Semantic Web technologies for digital libraries

York Sure and Rudi Studer
Institute AIFB, University of Karlsruhe, Karlsruhe, Germany

Abstract

Purpose – The purpose of this article is to provide an overview about the Semantic Web, its importance and history and an overview of recent Semantic Web technologies which can be used to enhance digital libraries.

Design/methodology/approach – The paper answers, at least partially, questions like "What is the Semantic Web?", "How could the Semantic Web look like?", "Why is the Semantic Web important?", "What are ontologies?" and "Where are we now?". Several pointers to further literature and web sites complete the overview.

Findings – Semantic Web technologies are valuable add-ons for digital libraries. There already exist numerous academic and commercial tools which can be applied right now.

Practical limitations/implications – The overview of Semantic Web technologies cannot be complete in such an article, therefore we limit ourselves to the most prominent technologies available. However, following the pointers given readers can easily find more information.

Originality/value – The article is of particular value for newcomers in this area.

Keywords Internet, Digital libraries, Generation and dissemination of information

Paper type General review

What is the Semantic Web?

Berners-Lee *et al.* (2001) describe the Semantic Web as: "…an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation". The key enabler of the Semantic Web is the need of many communities to put machine-understandable data on the web which can be shared and processed by automated tools as well as by people. Machines should not just be able to display data, but rather be able to use it for automation, integration and reuse across various applications.

The European Commission is funding numerous projects related to ontologies and the Semantic Web in its currently running Sixth Framework Research Programme, e.g. "Semantically Enabled Knowledge Technologies" (the SEKT project[1]). The worldwide Semantic Web community is growing rather fast and forces are being joined with other technology developments such as web services or multimedia. Last, but not least, vendors are already offering mature products and solutions based on semantic technologies. Thus, the Semantic Web is currently moving from being a vision to becoming reality.

How could the Semantic Web look like?

Even worse: "How would you explain the Semantic Web to your grandparents?" Answering this question is one of the challenges for participants of the Semantic Web Challenge[2]. It might be questionable whether grandparents would understand the aim of, for example, the winning application in 2003, namely that it "combines information from multiple heterogeneous sources, such as published RDF sources,

Library Management
Vol. 26 No. 4/5, 2005
pp. 190-195
© Emerald Group Publishing Limited
0143-5124
DOI 10.1108/01435120510596044

personal web pages, and data bases in order to provide an integrated view of this multidimensional space"[3]. Nevertheless, it offers the flavour of current Semantic Web technologies.

A very illustrative and at the same time amusing article gives a glimpse into the far future, namely: "August 2009: how Google beat Amazon and Ebay to the Semantic Web" (Ford, 2002).

Why is the Semantic Web important?

To illustrate the potential importance of the Semantic Web we will start with some quotes showing the relevance and awareness the Semantic Web already has at non-academic key players.

> The way software and devices communicate today is doomed. To interoperate on the X Internet, they'll use decentralized data dictionaries based on emerging Semantic Web technologies (Truog, 2001).

> While the industry is busy creating the underpinnings of open computing with standards like eXtensible Markup Language, still missing are what Plattner calls "semantic" standards, or how to make different computers recognize data about a business partner, a customer, or an order and know what to do with it. In other words, said Plattner, the software industry is building an alphabet but hasn't yet invented a common language (Hasso Plattner, SAP, in CNet News, 2002).

Little history of the Semantic Web

The advent of the world wide web (WWW) gave mankind an enormous pool of available information. The WWW is based on a set of established standards, which guarantee interoperability at various levels:, e.g. the TCP/IP protocol provides a basis for transportation of bits, on top HTTP and HTML provide a standard way of retrieving and presenting hyperlinked text documents. Applications could easily make use of this basic infrastructure, which led to the now existing WWW. However, nowadays the sheer mass of available documents and the insufficient representation of knowledge contained in documents make "finding the right things" real work for human beings. A major shortcoming of HTML is that it is well suited for human consumption, but not for machine-processability. As such, to interpret the information given in documents the human has always to be in the loop.

To overcome such shortcomings, ontologies recently have become a topic of interest in computer science. Ontologies provide a shared understanding of a domain of interest to support communication among human and computer agents, typically being represented in a machine-processable representation language. Thus, ontologies are seen as key enablers for the Semantic Web.

What are ontologies?

There are different definitions in the literature of what an ontology should be, the most prominent being published by Gruber (1995):

> An ontology is an explicit specification of a conceptualization. The term is borrowed from philosophy, where an Ontology is a systematic account of Existence. For AI systems, what "exists" is that which can be represented.

A conceptualization refers to an abstract model of some phenomenon in the world by identifying the relevant concept of that phenomenon. Explicit means that the types of concepts used and the constraints on their use are explicitly defined.

This definition is often extended by three additional conditions: “An ontology is an explicit, formal specification of a shared conceptualization of a domain of interest”. Formal refers to the fact that the ontology should be machine readable (which excludes for instance natural language). Shared reflects the notion that an ontology captures consensual knowledge, that is, it is not private to some individual, but accepted as a group. The reference to a domain of interest indicates that for domain ontologies one is not interested in modelling the whole world, but rather in modelling just the parts which are relevant to the task at hand.

