ONLINE Currents
Vol 18 (7) September 2003
Australia's journal for users of online services, CD-ROMs and the Internet

'OPEN THE POD BAY DOORS, PLEASE, HAL': HERE COMES THE SEMANTIC WEB

by Belinda Weaver

You may have heard mention of the Semantic Web in relation to Internet development, even if you have no real understanding of what the term actually means. If you have not, you are in good company, since even its developers and greatest advocates are not exactly sure about the final direction and scope of the project. The World Wide Web Consortium (W3C), headed by Tim Berners-Lee, the original creator of the World Wide Web, is working hard on Semantic Web development, and is doing its best to get the concepts out to ordinary Web users like us. But even so, the message is not always getting through, or is becoming garbled in the process.

The same thing happened to Berners-Lee, when he initially tried to demonstrate his concept of a 'Web' based on hypertext links to people who had barely pointed a mouse, much less clicked on a line to open a Web page. New concepts take time to get through to people, and it is often much further down the track that the Eureka moment kicks in; the 'Oh, now I get it' moment that proves an idea really works.

...these machines can then do all the hard work of sorting and sifting and weighing up

For committed Web users, the idea of being without the Web, of no longer having the convenience of online information on tap seven days a week, or of having to live without the quick to-and-fro of e-mail communication or instant messaging, would be truly terrible. Yet even those who most use and need the Web do not always find it very usable. Information exists in a bewildering array of formats. Despite enormous advances, finding tools are still relatively primitive, and the mechanisms for turning an avalanche of sometimes incompatible and unrelated information into usable, practical, targeted knowledge are still largely in our own heads. *We* do the sifting, *we* do the sorting, *we* make the judgment calls about what is relevant and what is not, what is reliable and what is rubbish.

Admittedly, different kinds of knowledge management systems have been adopted by institutions in an attempt to tame the tsunami of unrelated facts, hard data, research and stray morsels of knowledge that abound in any moderately-sized organisation. Yet each of these islands of coherence, however effective, is just that - an island – and, as such, cannot provide a generalised way forward to the question of making data usable.

The Semantic Web is an attempt to solve this problem. In the context of the project, 'semantic' simply stands for 'machine-processable'. If information can be made comprehensible to machines such as computers, these machines can then do all the hard work of sorting and sifting and weighing up that is currently done (very imperfectly) by humans, and, because they *are* computers, they can do it more quickly and on an unimaginably huge scale. In addition, they can learn on the job so that they can make an even better fist of it the next time around, and better again the time after that.

If it sounds like science fiction, then that is not too surprising since artificial intelligence has been the stuff of science fiction for decades. The worry that machines will become *too* smart, will – in fact – take over, as the wayward onboard computer HAL tried to do in the film *2001*, is still around, in movies, and in doomsday scenarios about self-replicating nano-machines taking over the planet or triggering global war. Yet the Web will never reach its full potential if the kinds of projects and actions proposed in Semantic Web work are not adopted.

Unfortunately, Berners-Lee may not be the best choice for the job of gospel spreader for the Semantic Web. This definition, produced by the W3C's Semantic Web Activity Group, states:

'The Semantic Web is the representation of data on the WWW. It is a collaborative effort led by W3C … based on the Resource Description Framework (RDF), which integrates a variety of applications using XML for syntax and URIs for naming. The Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.' *...cont p.4*

Published by Enterprise Information Management Pty Ltd A.B.N. 93 002 726 669

...cont from p.1 Open the Pod Bay Doors, Please, HAL

Read that and you may feel you are none the wiser about the real revolution that the Semantic Web could be poised to deliver, still less that the concept could be 'devastatingly valuable', nor that it 'excites people' (Berners-Lee, 2001). Berners-Lee admits that it 'takes a bit of imagination to realise that if all the databases in the world were linked together, there are all sorts of possibilities' (*Guardian*, 2003).

