



WHEN THE WEB STARTS THINKING FOR ITSELF

To start with, the semantic web will tag documents so search engines know what information they contain. But soon it will automatically process data intelligently – or, in other words, says David Green, the web will begin to think.

Through enabling easy, widespread publishing, the web has had massive social consequences – dramatically altering human behaviour and expectations in information retrieval, knowledge sharing and collaborative working. However, the web as it currently exists, makes searching and data exchange difficult.

In September 1998, Tim Berners Lee, the creator of the web, outlined a vision of how it could evolve to address this flaw. His 'Semantic Web Road Map' on the W3 Consortium website (W3C is the non-profit body that co-ordinates global web standards development) has helped spawn research all over the world into the standards and infrastructure that could deliver a web in which low-level information retrieval and processing is automated. The semantic web is an extension of the current web in which data is given meaning through the use of a series of technologies that are explored in detail below.

In the semantic web, data is given meaning through 'semantic markup' – that is, markup tags that are interpreted as an expression of a document's content, rather than tags that are interpreted for display (as with the web display format HTML).

First, documents will be published with 'semantic markup' – that is markup not interpreted for display (as with HTML) but as an expression of the document's content. This fundamental shift in web publishing will have far-reaching repercussions for web search engines. Rather than visit a search engine and trawl through a flat listing of possible matches, users will be able to issue high-level information requests and receive a distilled answer.

The semantic web is intended to complement humans in areas in which they do not perform well, such as processing large volumes of information quickly or analysing large texts for certain pieces of information. It will also extend to the 'real' world where appliances will advertise their functionality through smart chips and tags. For example, mobile phones could describe their display parameters so web content can be customised for them on the fly.

The semantic web is based on established technologies such as XML, RDF, ontologies and Intelligent Agents. XML is the successor to HTML (see *IWR*, December 2001). It is a semantically focused open technology that allows far greater possibilities than mere metadata because it allows a publisher to address the meaning of its content. XML enables powerful structured query searching on text web pages, allowing the user direct access to relevant segments of information within a document.

Through related formatting standards such as XSL, XML allows manipulation of how data should be formatted and output. This takes the web page beyond a flat display of data and allows the user to manipulate the data – creating the ability to write data once and then publish it to any device in real time.

Although publishers create their own arbitrary XML tag structure, XML schema explain the

Agents will dynamically adapt pages and add in links to related content. The 'intelligence' of this dynamic self-organising web, where popular links are prominent, and rarely used links will diminish, will gradually arise through the assembly of the limited intelligence of autonomous agents and systems.

publisher's structure by defining shared mark-up vocabularies and providing hooks to associate semantics with them. Every major player in the technology industry is touting this XML-driven interoperable future. Indeed, in the database arena, XML has already become the standard approach for distributing data from one application to another.

Another technology, the Resource Description Framework (RDF) provides meaning to the structure of XML documents. Just as in human language, where meaning is expressed in a sentence composed of subject, verb and object, RDF helps to express meaning and relationships between different web pages and concepts through a programming structure of things, properties and values.

For example, David Green (thing) is the author of (property) this and other *IWR* articles (value). Subject, object and verb (or thing, property and value) are encoded in the document through a uniform resource identifier (URI) which ensures that the words on the document are linked to a unique definition that everyone can access. This enables data interchange between systems.

However, while RDF allows a publisher to inform a visiting computer which terms it has used to tag the content in a document, different publishers will use different terms/identifiers to express the same concept. Ontologies provide a deeper level of meaning by providing equivalence relations between terms (ie, term A on my web page is expressing the same concept as term B on your web page). An ontology is a file that formally defines relations among terms, for example, a taxonomy and set of inference rules.

By providing such 'dictionaries of meaning' (in philosophy ontology means 'nature of existence') ontologies can improve the accuracy of web searches by allowing a search program to seek out pages that refer to a specific concept rather than just a particular term as they do now.

While XML, RDF and ontologies provide the basic infrastructure of the semantic web, it is intelligent agents that will realise its power. An intelligent agent can be best described as a piece of adaptive computer coding that is capable of reasoning and that learns from our behaviours and preferences (thus delivering what is called 'proactive personalisation').