In a nutshell, ontologies help to represent knowledge in a machine processable way; they express a shared view on a domain of interest.

How can Semantic Web technologies help digital libraries?

Digital libraries offer access to large amounts of content in form of digital documents. Many of them have evolved from traditional libraries and concentrated on making their information sources available to a wider audience, e.g. by scanning journals and books, thereby only taking limited advantage of the benefits modern computing technologies offer. To overcome this bottleneck, research and development for digital libraries include processing, dissemination, storage, search and analysis of all types of digital information.

Semantic technologies allow for the description of objects and repositories, i.e. the need to establish common schemes in form of ontologies, e.g. for the naming of digital objects. A main goal is to enable interoperability, i.e. the ability to access, consistently and coherently, similar classes of digital objects and services, distributed across heterogeneous repositories.

Typical usage scenarios for Semantic technologies in digital libraries include among others user interfaces and human-computer interaction (displaying information, allowing for visualization and navigation of large information collections), user profiling (taking into account the overall information space), personalization (balancing between individual and community-based personalization), and user interaction.

These and other challenges are addressed by the SEKT project guided by the vision that in future, while there would still be many digital repositories, a digital library system should provide a consistent view of as many repositories as possible. From a user’s perspective, they should appear to be a single digital library system. Even more, a digital library system needs to extend smoothly from personal information sources, workgroup and corporate systems, out to personal views of the content of more public digital libraries.

Further information about this SEKT case study can be found in the following article in this issue, “Applying semantic technology to a digital library: a case study”.

Which are prominent Semantic Web technologies?

Ontobroker (initially developed at the Institute AIFB/University of Karlsruhe, and now commercialized by the company Ontoprise) and SHOE (University of Maryland) were two ontology-based systems ahead of their time. Both systems relied on additional semantic markup which was put into regular web pages, so-called annotations. The systems showed very early the feasibility of adding machine-processable semantics to web pages. Many ideas of this work made it into current Semantic Web standards of the W3C (see the later section on Standards).

Both systems also heavily influenced the development trends of semantic technologies. In the following we will briefly characterize typical Semantic Web tools and give examples of existing commercial and academic tools. It is quite noteworthy that most tools are currently not only being used to build and maintain WWW applications, but also corporate intranet solutions.

Ontology editors

Ontology editors allow for creation and maintenance of ontologies, typically in a graphically oriented manner. There exists a plethora of available implementations, each having its own specialty and different functionalities. Common to most editors is the ability to create a hierarchy of concepts (such as "Car is a subconcept of Motor Vehicle") and to model relationships between those concepts (such as "A car is driven by a person". More advanced editors also allow the modelling of rules, but to explain this is beyond the scope of this paper.

OntoEdit is the most prominent commercial ontology editor (available at: www.ontoprise.com). Unlike most other editors, OntoEdit comes with a strong inferencing backbone, Ontobroker, which allows the modelling and use of powerful rules for applications. Numerous extensions, so-called plug-ins, exist to adapt OntoEdit flexibly to different usage scenarios such as database mapping. Last, but not least, a full-fledged tool support is provided by the Ontoprise team which makes it attractive for companies.

Protégé is the most well-known academic ontology editor with a long development history (available at: http://protege.stanford.edu/). Similar to OntoEdit it is based on a flexible plug-in framework. Numerous plug-ins have been provided so far which nicely demonstrate possible extensions for typical ontology editors. An example is the PROMPT plug-in, which allows for merging of two given ontologies into a single one.

KAON (http://kaon.semanticweb.org) is not only an ontology editor, but rather an open-source ontology management infrastructure targeted at business applications. It includes a comprehensive tool suite allowing easy ontology creation and management, as well as building ontology-based applications. An important focus of KAON is on integrating traditional technologies for ontology management and application with those used in business applications, such as relational databases.

Annotation tools

Annotation tools (see also Handschuh and Staab, 2003) allow for adding semantic markup to documents or, more generally, to resources. The great challenge here to automate the annotation task as much as possible to reduce the burden of manual annotation for large-scale resources. A good place to find further information on annotation and authoring, a quite related topic, is http://annotation.semanticweb.org/.

Annotea (http://www.w3.org/2001/Annotea/) is a LEAD (Live Early Adoption and Demonstration) project enhancing the W3C collaboration environment with shared annotations. By annotations we mean comments, notes, explanations, or other types of external remarks that can be attached to any web document or a selected part of the document without actually needing to touch the document. When the user gets the document he or she can also load the annotations attached to it from a selected annotation server or several servers and see what his peer group thinks.

OntoMat-Annotizer (http://annotation.semanticweb.org/ontomat) is currently the most prominent annotation tool. It is based on a full-fledged annotation framework called CREAM, which is already being extended to support semi-automatic annotations of documents as well as annotation of databases.

KIM (http://www.ontotext.com/kim) provides a knowledge and information management (KIM) infrastructure and services for automatic semantic annotation, indexing, and retrieval of unstructured and semi-structured content.