Edd Dumbill's article, *Building the Semantic Web*, adapted from his closing keynote address to the Knowledge Technologies conference in 2001, makes a better case and is a good general introduction to the whole concept of the Semantic Web. (Dumbill, 2001)

> 'The essential aim of the SW vision is to make Web information practically processible by a computer. Underlying this is the goal of making the Web more effective for its users. This increase in effectiveness is constituted by the automation or enabling of things that are currently difficult to do: locating content, collating and cross-relating content, drawing conclusions from information found in two or more separate sources … Speaking personally, I have a fundamental excitement at being able to recover and integrate my data from disparate sources and proprietary formats. This springs from constraints on my time, the difficulty of finding information, and the redundancy of having my data scattered across multiple devices.'

But even that summation is still a little dry and lacking in detail.

A scenario might help. Imagine you have found a Web page online for a forthcoming conference in your city. Your Semantic Web software agent knows that you are interested in such things, so it books you into the conference, notifies your electronic diary of the dates it is on (after checking first that there are no clashes with other important meetings), and alerts your bank to make the payment by the scheduled date, after checking your memberships to see whether you qualify for a discount. The day before, an instant message is sent by your diary to your mobile phone to alert you to the event. Travel schedules are then sent to the mobile to allow you to choose how you will make your way to the venue.

Seamless. Effortless. All of the above steps are done now when we see an event we want to attend. But, in the world of now, we are forced do every step manually. We type the entry in the online diary, we arrange the payment, perhaps by telephone, perhaps online, or maybe even through some complicated invoicing system at work. Although all the information needed is probably computerised – in an online diary such as Outlook, in our Web-enabled bank accounts, in spreadsheets or Palm Pilots – nevertheless each component is walled off from the next by the difference in the applications that hold the data. What the Semantic Web aims to do is to bring down the walls between applications, so that data in one type of system can be used and re-used by a completely different kind of application. Data can be described in such a way that one man's zipcode can equal another man's postcode, making previously incompatible data able to be shared.

To some extent, this is already happening in some places. Consultants make a good living writing 'bridge' software to help organisations share data between previously incompatible kinds of database systems. Yet, different kinds of 'bridge' software cannot deliver the seamless world of information sharing and management in the scenario above, since each bridge is essentially a one-off, a quick link from island to island, to solve one specific problem. To achieve the seamlessness that the Semantic Web project proposes means working towards standards for the description of data, so that data can be shared on a truly grand and global scale, not just between Joe in Accounts and Megan in Customer Service, but between businesses, governments, universities and research centres, NGOs – in short, by anybody who wants to put their data into the loop.

Aaron Swartz (2002) says:

> 'One of the best things about the Web is that it's so many different things to so many different people. The coming Semantic Web will multiply this versatility a thousandfold. For some, the defining feature of the Semantic Web will be the ease with which your PDA, your laptop, your desktop, your server, and your car will communicate with each other. For others, it will be the automation of corporate decisions that previously had to be laboriously hand-processed. For still others, it will be the ability to assess the trustworthiness of documents on the Web and the remarkable ease with which we'll be able to find the answers to our questions - a process that is currently fraught with frustration.'

What the Semantic Web will do is 'bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users … For the Semantic Web to function, computers must have access to structured collections of information and sets of inference rules that they can use to conduct automated reasoning' (Berners-Lee, 2001). The kind of knowledge representation necessary has been around for a while, as people have tried to build artificial intelligence machines. As Berners-Lee notes: 'The challenge for the Semantic Web therefore is to provide a language that expresses both data and rules for reasoning about the data and that allows rules from existing knowledge-representation systems to be exported onto the Web'.

Two of the important building blocks for the Semantic Web are already in place – eXtensible markup language (XML) and the Resource Description Framework (RDF). HTML, the language of the Web, is very good at

➥

allowing users to visualise information online, since it manages the display of information in an orderly fashion. However, HTML cannot really provide enough information for software programs to easily find and interpret that information. XML was developed by the W3C to allow information to be more meaningfully described using tags. While XML allows users to create any number of their own tags which, though hidden from view, annotate some or all of the text on a Web page, it alone is not the answer to creating the Semantic Web, since it 'has a limited capability to describe the relationships (schemas or ontologies) with respect to objects' (DAML, 2003). While programs can be developed to make use of such tags, the program writer first needs to know what each tag meant to the person who put it there. As Berners-Lee states: 'XML allows users to add arbitrary structure to their documents but says nothing about what the structures mean.'

