

Semantic Modeling

Jeroen Hoppenbrouwers

March 21, 2003

Semantics, the science of describing meaning in some machine-readable form, has been a popular research topic for decades. Despite some promising results, thinking machines or real artificial intelligence did not see the light yet. But by making use of typical features of today's World-Wide Web infrastructure, which was not designed to carry meaning at all, we might be able to create a knowledge platform which offers much more potential to change the information world than all previous attempts (Berners-Lee, 2001). The principle is comparable to that which made Google¹ such a powerful service. It is neither the engine that makes the difference, nor the document collection. It is the Web.

1 The Semantic Web

After the explosive growth in popularity of the World-Wide Web (www) in the early 1990s, researchers quickly started drawing up ideas for enhancements to the www. They started from a few key points that made the www so much bigger and more popular than any of its predecessors (Berners-Lee 2001):

1. Universality, "anything can link to anything". There is no discrimination between the scribbled draft and the polished performance, between commercial and academic information, or among cultures, languages, media and so on.
2. Decentralisation. This enables unchecked exponential growth of the network, at the expense of throwing away the ideal of total consistency of all of the interconnections. Web links are unidirectional, not bidirectional; there is no mechanism to weed out redundancy, or assure mutual updates.
3. Versatility. Paradoxes and unanswerable questions are a price that must be paid to achieve versatility. The Web is not a well-organised library, it has no tree structure, one is never sure of finding everything. But due to this lack of top-down organisation, it is remarkably effective in providing us with unprecedented amounts of raw information.

¹<http://www.google.com/>

With these three basic principles firmly in their mind, researchers looked at the scientific field of Knowledge Representation. The current state of the art in Knowledge Representation is comparable to that of Hypertext before the Web: it is clearly a good idea, and some very nice demonstrations exist, but it has not yet changed the world. By adapting the existing systems to be less strict in their world view, i.e., by dropping the central definitions, logical provability, decidability, etc., that have been the main goals of research over the last decades, they expect to gain so much more universality and versatility that it is worth the sacrifice.

Knowledge representation research long focussed on what can now be called heavyweight knowledge: mathematically sound, formal descriptions of (parts of) the world, close to formal logic and computer programming.² Although very nice to have, acquiring this type of knowledge about the world is very hard since you need formal analysis and experienced programmers to write it down. Traditional formal knowledge representation systems therefore were small and limited in scope. Few projects tried to scale up, with the CYC project on top.³

From several corners of the scientific world, lightweight knowledge representation systems have been appearing that are much easier on the acquisition work. A good example is a classic library thesaurus. With only a handful of clear rules, people have been building impressive knowledge structures that have proved their usefulness for decades. The WordNet community also works with lightweight structures, though richer in expressiveness and therefore more difficult to compile than a plain thesaurus.⁴

One of the ideas behind the Semantic Web is that the latter type of lightweight knowledge, although it is not fully descriptive, is so much easier to work with that it is preferred over the traditional heavyweight approaches.

2 How to Enable the Semantic Web

The two basic knowledge items of the Semantic Web are data and rules (Berners-Lee 2001). *Data* can be compared to terms or subject headings: unique descriptions in some language that mean something to a group of people. They are not formally defined, just presented. And they can be uniquely referred to, so that several objects can share the term. *Rules* link up the terms, by connecting two terms with a descriptive relationship. If you think you now see a classic thesaurus, you are nearly right. The relationships may form a mesh, not necessarily a tree, and the relationship types are not restricted to the limited standard thesaurus set. Just as anybody may add a new term, anybody may also add a new relationship type.

In the Semantic Web, terms are attached to (WWW) documents by using Extensible Markup Language (XML) tags, so that machines can find out extra information about a document.⁵ Terms are linked up by the Resource

²<http://www.semanticweb.org/inference.html>

³<http://www.cyc.com>

⁴[http://www.cogsci.princeton.edu/~ wn/](http://www.cogsci.princeton.edu/~wn/)

⁵<http://www.w3.org/XML/>

Description Framework (RDF),⁶ also expressed in XML. Instead of a central term and rule database, many individual Web pages exist with terms and rules, and these can be referred to by normal Web URLs.

