

ICT Results ... Results that lead the way

Features

Semantics gives the web meaning – for machines

Where would we be without the web? It is such an immense and rich source of information; we feel that every answer is out there. All it takes is a bit of searching...

But internet searches are often fruitless – even Google's eight billion indexed web pages and vast store of data and documents. Text-based searches do what they say on the box: they find keywords within documents. But what kind of web search could quickly give you a list of foods triggering adverse reactions in elderly women taking medication for high blood pressure?



"The current web is a web of text and pictures," says Frank van Harmelen, a researcher in the Department of Artificial Intelligence at the Free University of Amsterdam. "Data is everywhere, but most of it is locked in inaccessible databases behind websites, locked within documents, or held within silos so it can't be linked to related data elsewhere."

What's more, computers are unable to understand the data that they find. To a computer, the number 00352 is just a series of digits. Type this into Google and the top hits are an eclectic collection of unrelated pages (usually because the number 0352 features somewhere within a code number or filename).

But a new era of the web is upon us. The 'semantic web' codes data in a way that gives meaning to words and digits in a way that computers can understand. Given just a little bit of context, applications can now recognise when 00352 is actually an international dialling prefix.

And by linking unrelated data sources, the semantic web might be able to tell you the name of the person you are calling, their email address, the cheapest carrier for calls to Luxembourg, and even link to profiles of several of your contact's friends and colleagues.

"The semantic web is a web of data," van Harmelen continues, "It realises the vision for interoperability between data sources on the web and it gives the data meaning in a way that computers can understand and reason with it."

From trumped-up to joined-up data

Tim Berners-Lee, the beknighted father of the web, famously said: "I have a dream for the web [in which computers] become capable of analysing all the data on the web – the content, links, and transactions between people and computers."

European research has long sought to materialise that dream. Since before 2000, Europe has driven research on the semantic web and taken a strong lead on the development of concepts, underlying standards and, more recently, software and services for the semantic web. But the semantic web is not just some European pipedream. Other big forces in the world are also hearing the 'semantic call'.

Indeed, in February 2008, Yahoo announced that its indexing would support the semantic web's RDF format (see below) in web-page metadata and open database sources. Yahoo undoubtedly hopes that, by using semantic datasets, its searches will become much more powerful, finding links between disparate data sources, and delivering better results in the top few search hits. Better because they will be better structured with all the relevant bits

presented in a coherent whole after having been extracted from diverse independent sources.

RDF (Resource Description Framework) is an internationally agreed model for representing the meaning of data (e.g. 'telephone number') and how it relates to (e.g. 'has the value') the data value (e.g. '00352...'). These are known as RDF triples; and a single item of data could have several triples (e.g. 'telephone number'... 'is the switchboard number of'... 'the European Commission').



Wise moves

Europe is perfectly poised to meet the growing demand for semantic technologies that Yahoo's move will undoubtedly stimulate. The European Union's investment in semantic web research has already far surpassed other regions and countries, including the USA.

The EU's Sixth Framework Programme (FP6, 2002-2006) for research has funded 17 semantic web projects and about €50m annually is allocated to continued research in this area under FP7, which runs until 2013. With some ten years of experience in this field, Europe has a firm 'first-mover' status on the semantic web and a large pool of

talent working on it.

One of Europe's biggest contributions in the early days has been its involvement in the engineering of ontologies. Ontologies are like the vocabulary of the semantic web, collections of related concepts used to assign meaning to data and describe the relationships between them.

Several earlier projects (e.g. [OntoWeb](#), [WonderWeb](#)) focused on how to build ontologies and were closely linked with international efforts to standardise the increasingly prevalent OWL ontology language.

Ontology research in Europe remains strong; ongoing projects focus on making ontologies 'user friendly', for example by developing tools (e.g. [NeOn](#)) that reduce the costly and time-consuming process of ontology construction and maintenance – ontologies evolve over time just like computer programs.

Information, extraction not abstraction

But an ontology is just the beginning because it is very time consuming and expensive to convert data manually into RDF or OWL formats. With structured information (i.e. information in databases) automatic conversion is relatively straightforward. But what about the data embedded in unstructured sources, like free text?

As the market ripens for semantic web tools and services, the rest of the world is starting to understand why this conversion makes good sense. Yahoo's announcement earlier this year and Reuters' recently launched 'Open Calais' deal with the supply-side of the semantic web are testimony to this.

