The Semantic Web in Practice: Opportunities and Limitations

Matthew Hirons
Centre For Internet Computing, University of Hull, Scarborough Campus
m.hirons@dcs.cic.hull.ac.uk

Abstract

The Internet and the World Wide Web have brought a revolution to information technology and the daily lives of most people. However, most of the current forms of web content are designed and structured for use by people but are barely understandable by computers. The lack of semantic mark-up is a major barrier to the development of more intelligent document processing on the Web. Current HTML markup is used only to indicate the structure and lay-out of documents, but not the document semantics.

The goal of the next generation web – the Semantic Web, with its vision by Berners-lee (1998), is to develop expressive languages to describe information in forms understandable by machines. It will bring structure to the content of Web pages, being an extension of the current Web, in which information is given a well-defined meaning.

There are many important technologies for developing the Semantic Web to replace HTML, which is no longer capable of standing up to the new challenges of Internet-based computing.

1. Introduction

This paper begins by discussing what the Semantic Web is and XML *(Extensible Markup Language, 1998)*, which is a major technology in moving towards the vision of the semantic web. However XML is one of many technologies being developed and has its limitations, which are discussed and how they can be solved in conjunction with other technologies. Opportunities that the Semantic Web will bring along with problems in its widespread adoption are discussed. Two case studies are provided of the Semantic Web in Practice: Google.com and Amazon.com. User's Views on the Semantic Web obtained via a simple questionnaire are also discussed.

2. What is the Semantic Web?

The Semantic Web is a mesh of information linked up in a way that is easily processable by machines, on a global scale. You can think of it as being an efficient way of representing data on the World Wide Web, or as a globally linked database. It is not a separate Web but an extension of the current one, which encompasses efforts to build a new WWW architecture that, enhances content with formal semantics [Palmer, 2001]. It isn't about pages and links, it's about relationships between things - whether one thing is a part of another, or how much a thing costs, or when it happened. This will better enable computers and people to work in cooperation.

Tim Berners-Lee, who along with Hendler and Ora Lassila first mentioned the Semantic Web in May 2001, told W3C members earlier this year that it is going to be very powerful, and fun. There is a dedicated team at the World Wide Web consortium (W3C) working to improve, extend and standardize the system, and many languages, publications and tools have already been developed. However, Semantic Web technologies are still in their infancies, and although the future of the project in general appears to be bright, there seems to be little consensus about the likely direction and characteristics. [Palmer, 2001]

3. Opportunities

Like other technologies, the interest in creating and developing the Semantic Web is motivated by the opportunities it might bring. Most of the Web's content today is designed for humans to read, not for computer programs to manipulate meaningfully. Computers can accurately parse Web pages for layout and routine processing—here a header, there a link to another page, but in general, they have no reliable way to process the semantics. Data is generally hidden away in the visually bloated HTML tagging language.

The question of whether a certain piece of information is on the Web has become the problem of how to find and extract it. The problem will become even more serious when the growth of the Web maintains its high speed as expected by the W3C. Therefore there a clear need for this next-generation Web now.

”Expressive meaning” is the main task of the Semantic Web [Berners-Lee, Hendler, Lassila, 2001]. It will enable automated agents to reason about Web content, and carry out more intelligent tasks for the user. Documents will be able to be queried based on their semantics, rather than on strings of characters that may occur in them. A Semantic Web is like the librarian who has read the books [Brennan, Petrosillo, 2003]. It has the potential to go beyond mere information retrieval to intelligent decision-making, dramatically improving your website's ability to meet client needs.

Users would be welcomed with accurate and relevant search results as opposed to the often frustrating current keyword-based matching searching techniques, where they frequently experience one of two problems: they either get back no results or too many irrelevant results. This is because from an end user perspective, a semantic website is better because it thinks and can interact more like a person [Brennan, Petrosillo, 2003]. It understands synonyms, but it also knows that all synonyms ultimately point to a single concept not an unrelated collection of concepts. For example, it would understand that the users request for mountain bike means the same thing as bicycle, off-road. As a result the search program can look for only those pages that refer to a precise concept instead of all the ones using ambiguous keywords. This raises the question of how much new business could you capture if every potential client who visited your website found exactly what they were looking for?

Furthermore, the same information can be delivered over a WAP interface, in writing, using speech synthesis over a mobile phone, or even via methods not yet invented. A semantic model is delivery independent.

Two important technologies for developing the Semantic Web are already in place: eXtensible Markup Language (XML) and the Resource Description Framework (RDF). These will be discussed in the following two sections respectively.