There are many thousands of different agents (or bots as they are also known), each performing specific, specialised tasks (for example, search bots, chatter bots and shopping bots, etc). An

important aspect of agents is that they are sociable – they can interact and communicate with humans and other agents. In the semantic web, different agents work together to create an information value chain in which the user's search request is 'packet processed' through sub assemblies of information passed between agents – each adding value to construct the user's answer.

The process works as follows. The user will issue a high-level information request. An intelligent agent will then analyse this request and delegate it to other appropriate agents and services that it has each identified through service directory ads on the web. These agents will distil large amounts of data distributed across the web and progressively reduce it to a much smaller amount of high-value customised information – in other, words the answer.

When broken down into a series of explicit search statements and appropriate content sources to search, a simple user information request is revealed to be a complex task. Automating such tasks will result in an ever-larger role for artificial intelligence technologies such as agents.

One key concern about the brave new world of bots is that by increasing their autonomy their accountability will be lost (see *IWR*, Dec 2000). The question is, how much information about our behaviours and content preferences should agents digress to other agents, databases and systems? There is a need to construct boundaries such as user-determined privacy settings to safely contain such interactions.

Similarly, agents will need to authenticate the veracity of content sources and other agents they meet through the use of digital signatures – this is of particular concern when much future crime will involve the more profitable theft of personal details rather than artefacts. Recognising this, the Joint Research Centre of the European Commission is building an experimental privacy protection agent using semantic web technology and W3 Consortium's P3P privacy protocol. The agent will partially automate the process of protecting a user's privacy by comparing privacy policies and a user's privacy preferences.

The semantic web and other developments such as grid computing (where any one computer can tap the power of all computers) and an internet operating system, have given rise to the concept of a 'global brain'. Populated with adaptive, reasoning agents, the web will act as a global super-organism – the brain of society. It will open up humanity's collective knowledge to

meaningful analysis by agents, identifying undiscovered relations between concepts and enabling communication of concepts even where there is no commonality of terms.

Agents will also dynamically adapt pages and add in links to related content – identifying connections between concepts. The 'intelligence' of this dynamic self-organising web, where popular links are prominent, and rarely used links will diminish (just like neurons), will gradually arise through the assembly of the limited intelligence of autonomous agents and systems, each operating within defined context parameters.

The Open Directory project's slogan is 'Humans do it better'. Tim Berners Lee's vision is that the semantic web will do it better ('it' being low-level information discovery and exchange), thus enabling humans to do better things. This symbiotic intelligence of people, plus computers, plus AI agents that offer immediate access to humanity's collective knowledge does sound somewhat utopian. Equally it raises the spectre of a self-adaptive intelligence that quickly surpasses our ability to comprehend it.

Would such a global brain act as a digital dictator to whom individuals would have a secondary role to society's demands? Two papers published in the 9 September 1999 issue of the scientific journal *Nature*, revealed the internet appeared to be 'evolving' rather than following

the expected model of random inanimate networks. Quoted in the June 2000 issue of *New Scientist*, Daniel Dennett, director of the Centre for Cognitive Studies at the University of Medford, Massachusetts, commented that 'the global communication network is already capable of complex behaviour that defies the efforts of human experts to comprehend'.

The upshot is that the semantic web may act as a 'collective memory' augmenting individual brain power and accelerating the pace of human learning and discovery – but we will need to be careful about controlling its development and our dependence on it if we wish to avoid a dystopian digital dictator scenario.

David Green is the web and e-marketing manager, EMEA at Deloitte Touche Tohmatsu. His personal website is www.davidgreen.me.uk

- Useful links**
www.semanticweb.org
www.ontoweb.org
W3 Consortium Semantic Web activity:
www.w3.org/2001/sw
DARPA Agent Markup Language (DAML):
www.daml.org
Semantic Web article in Scientific American:
www.scientificamerican.com/2001/0501issue/0501berners-lee.html



THE PRESENT & FUTURE OF SEARCH	
Search today	Search tomorrow
Search for term	Search for concept
Historical Indexes	Real time environment
Manual process - explicit instructions via search engine site	Automated process - state high-level goal via PC or other device
Machine display: HTML = content presentation	Machine processing: XML = content meaning
Results - as published on each page	Results - distilled from many sources
Flat listing - relationships between data sets not presented	Visualisation of concept space - data relationships presented
Defined data types - for example, HTML and PDF	Many data/file types
Constellation of computers via PC client/server model	Anything anywhere - distributed, peer-to-peer