Meaning is supplied by the 'RDF, which encodes it in sets of triples, each triple being rather like the subject, verb and object of an elementary sentence … In RDF, a document makes assertions that particular things … have properties … with certain values.' The subject, verb and object are all identified by individual Uniform Resource Identifiers (also known as URIs). (Uniform Resource Locators, or URLs as they are more commonly known, are the most common form of URIs.) Once these triples have been created, they create 'webs of information about related things' (W3C FAQ about RDF, 2003). Since anyone can use the URIs to find this information, the data is not locked away in a document but becomes accessible to anyone. The aim of RDF is to provide interoperability across applications, for example, allowing you to import your bank statements into your calendar. It began as a 'framework for metadata, thus providing interoperability between applications that exchange machine-understandable information on the Web' (W3C FAQ about RDF, 2003). The W3C believes RDF has a bright future, with multiple uses – for resource discovery, cataloguing, for content rating, for describing collections of documents at a site or the intellectual property rights that govern their use, and for use by intelligent software agents in knowledge-sharing and exchange. Coupled with the increasing use of digital signatures, the W3C views RDF as the key to building a 'Web of Trust' for collaborative ventures, such as research and ecommerce.

Since trustworthiness of data is such an issue in the 'lucky dip' Web world we use now, anything that can identify and guarantee provenance will be helpful. Digital signatures can be used to establish the provenance not only of data, but also of ontologies, helping people understand where all kinds of different information has come from, so they can make their own decisions about whether or not to trust the information.

Another Semantic Web building block is the use of ontologies to describe the relations among terms in use. An ontology is the tool that allows one man's zipcode to equal to another man's postcode, since it contains a taxonomy and a set of inference rules. As Dumbill states: 'Ontologies provide the ability to say "my world is like this" and are the foundation that will enable programs to reason about different worlds and environments and make connections between them' (Dumbill, 2001). The DAML language (DARPA Agent Markup Language) is an extension to XML and RDF, which provides a set of tools to create ontologies and to mark-up information, so that it is both machine-readable and machine-understandable. (DARPA is the US Defense Advanced Research Projects Agency whose original network, ARPANet, marked the beginnings of the Internet. DARPA's DAML activity will play a large role in Semantic Web development.)

There is a lot happening on this front now that both the European Union and the US government have earmarked funding for the project. Will it develop in the way we want it to?

Long ago, Berners-Lee wrote a Semantic Web Road Map (Road Map, 1998), that sets out in detail the W3C's concept, and the different layers that make it up. Berners-Lee's diagram, at **http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html**, shows how the layers will eventually build up to form the Semantic Web. Activity now is concentrated more in the lower layers of XML, RDF and so on, but as these bed down and become more universally used, then higher level development can occur. Digital signatures are important at all levels in creating a Web of more trustworthy data.

Dumbill (2001) states three criteria for getting Semantic Web activity right:

> 'Simple protocols, concepts and syntax: the easier the component parts of the SW are to learn, the quicker they will spread in adoption. Of course there is a tension here, but on the Web widespread adoption is something that can be set against complexity. There is ultimately more power in a simple technology universally adopted than a more powerful one with patchy or little adoption.
>
> Low barrier to access: the SW should be something which normal users have easy access to, in the same way that it's very easy to read the Web, and relatively easy to set up and publish a Web page. We run into tool-dependencies here, but that's not a blocker, as most non-HTML-savvy folk use an authoring tool to publish. The point is that SW technology must become commoditized.
>
> Tangible utility: this may seem obvious, but the Web actually does something people want. There's a danger with the SW, as with any technology, that its developers get carried away with ideas that end up being clever but hardly useful. The use cases for the SW must begin at home and describe practical problems.'