4. What is XML (eXtensible Markup Language)?

XML has brought great features and promising prospects to the development of the Semantic Web. It will have a profound impact on the way data is exchanged on the Internet [Shiyong, Ming, Farshad, 2002]. An important feature of this language is the separation of content from presentation, which makes it easier to select and/or reformat the data. Web-page creators use their own set of markup-tags, which can be chosen to reflect the domain specific semantics of the information, rather than merely its lay-out. For example:

```
<LOCATION>
His tel.nr. is <TEL>877834<TEL>,
room nr. <ROOM>145a</ROOM>
</LOCATION>
```

XML allows us to structure Web-pages as labeled trees, where the labels can be chosen to reflect as much of the documents semantics as is required. Although XML allows the use of any tags as long as they are properly nested in the document, it's possible to define restrictions on the set of tags that can be used in document. This is done in a Document Type Definition (DTD).

5. RDF (Resource Description Framework) – a solution to XML's Limitations?

XML will continue to play an important role in the development of the Semantic Web. However, it does not provide a full solution to its requirements. In short, XML allows users to add arbitrary structure to their documents but says nothing about what the structures mean, therefore further means are required for the Semantic Web and the role of XML is reduced to a syntax carrier [Harmelen, 2001]. This is where RDF comes into play.

RDF (Lassila and Swick, 1998) provides a means for adding semantics to a document without making any assumptions about the structure of the document. RDF encodes information in sets of triples, each triple being rather like the subject (statement), verb (property) and object (resource) of an elementary sentence. These triples can be written using XML tags. Therefore information is mapped directly to a model.

In RDF, a document makes assertions that particular things (people, Web pages or whatever) have properties (such as "is a sister of," "is the author of") with certain values (another person, another Web page). Subject and object are each identified by a Universal Resource Identifier (URI). The verbs are also identified by URIs, which enables anyone to define a new concept, a new verb, just by defining a URI for it somewhere on the Web. A simple example is:

```
Author(http://www.blog.com/matth) = Matt
```

This states that the author of the named Web document is Matt.

Values can also be structured entities:

```
Author(http://www.blog.com/matth) = X
Name(X) = Matt
Email(X) = matth@blog.com
```

Where X denotes an actual (i.e., the homepage of Matt) or a virtual URI.

There are differing views regarding RDF though. For example, "This is different from my personal, long standing view of RDF as a simple and rather awkward logic language. I *think* it explains why some people get so excited by the "graph nature" of RDF, which I just saw as a (personally uninteresting) notation." [Parsia, 2002]

6. Ontologies

There are further considerations because, because two databases may use different identifiers for what is in fact the same concept, such as zip code. A program that wants to compare or combine information across the two databases has to know that these two terms are being used to mean the same thing. Ideally, the program must have a way to discover such common meanings for whatever databases it encounters. [Berners-Lee, Hendler, 2001]

A solution to this problem is provided by the third basic component of the Semantic Web, collections of information called ontologies. According to Artificial-intelligence and Web researchers an ontology is a document or file that formally defines the relations among terms. Most kinds of ontologies for the Web have a taxonomy and a set of inference rules. The taxonomy defines classes of objects and relations among them. For example, an address may be defined as a type of location, and city codes may be defined to apply only to locations, and so on. Inference rules in ontologies supply further power. An ontology may express the rule: Cars are a type of vehicle and are associated with an engine. BMW vehicles are cars therefore BMW vehicles have an engine.

7. Problems with the Semantic Web

There are numerous challenges regarding the Semantic Web including the development of ontologies, and the development of the formal semantics of Semantic Web languages, and the development of trust and proof models.

The *cultural* future of the Semantic Web is tricky. Privacy is a huge concern, but too much privacy is unnerving. For example a group of people could come up with"ghost taxonomy" - a thesaurus that seemed to be a listing of interconnected yacht parts for a specific brand of yacht, but in truth the yacht-building company never existed except on paper - it was a front for a money-laundering organization with ties to arms and drug smuggling. When someone said "rigging" they meant high powered automatic rifles. Sailcloth was cocaine and an engine was weapons-grade plutonium. [Ford, 2002]

This could be possible and enable criminals to sell plutonium as smooth, easy and anonymous as selling laptops! Therefore it is vital that all RDF be referenced to a public taxonomy approved by a special review board.

A key reason for the apparent lack of progress with the Semantic Web according to Tim Berners- Lee is that "Human endeavour is caught in an eternal tension between the effectiveness of small groups acting independently and the need to mesh with the wider community." [Berners-Lee, Hendler, Lassila, 2001]

In short some call the movement a vision, while others note that the pieces are all there, but someone has to assemble the puzzle, and agree on standards to make sure all the pieces fit together. As a result it seems (to the general population) that not much progress has occurred since the idea of the Semantic Web was first suggested.

