systems to which the label of "metadata vocabulary" is applied. These vocabularies range from basic descriptive metadata systems, with limited or single-focused functionalities, to more complex "member/class" semantic vocabularies. An example of a simple ontology is the Dublin Core Metadata Initiative (DCMI) Elements and Element Refinements (www.dublincore.org/usage/terms/dc/current-elements/), a metadata schema developed mainly to facilitate resource discovery. A more complex metadata vocabulary to which the word *ontology* is applied is the Ariadne Genomics ontology (www.ariadnegenomics.com/technology/ontology.html). This ontology is used to formalize data on cell proteins for computer analysis: the ontology defines the various proteins, classifies them into taxonomic trees and defines the semantic relationships among them. While many metadata vocabularies in operation were not necessarily created with the Semantic Web in mind, they may be able to play a significant role in its development. For existing and developing ontologies to be used and function fully in the Semantic Web environment, they need to adhere to standards supported by enabling technologies.

Enabling Technologies for Metadata

Although metadata is integral to the Semantic Web, metadata on its own is far from sufficient. Needed are standards to syntactically encode and represent semantic knowledge so that agents can perform tasks in an efficient and comprehensible manner. A variety of enabling technologies have developed over the last few years that are proving significant to the

construction of the Semantic Web. Some have developed prior to the conceptualization of the Semantic Web, while others are being developed and refined specifically with the Semantic Web in mind. Among four key developments that are critical for metadata encoding and manipulation by agents are:

- XML and XML schema
- RDF and RDFschema
- DAML+OIL (DARPA Agent Metadata Language/Ontology Inference Layer) (www.w3.org/TR/daml+oil-reference)
- OWL (Web Ontology Language) (www.w3.org/2001/sw/WebOnt).

The other articles in this special section of the *Bulletin of the American Society for Information Science and Technology* give attention to each of these and other important technologies.

Semantic Web Authoring and Annotation

As we move on with our discussion of the Semantic Web, we need to account for the notion that all of these impressive developments assume that metadata will exist – that authors and third parties will take the time and trouble to create metadata for Web content which can be read, understood and harvested by intelligent agents. This is a daunting proposition since accurate, consistent metadata is notoriously difficult to create, as librarians and information scientists over the world will verify. The challenge, then, is to decentralize a task that has traditionally been centralized in libraries and other information centers, carried out by professionally trained catalogers. The glories of the Semantic Web will ultimately depend on tools that will enable authors to create with very little effort RDF annotations and other useful semantic metadata on their Web pages.

Annotation has been one of the slower developments on the World Wide Web: Berners-Lee's vision of a Web that permits collaborative authoring, in addition to hyperlinked pages, has not yet materialized. At last, however, an encouraging number of annotation tools are appearing, ranging from simple captioning systems to ambitious and sophisticated systems that provide multiple views of annotation in multiple formats. At the most basic level, programs such as Annotea enable multiple users to comment on and provide metadata for a single pool of documents for purposes of collaborative writing and research. Other programs more directly geared to the Semantic Web provide ways in which RDF and ontological data may be easily created and stored in the headers of the document.

Painless creation of RDF metadata depends on two things: a predefined ontology that spares the author the task of creating terms and relationships and a user-friendly interface that permits the author to create metadata instances intuitively. One such annotation program, OntoMat, permits the author to download a predefined ontology that appears in one window, while the HTML document being annotated appears in the other. The author highlights elements of the document to be annotated and places them into a third window, and then uses the ontology to

define the data elements and their relationships to each other. The program then generates highly detailed RDF metadata without the author having seen a single angle bracket. With tools such as these, conference organizers, for instance, can require contributing authors to annotate their abstracts with RDF metadata or members of a particular community can annotate their Web pages using a common ontology. Equally important, the tasks can be integrated fairly painlessly into the normal workflow of Web authoring.

A variety of annotation tools are listed at the Semantic Web Authoring and Annotation site: <http://annotation.semanticweb.org/tools>. They include

- OntoMat – <http://annotation.semanticweb.org/tools/ontomat>
- Annotea – www.w3.org/2001/Annotea/
- Annozilla – <http://annozilla.mozdev.org/>
- COHSE – <http://cohse.semanticweb.org/software.html>
- SMORE – www.mindswap.org/~aditkal/editor.shtml

These tools, and the tools which improve on them, will hopefully provide the simplicity that is essential for the Semantic Web to grow. The Semantic Web, after all, was not envisioned as a tool for information professionals and computer scientists, but as a tool for everyone. And if Pete and Lucy are ever going to prefer this new Web to their date books and palm pilots, it has to stay simple.

Conclusion

Our perceptions of metadata's role in both the vision and architecture of the Semantic Web are not yet fully focused. While the vision embraces metadata as a first-order prerequisite to that architecture, its roles and mechanisms currently resonate with the evolution of earlier technological achievements. At the turn of the last century, as ships began to pass through the Panama Canal, the fledgling horseless carriage traversed other byways in forms we would hardly recognize today. They asked then, Shall we steer with a wheel or with a stick? Will the brake be on the left or the right of the steering column? Competing variations on the vision's theme sought to dominate the architecture of the evolving automobile. In many ways, we stand in a place quite like that occupied by those earlier pioneers when we view the potential roles and forms of metadata in the emerging architecture of the Semantic Web.

Further Reading

- Berners-Lee, T. (1997). Axioms of Web architecture: Metadata: Metadata architecture. In *Design issues: Architectural and philosophical points (Semantic Web roadmap)*. Available at www.w3.org/DesignIssues/Metadata.html
- World Wide Web Consortium. *Metadata activity statement*. Available at www.w3.org/Metadata/Activity.html
- Semantic Web authoring and annotation*. Available at <http://annotation.semanticweb.org/>