It might be timely to ask 'What's in it for me?' Certainly, no-one will go to the bother of coding material in XML,

building ontologies or describing their data in RDF for no return. There have to be tangible gains for people to adopt new technologies, and the paths towards adoption have to be relatively smooth. Should you wish to experiment a little, one place to find out more is at the Semantic Web Community Portal, where news of initiatives, development and projects abounds. Just about everything you need to know about ontologies, markup languages such as DAML, annotation tools, standards and Semantic Web resources is there. Between the W3C's own site and this, you should be awash in information. A look at the vast array of events and projects listed in the resources page alone is proof enough that the Semantic Web work is alive and well. There are annual international conferences on the Semantic Web, the next planned for Florida in October this year. Papers from earlier conferences as well as information on the latest conference are available online, at **http://iswc.semanticweb.org/**. Anyone who thinks that software agents are a thing of the future should visit AgentLand (**http://www.agentland.com/**) or BotSpot (**http://botspot.com/**) and see the vast range of existing agents already developed for uses as diverse as Net searching, software downloads, gambling, shopping, virtual assistants, monitoring agents, and Web agents of various kinds. Anyone feeling brave enough can try to develop their own little Semantic Web helper application.

What the portal provides too, is practical help, by pointing users towards Semantic Web projects of use to their field, for example, the standardization efforts in human resources, telecommunications, business processes and paper supply. It is in this type of concrete, practical work that the usefulness of the Semantic Web will be tested. If specific problems can be solved, then that information can be passed on for adaptation elsewhere. Solving the problems of interoperability (or not) is what the Semantic Web is all about. If it can deliver that, we can all benefit. As Dumbill rightly says:

'The SW represents an enormous opportunity not just to solve our problems with information management, but also to solve them in an interoperable environment, so we can all share solutions and enjoy the network effect. But always the goal should be to make the Web more effective for the user, and it is by such that it will be judged.'

References

Berners-Lee, Tim (n.d.) Scientific publishing on the "semantic web" ', *Nature* web debates. **http://www.nature.com/nature/debates/e-access/Articles/bernerslee.htm**

Berners-Lee, Tim et al. (2001) 'The Semantic Web', *Scientific American*, 17 May. **http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21**

Berners-Lee, Tim (1998) The Semantic Web Road Map. **http://www.w3.org/DesignIssues/Semantic.html**

DAML.org (2003) About the DAML Language. **http://www.daml.org/about.html**

Dumbill, Edd (2000) 'Berners-Lee and the Semantic Web Vision'. **http://www.xml.com/pub/a/2000/12/xml2000/timbl.html**

Dumbill, Edd (2001) 'Tim Berners-Lee on the W3C's Semantic Web Activity'. **http://www.xml.com/lpt/a/2001/03/21/timbl.html**

Dumbill, Edd (2001) 'Building the Semantic Web'. **http://www.xml.com/pub/a/2001/03/07/buildingsw.html**

Dumbill, Edd (2001) 'The Semantic Web: a Primer'. **http://www.xml.com/pub/a/2000/11/01/semanticweb/index.html**

Gibson, Owen (2002) 'The Next Step', *Guardian*, 28 October. **http://www.guardian.co.uk/Print/0,3858,4533430,00.html**

Hendler, James et al. Integrating Applications on the Semantic Web. **http://www.w3.org/2002/07/swint**

Pease, Adam (2002) 'Why Use DAML?' **http://www.daml.org/2002/04/why.html**

SemanticWeb.org (2003) Semantic Web Community Portal. **http://www.semanticweb.org/**

Swartz, Aaron (2002) The Semantic Web in Breadth. **http://logicerror.com/semanticWeb-long**

World Wide Web Consortium (2002) Frequently Asked Questions About RDF. **http://www.w3.org/RDF/**

World Wide Web Consortium (2001) Semantic Web. **http://www.w3.org/2001/sw/**

Belinda Weaver is the Coordinator of ePrints@UQ, the University of Queensland Library.

NET NOTE

AUSTRALIAN ISP TAKEOVER

Chariot (**http://www.chariot.com.au**) is an Adelaide-based, publicly listed, Internet Service Provider (ISP). It has recently purchased two ISPs, Armidale-based Blue Pin and Brisbane-based Squirrel Net. Chariot is one of Australia's top 10 listed ISPs. Its previous purchases were of Picknowl, Cyberwizards, Mr Bean, better.net and Ecite ISP businesses. Its coverage now extends from Cairns, Mackay, Brisbane, Sydney and north eastern NSW, Melbourne and western Victoria, to Adelaide.