8. The Semantic Web in practice

Beyond the great wall of data on the Internet lies a goldmine for enterprises called the Semantic Web. W3C Semantic Web Activity Lead Eric Miller, who is spearheading the project, says bloggers are some of the first end users immersed in the social network of the Semantic Web. "Some of the tools here are things like TrackBack and syndication. If you use any consistent blogging system, that system is available to RDF; you can leverage RSS tools and ask questions like 'show me all the people who are talking about grid technology.' What you get back is a more relevant response regardless of the data set." [Singer, 2003]

Google.com and Amazon.com are two companies that have taken advantage of the Semantic Web to some extent.

8.1 Google

According to Google they, "search more sites more quickly, delivering the most relevant results". It's hard to believe Google, which is now the world's largest single online market, came on the scene not much more than a decade ago. Part of the reason for this is due to the Semantic Web. Google takes advantage of what might be called a limited semantic model. They derive a semantic model of the entire WWW, which empowers their search engine with an ability to perform queries with far more intelligence than any of their competitors [Brennan, Petrosillo, 2003]. This is reflected in Google's significant leading position in the search engine wars (see Figure 1).

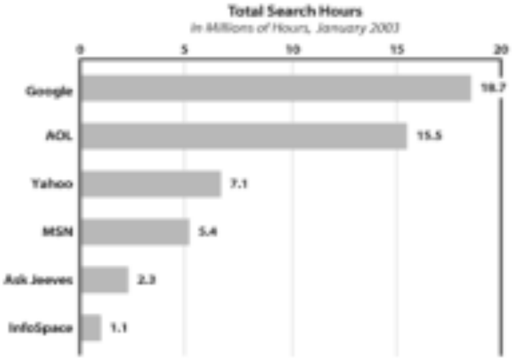

Figure 1. Search Engines compared: total search hours

"Google makes the Web a vastly nicer place to be" [Parsia, 2002]. They *reason* about hyperlinks, augmented by some heuristics about page composition. They combine PageRank (the heart of their software for ranking web pages) with sophisticated text-matching techniques to find pages that are both important and relevant to your search. Google goes far beyond the number of times a term appears on a page and examines all aspects of the page's content (and the content of the pages linking to it) to determine if it's a good match for your query. [Google, 2004]

However, because Google's model is derived from the Web rather than driving it, Google is unable to take advantage of some of the best features of semantic modeling: they cannot approve a single vocabulary for the entire Web. Even more importantly, they cannot decide which relationships should define a page. This fundamentally limits the sorts of thing Google can reason about.

If there were more information "in" the links than their presence, Google would be able to do much more. Plus the more machine understandable we make the content of the pages, the more likely search results will combine satisfyingly with link derived information.

It may be that a Semantic Google would be more vulnerable to trash input. Or that good typed links will be too hard to add, etc.

A final point to note, is that, when companies have power - and Google is getting *real* power over the way that information is disseminated - they need to be watched *carefully*. [Ford, 2002]

8.2 Amazon

Like Google, Amazon takes advantage of a limited semantic model. But instead of working from a derived semantic model, they add a semantic layer on top of what is essentially a syntactic database. Amazon begins with a standard database of books indexed by title, author, publisher, and ISBN number. They add value to this database by adding a layer of semantic modelling of their customers buying habits, using what they call recommendation algorithms. According to Amazon, these personalize the online store for each customer, radically changing it based on customer interests, showing programming titles to a software engineer and baby toys to a new mother for example. Amazon has significantly expanded over time selling new items (from books to electronics) and allowing thousands of users to sell them used as well.

Given that Google and Amazon, two of the greatest success stories of the Internet, owe their success partially to a partial semantic model, how much new business would be generated by a fully semantic website? And how do we get there? [Brennan, Petrosillo, 2003]

9. User's Views on the Semantic Web

A simple questionnaire was conducted to obtain the views of a selection of Web users (20 final year BSc Internet Computing students from The University of Hull, Scarborough Campus) regarding the Semantic Web including its relation with search engines and e-commerce sites. The results and explanations to each question asked are outlined below:

9.1 Which search engines do you use most?

As expected Google came out on top by a significant margin (100% chose it compared with 35% selecting Yahoo putting it in second place). Practically all the others went unnoticed. Google must be doing something right if it attracted all the users questioned and the partial semantic model Google uses is most likely the underlying reason.

9.2 Why do you use your selected search engines?

Many less experienced web users would likely just follow the crowd and select popularity as their reason for using a search engine. However, the group questioned can be classed as experienced Web users and their main reason was the relevance of the search results (65% chose this option). This clearly expresses the need for widespread adoption of the Semantic Web as it will greatly improve the relevance of hits (see section 3. Opportunities). Number of hits were less important (10%) because many hits are no good if they aren't relevant and take ages to go through as is sometimes the case. Speed, as expected wasn't particularly important to the users because none of the popular search engines are noticeably slower than each other.