This Semantic Web structure therefore requires three resources:

1. Lists of terms to tag documents with. Each term should have its own Web page, where further information is available so that humans can actually try to understand the meaning of the term. For machines, the information is less useful. They need:
2. Lists of links between terms, also available on Web pages. These links form a structure of terms, a conceptual space, which can be used to look up the meaning of terms expressed by other terms.
3. Formal, coherent collections of terms linked up to form conceptual structures called ontologies (Guarino 1998). These ontologies form the knowledge backbone of the Semantic Web, as they organise terms in classes, subclasses, and relationships between them. With these formal relationships, machines can make inferences which allow certain forms of knowledge deduction.⁷

The millions of Web pages (documents) out there should obviously be tagged with appropriate terms. This is a gigantic job, which hopefully will be undertaken by the individual authors of the pages, and brings back memories of efforts to have authors tag their own library index cards. But these pages do not form the Semantic Web. The Semantic Web is a separate structure that contains references to the World-Wide Web, while also being part of it.

3 How to Manage the Semantic Web

As you will have noticed, many of the structures that make up the Semantic Web are not entirely unknown to librarians. Collections of references, taxonomies of terms, and documents marked up with keywords all sound very familiar. This is no coincidence; ages of librarianship have produced some practical ways to organise growing collections of material and there is no reason to assume that these practices are invalid in the electronic information age.

A key difference between library practice and the Semantic Web, however, is the way in which these knowledge resources are managed. Library knowledge resources traditionally have seen a rigid, central management system, led by highly trained and skilled people in order to maintain the required systematic organisation. With a small hiccup in the systematic approach, whole knowledge areas ran the risk of becoming inaccessible, since the single pathway into the area could get obscured. And still, intimate knowledge of the reasoning behind the thesauri and subject heading systems was the key to a truly efficient and effective use of these resources. In other words, an experienced librarian performed better in searching material than an average library user.

⁶<http://www.w3.org/RDF/>

⁷<http://www.semanticweb.org/knowmarkup.html>

The Semantic Web abandons much of the central rigidity in favour of a less formal, 'wilder' approach, for the reasons mentioned in the first section of this paper. Nonetheless, existing resource management practices will still be appropriate, especially for the ontologies. The fact that the Semantic Web can be maintained by unorganised volunteers does not mean that there is no place for professionals! On the contrary, the better the resource, the more likely that Semantic Web authors/annotators will use it, so there is a compelling reason to provide high-quality semantic resources for free on the Web – a task uniquely suited to libraries.

In a way closely related to the MACS distributed crosslink database management (Landry 2000), professionals from all over the world could join forces to construct ontologies, term, and link databases which are freely accessible via the Web. Much of the content of these database is already available in the form of thesauri or subject heading catalogs. MACS and the like even provide for crosslinks between ontologies, a much-needed resource for further growth of the Semantic Web. It all is a matter of conversion to the appropriate standards (XML and RDF), publishing of the resource and a certain amount of advertising it, and most importantly, keeping it up to date. Quality resources that cover a specific area will be the first candidates for world-wide use in the Semantic Web, but there definitely is room for more general, broader-coverage resources, too.⁸

Eventually, joining the world's semantic resources into one, massive resource that is not necessarily one hundred percent correct or complete, but so large that it serves the majority of questions, would fulfill one of the long-standing goals of libraries worldwide: to assist the evolution of human knowledge as a whole (Berners-Lee 2001). It would be a pity, a shame, and a crime if libraries were not involved in this task.

4 References

Berners-Lee, T., Hendler, J. and Lassila, O. (2001): The Semantic Web. In: *Scientific American*, May 17, 2001.

Guarino N. (1998): Formal Ontology and Information Systems. In N. Guarino (ed.), *Formal Ontology in Information Systems*. Proc. of the 1st International Conference, Trento, Italy, 6-8 June 1998. IOS Press (amended version) <http://www.ladseb.pd.cnr.it/infor/Ontology/Papers/FOIS98.pdf>

Landry, P. (2000): The MACS Project. In: proceedings of the 66th IFLA Council and General Conference, Jerusalem, August 13-18, 2000.

Papazoglou, M. and J. Hoppenbrouwers (1999): Knowledge Navigation in Networked Digital Libraries. Keynote speech for the 11th European Workshop on Knowledge Acquisition, Modeling, and Management (EKAW'99). In: *Lecture Notes in Computer Science*, LNAI 1621, Springer Verlag, Heidelberg.

⁸<http://www.semanticweb.org/knowmarkup.html>