



**CENTRAL
CO-ORDINATION:**
Bente Maegaard
Center for Sprogteknologi
Njalsgade 80
DK - 2300 Copenhagen S
Tel: +45 35 32 90 90
Fax: +45 35 32 90 89
euromap@cst.dk
www.hltcentral.org/euromap

HLT-Powered Web Technology

Human language without meaning is, ... well, ... meaningless. Andrew Derrington shows how the Semantic Web adds meaning to web pages to support the next generation of HLT-powered web technology.

The web is too much for humans to handle.

An optimist looks at her drink and smiles with satisfaction that it is half full. Her pessimistic companion frowns because his is half empty.

If they were computer scientists contemplating the World Wide Web as it is today, the optimist and the pessimist would also disagree. She would be delighted at the staggering quantities of data from web sites scattered across the planet that can be sorted and delivered to her screen at the click of a mouse. He would complain that once information has been delivered to his screen, he has to read, interpret, collate and process it.

All these jobs could be done better and more reliably by a computer than by a human.

They are both right of course. And they are getting righter. The web is growing exponentially. That generates a massive explosion of information that poses problems for optimist and pessimist alike.

You can easily see the scale of this problem for yourself. Type a common word into your favourite search engine. In a fraction of a heartbeat it will offer you more web pages than you can read in the whole of the rest of your life.

So let computers handle it.

Fortunately, there will be a way of handling this explosion of information. It is called the Semantic Web.

Conceived by the architects of the original Web, the Semantic Web is still a babe in arms. When it is fully developed, it will enable computers to talk meaningfully to each other.

In their digital conversations our computers will accomplish many of the transactions that today we have to do ourselves. And they will do it better, quicker and more reliably than we can.

So exactly how will this be different?

To understand how it will be different, imagine you are booking flights for a Christmas holiday in Tenerife. If, like me, you leave the booking until mid-December, the only option the web travel agent offers from your nearest airport costs five times what you want to pay and involves changing planes in Amsterdam and Madrid.

You then realise that the budget airlines, the ones that offer really cheap flights, only sell direct to the customer. You have to look on each of their web sites. Eventually, after several hours of trying dozens of different date combinations on the web sites of three different airlines – all of which work differently - you find an affordable direct flight from another airport. You then need to look up train and bus timetables and car-parking prices to arrange your ground transportation to the airport.

This is a huge advance on pre-web days. All the information you need is out there now and can be delivered to your desktop. Most of the transactions you need can be carried out over the web. You can do it. But why is it so difficult? It takes hours of sitting at the computer, reading the screen, clicking your mouse.

The world of the semantic web will be much easier. You won't touch the mouse once. Your computer - which will probably be inside your mobile phone - will do all the searching, make the reservations and buy the tickets. It will know your favourite airline, where you like to sit, and what dates in your diary are free for a holiday. It will even check the punctuality records of the airlines, so you won't, as I did two years ago, end up driving 150 miles to an airport for a midnight flight that didn't leave until noon the following day.

And why can't it happen now?

According to James Hendler, full professor in the University of Maryland Institute for Advanced Computer Studies, the main reason this can't happen now, despite the fact that all the relevant information is on the web, is that the information is not in a form that allows computers to interpret its meaning. "There is no world wide web of data" he says.

HTML, hypertext mark-up language, which is what web pages are written in, defines the appearance of the page and the links to other pages. It does not define the meaning of the text on the page.

We can read the text and interpret its meaning almost without effort. Programming a computer to do the same thing would be a gargantuan task.

How to make web pages intelligible to computers: XML and RDF.

So we need a way of putting computer-understandable data on web pages. There are two parts to this. First, we need a mark-up language that, unlike HTML, allows tags that represent computer-understandable meaning to be inserted into web pages. Then we need a way of defining the meanings.

Both of these things already exist in principle, so now all we need is a practical way of putting them together to make it possible for computers to look up the relevant definitions when they read an appropriately marked web page. "That's easy to say but hard to do." says Hendler, who is one of the leading lights in the spinning of the global semantic web.

On the face of it, it seems straightforward enough. The two key components are XML, or extensible mark-up language, and RDF, or resource description framework.

XML allows new tags to be created that give defined, computer-understandable meanings to entries on web pages. For airline booking pages these would be entries like <ticket price> or <departure airport>. XML makes it possible to create the words of the language for transactions between computers on the Semantic Web.

Creating the right words makes it possible to write meaningful data on the semantic web. But how will the computers that read the data know what it means? There has to be a way for them to look up the meanings of the tags.