9.3 Do you have problems finding relevant search results such as getting back no results or too many irrelevant results?

With all search engines (including Google) users are likely to have some problems searching for exactly what they want. This is reflected by the results from the users questioned as 100% agreed to having problems sometimes. The reason for this is because existing keyword-based search retrieves irrelevant information that uses a certain word in a different meaning or it may miss information where different words about the desired content are used. Again this expresses a need for the Semantic Web to have an impact, to make using the Web less frustrating and be more productive.

9.4 How important are personalisation features to you with e-commerce sites such as Amazon?

As described above Amazon uses semantic modeling of their customers buying habits to create a personalized online store. The results obtained from the questionnaire (60% chose either 3 or 4 out of 5) clearly show that this is welcomed by users and makes e-commence a more pleasurable experience. The Semantic Web of the future has the opportunity to allow personalization features to a larger extent than Amazon at present. However nobody selected 5/5 so still 40% chose 1 or 2 out of 5. This shows that some users aren't particularly bothered about fancy personalization features and just want to get the purchase over with quickly. Plus recommendations can be inaccurate and sometimes annoying.

9.5 Do you recognise the advantages in migrating web documents towards the Semantic Web?

As outlined above the Semantic Web will bring many opportunities but as many of them are still a vision and not in practice many users cannot clearly understand them at present. The user sample is biased though, as they have been introduced to the advantages. However 20% still agreed to not recognising them.

9.6 How confident would you be in developing a web site using Semantic technologies such as XML and RDF rather than just HTML?

Even though the entire user sample has had some practical experience in using XML, RDF and some related technologies, few have the confidence to migrate from using the mature HTML language. This is most likely related to the previous question in that the advantages of migrating are not clear. There are many technologies being suggested and this may overwhelm developers as which to learn and use. The learning complexity and learning curve is greater than the relatively simple HTML. Only 20% rated their confidence as 4 or 5 out of 5.

10. Conclusion

The structure of the Semantic Web will open up the knowledge and workings of humankind to meaningful analysis by software agents, providing a new class of tools by which we can live, work and learn together. [Berners-Lee, Hendler, Lassila, 2001] In this paper, Semantic Web concepts and technologies were discussed and in particular the opportunities that this new revolution will bring to us were considered along with case studies. The challenges that we are facing during the development of the Semantic Web were presented. User's views analysed using a simple questionnaire stressing that they think the Semantic Web will and already is having a positive effect, but the technologies are quite confusing and barely used at present. A potentially unrepresentative sample of people were questioned. If non-computing students were involved a much greater lack of understanding (or even complete) would likely be expressed regarding the Semantic Web.

The Semantic Web is still a vision, but the Web will likely grow towards this vision in a way like the development of the real world: Semantic Web communities will appear and grow first, and then the interaction among different communities will finally interweave them into the Semantic Web. [Shiyong, Ming, Farshad, 2002]

11. References

[Berners-Lee, Hendler, Lassila, 2001] Berners-Lee, T., Hendler, J., Lassila, O., The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities, Scientific American, http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21, 2001.

[Brennan, Petrosillo, 2003] Brennan K, Petrosillo S, Demystifying the semantic web: is migration right for you? 2003.

[Fischer, 2000] Fischer, P., "Migrating from HTML to XML", http://www.newarchitectmag.com/archives/2000/07/, 2000.

[Ford, 2002] Ford, P., August 2009: How Google beat Amazon and Ebay to the Semantic Web, http://www.ftrain.com/google_takes_all.html, 2002.

[Ford, 2002] Ford, P., A bit of commentary on Google and the Semantic Web, http://www.ftrain.com/google_semweb_commentary.html, 2002.

[Harmelen, 2001] Harmelen, F., Practical Knowledge Representation for the Web, 2001.

[Palmer, 2001] Palmer, S.B., The Semantic Web: An Introduction, 2001.

[Parsia, 2002] Parsia, B., A simple, prima facie argument in favor of the Semantic Web, http://monkeyfist.com/articles/815, 2002.

[Shiyong, Ming, Farshad, 2002] Shiyong, L., Ming, D., Farshad, F., The Semantic Web: Opportunities and challenges for next-generation Web applications, http://InformationR.net/ir/74/paper134.html, 2002.

[Singer, 2003] Singer, M., Semantic Web: Out of the Theory Realm, http://siliconvalley.internet.com/news/article.php/307696, 2003.

[W3C, 2001] W3C Semantic Web Activity, http://www.w3.org/2001/sw, 2001.