Inference engines

Inference engines allow for the processing of knowledge available in the Semantic Web. In a nutshell, inference engines deduce new knowledge from already specified knowledge. Two different approaches are applicable here: having general logic based inference engines, and specialized algorithms (problem-solving methods). Using the first approach one can distinguish between different kinds of representation languages such as higher order logic, full first order logic, description logic, datalog and logic programming (see also http://semanticweb.org/inference.html). Recently, in a contest-like project three state-of-the-art inference engines were evaluated with quite interesting results (see also http://www.projecthalo.com/). Inference engines are *per se* very flexible and adaptable to different usage scenarios such as information integration or, to show the bandwidth of possible scenarios, intelligent advisors.

Ontobroker (http://www.ontoprise.com) is the most prominent and capable commercial inference engine. It is based on frame logic, tightly integrated with the ontology engineering environment OntoEdit and provides connectors to typical databases. It was already used in numerous industrial and academic projects.

FaCT (http://www.cs.man.ac.uk/ ~ horrocks/FaCT/)is one of the most prominent Description Logics based inference engines. In a nutshell, FaCT (fast classification of terminologies) is a description logic classifier that can also be used for modal logic satisfiability testing. It is based on the tableaux calculus.

KAON2 (http://kaon2.semanticweb.org/) is a new description logics based inference engine for OWL-DL and OWL-Lite reasoning. Reasoning is implemented with novel algorithms which reduce a SHIQ(D) knowledge base to a disjunctive datalog program.

Where are we now?

Standards activities for Semantic Web languages are mainly driven by working groups of the W3C (http://www.w3c.org/). The Semantic Web layer cake (see Figure 1) by Tim Berners-Lee shows the layering of the current state-of-the-art and future planned standards. On the right side can be seen the current status of each layer. While XML as a baseline allows for a syntactical description of documents, the layers RDF, Ontology and Logic are adding machine-processable semantics – a necessary prerequisite for, for example, shareable web resources.

On top of the core standards for XML (eXtensible Markup Language) and RDF (Resource Description Framework) the W3C WebOnt working group (http://www.w3.org/2001/sw/WebOnt) released early 2004 the OWL web ontology language standard

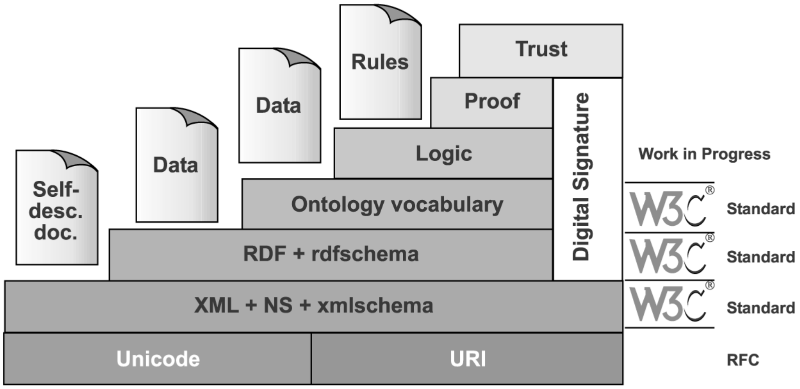

Figure 1.
The Semantic Web layer cake

(http://www.w3.org/TR/owl-ref). Future work remains to be done for the logic, proof and trust layers.

The community is growing quickly, also attracting researchers and practitioners from other areas such as information systems (e.g. the AIS SIG on Semantic Web and information systems (see http://www.sigsemis.org/). The upcoming conferences ESWC2005 (2nd European Semantic Web Conference (see http://www.eswc2005.org)) and ISWC2005 (4th International Semantic Web Conference (see http://iswc2005.semanticweb.org)) will be good places to find latest research results and industrial applications of Semantic Web technologies.

Notes

1. EU IST SEKT project, see www.sekt-project.com
2. Semantic Web Challenge, initiated in cooperation with the International Semantic Web Conference in 2003, continued in 2004, see also http://challenge.semanticweb.org/
3. CS AKTiveSpace Tour, see also http://triplestore.aktors.org/SemanticWebChallenge/

References

Berners-Lee, T., Hendler, J. and Lassila, O. (2001), "The Semantic Web", *Scientific American*, May, available at www.sciam.com/2001/0501issue/0501berners-lee.html

CNet News (2002), "CNet News SAP calls for software that works together", *CNet News.com*, 27 March.

Ford, P. (2002), "August 2009: how Google beat Amazon and Ebay to the Semantic Web", *Ftrain.com*, 26 July, available at: www.ftrain.com/google_takes_all.html

Gruber, T.R. (1995), "Towards principles for the design of ontologies used for knowledge sharing", *International Journal of Human-Computer Studies*, Vol. 43 Nos. 5/6, pp. 907-28.

Handschuh, S. and Staab, S. (Eds) (2003), "Annotation for the Semantic Web", *Frontiers in Artificial Intelligence and Applications*, Vol. Vol. 96,.

Truog, D. (2001), "How the X Internet will communicate", *Forrester Report*, December.