Web developers can 'plug in' to Open Calais for free and use the service to extract company and financial data from documents and text, and embed the extracted information as semantically expressed data.

Open Calais and similar commercially available tools use natural language processing to make sense of words and numbers – extracting 'meaning' – and encode the information according to semantic web standards.

European research into natural language processing stems back a decade. Today, projects are building automated semantic extraction engines. [BootStrep](#), for example, is using medical dictionaries, thesaurus-like features and biological fact databases to build vast biological lexicons which, combined with natural language processing, can pull facts and semantic entities from documents.

Yet more European teams have developed technologies to analyse and mark different media formats (for example, image and audio files) automatically in order to produce semantic

metadata tags.

Unlock, define... and now reason

Unlocked databases, better-defined 'entities' and their relationships, information extracted from text and multimedia files... Is that enough for the semantic web to really meet its potential? Not really.

With all this data now on the web, computing now turns to the question of 'reasoning', or 'reasoning engines', to be precise. And European research is on this case, too.

Two new projects look set to make a global impact on the semantic web.

First, the REVERSE network has worked on a set of interoperable reasoning languages for advanced web systems and applications. These languages – the first of their kind – will be submitted to bodies, such as the World Wide Web Consortium (W3C), as the main basis for international standards.

Second, the LarKC project is looking at semantic reasoning to solve a fundamental problem of the semantic web: its size.

"Now the semantic web is taking off for real – with billions of facts available – scale is becoming a problem," explains Frank van Harmelen who is working on LarKC. "The tools we have are fine for small-scale applications, but we need large-scale infrastructure, to break away from toy applications."

LarKC will be a platform for massive, distributed, incomplete reasoning. It will achieve scalability both through its lack of completeness (it decides when it has queried 'enough' data) and its parallel processing (on clusters of high-performance computers or through a distributed network of 'home' computers).

Van Harmelen hopes that this kind of approach could get web technology to the point where, as you drive into town, an application spots a space in a nearby car park, calculates how long and what route to get there, and that no one else is closer.

We are not there yet, but as Yahoo's support for the RDF format shows, the semantic web is about to make it big. Companies will find it easier to integrate datasets and access information – internal and external – while consumers should find their web searches are more fruitful and web services more functional.

The technology behind the semantic web may remain hidden to most, but the results it achieves will not fail to impress.

Your future experience of the web may not be labelled Semantic Inside but the technology will almost certainly be there. And there is a good chance that it will say Made in Europe on the box, too.

Selected EU-funded projects and links on the semantic web

SEKT (January 2004): development and testing of tools for semi-automated ontology learning and metadata extraction, but with built-in mechanisms to cope with inconsistent data.

aceMedia (January 2004): image recognition and knowledge analysis tools to supply metadata annotations on static and moving images, and on specific parts of those images. It can be used to identify and annotate different sections of a photograph, for example differentiating beach, sea and sky.

SIMAC (January 2004): semantic processing of audio content, based on their actual musical properties – rhythm, timbre, harmony, structure and instrumentation. It allows comparisons between songs to be made. Listeners can find little-known tracks that suit their tastes but may otherwise go unnoticed.

NeOn (March 2006): developing tools that make it easier for people to construct and maintain ontologies, lowering the barriers to wider development of semantic web applications.

VIDI-VIDEO (February 2007): improving semantic annotation of videos using a system that detects a large number of weak correlations between many different aspects of a video's

contents.

ACTIVE (March 2008): increasing the productivity of knowledge workers, based on Web 2.0 (social networking and collaboration) and semantic technologies. Trial applications to help businesses convert tacit and unshared information – the "hidden intelligence" of enterprises – into transferable, interoperable and actionable knowledge.

OKKAM (January 2008): developing the decentralised infrastructure and underlying technology making it possible to allocate unique identifiers to every semantic web entity and removing ambiguity and incompatibilities between unlinked semantic datasets.

[BBC interview with Tim Berners-Lee on the future of the internet](#)

INFORMATION :

DATE : 16 Jul 2008

TECHNOLOGY AREA :

[Internet protocol](#)



[Language/speech](#)

MARKET APPLICATION:

[Exchange of information](#)



[Publishing/media](#)



RECENT ARTICLES :

- [Special feature: SMEs make Europe's innovation clock tick](#)
- [Special feature: Un-masking a faster solution for chip-making](#)
- [Collective solution to accessing the internet via satellite](#)
- [Pooled data towards cervical cancer cure](#)
- [Semantic desktop paves the way for the semantic web](#)