This is where the RDF comes in. It allows the tags to be defined by stating their properties in terms of their relationships to one another. The properties are stated as assertions of the form "subject X" "has the property" "Y". So a computer-readable version of the page you are reading might have the assertion that "Andrew Derrington" "is the author of" "this story".

The three elements of the assertion, the subject, the verb and the object, would all be in the form of Universal Resource Identifiers. URI's, are analogous to the URLs in HTML. They are references to the locations of the original definitions of the terms on the web.

A tower of babel for computers? Enter the ontology.

The combination of XML and RDF makes it possible to encode information and to specify how it may be decoded. But it also sounds like a good way to build a tower of babel for computers. If anybody can define new terms just by placing the URIs on the web, how do we keep track of all the new terms that have been coined? How do we decide when two different terms mean exactly the same thing?

This is where the ontology comes in. Ontologies are the Rosetta Stones of the semantic web. They provide two things, taxonomies, and inference rules. Both are important devices for representing and manipulating knowledge in computers.

A taxonomy is a classification of objects and their relationships to each other. So if two terms mean exactly the same thing, an ontology may define them as equivalent and they can be used interchangeably. Similarly, one term may be defined as a

subcategory of another. For example “twin”, “brother” and “sister” are subcategories of “sibling”. This means that everything that is true of the category “sibling” is also true of its subcategories “twin”, “sister” and “brother”.

Inference rules specify what may be deduced by combining specific types of information about specific categories of object. They provide a computer with the basis for processing the information on semantic web pages to derive knowledge.

Can we trust knowledge derived from the web?

All this sounds pretty straightforward. Representation and manipulation of knowledge are well-established activities within computer science and artificial intelligence. Surely this means that we have lots of techniques already available for exploiting the semantic web?

It is not quite so simple. There may be established techniques for manipulating knowledge, but there are also established problems in ensuring the reliability of that knowledge. Inferences that work consistently on well-controlled datasets may give completely the wrong result when applied in the ramshackle world of the web. According to Nigel Shadbolt, Professor of Artificial Intelligence at the University of Southampton and Director of the UK collaborative project, ‘AKT’, we simply do not know whether it is possible to scale ontologies to encompass large numbers of users and large amounts of content or to transfer them from one context to another.

Hendler gives an example of the problem of transferring the same terminology between contexts. In the US airforce both flight schedulers and maintenance depots talk about airplanes. We would expect the word “airplane” to mean the same thing to both groups. But it is not so. For the flight scheduler, the most important component of the airplane is the munitions it carries. An airplane without munitions might as well not exist. For the maintenance depot it is exactly the opposite. If the munitions are not removed from the plane before it enters the depot, the whole depot has to be shut down and decontaminated. Clearly, the question of whether munitions are part of an airplane has more than one answer!

At a simpler level, the data on the web may simply be wrong. People can say whatever they want to on the web, and they do. There is no policeman. So computers on the semantic web will need access to services for establishing the reliability of data.

Will the semantic web ever work consistently?

With all these difficulties of establishing the meaning and the reliability of data, will the Semantic Web ever take off? Both Shadbolt and Hendler think it will, but that it will be rather heterogeneous.

Shadbolt suggests that the commercial imperative will ensure the start of Semantic web activity in some areas. There is a strong financial incentive for some businesses to carry out transactions and disseminate information about products automatically over the web. These businesses and their users will be early foci of activity on the semantic web.

Hendler suggests that the Semantic Web will contain many small communities of users with common interests that enable them to share ontologies. Their ontologies will often be highly technical and mission-specific.

He thinks that larger communities of users will form around shared terms. The end result will be a ramshackle system that has multitudes of partial mappings between ontologies. The system will be distributed and inconsistent, but flexible – like human knowledge.

But then Hendler would say that. He’s an optimist. You can tell it from his web page: it’s already marked up for the semantic web!

URLs

Nigel Shadbolt <http://www.ecs.soton.ac.uk/~nrs/>
James Hendler <http://www.cs.umd.edu/users/hendler/>

Semantic Web Links

<http://www.cs.umd.edu/projects/plus/SHOE/>
<http://www.w3.org/DesignIssues/Semantic.html>
<http://www.w3.org/2001/sw/>
<http://www.aktors.org/>

<http://www.itri.brighton.ac.uk/events/hendler/hendler.pdf>

Andrew Derrington is professor of Psychology at Nottingham University and a freelance journalist. His research investigates the way the brain processes visual information. He has written over 100 articles on topics in science and technology for the Financial Times.