It's All Semantics

Wendy Hall & Kieron O'Hara

Intelligence, Agents, Multimedia Group
School of Electronics and Computer Science
University of Southampton
Highfield
Southampton SO17 1BJ
United Kingdom
{wh,kmo}@ecs.soton.ac.uk

The Parliament Magazine's European Research and Innovation Review, 7 (November 2008), 31-32

The World Wide Web is one of the most transformative technologies of our time, a giant network of documents and resources allowing access to undreamt-of quantities of information. But as its inventor Tim Berners-Lee argued as long ago as 1994, its potential will be achieved when the links are between not documents, but *data* – enabling information to be gathered from distributed sources, in heterogeneous formats, and processed by machines. This vision was originally dubbed the Semantic Web but is perhaps more accurately referred to as the *Linked Data Web*. This is the Web's future, allowing *serendipitous reuse* of information to emerge through its sheer abundance, often in new and unexpected contexts. Early adopters include e-science and e-social science, which depend on the integration and automatic interrogation of large quantities of distributed data.

Box: The Linked Data Web

The Linked Data Web, also referred to as the Semantic Web, is a Web of actionable, linked data. Data become actionable by being interpreted through semantic theories (*ontologies*) which attach significance to the data's symbols. A basic symbol, '6', say, could then be understood as a temperature, an age or a distance in parsecs. Through those semantics we achieve interoperability between systems. Building the Linked Data Web is analogous to making the World Wide Web a giant database, allowing us to query distributed and heterogeneous sources.

But it's not just big science. Another central development is the creation of large amounts of information by individuals, using Web 2.0, or social networks, or blogs, while Moore's Law makes it possible to store it all. Information overload is a problem, yet also an opportunity to provide rich content to support personal or community memory. The investigation of this new frontier in personal computing and information management has been recognised as a grand challenge by the United Kingdom Computing Research Committee, under the banner *Memories for Life* (M4L), and the research goal is being widely adopted across the world. Creating tools to manage this personal information will need interdisciplinary work by computer scientists, psychologists, sociologists, HCI researchers, information managers and more – and will also need to cross digital divides of age, education and gender. We expect to see major developments in terms of applications and devices in the near future.

As we build more complex computational artefacts and infrastructures, we observe large-scale behaviour emerging from small-scale and local regularity. We need engineering methods to ensure that our structures conform to reliable and repeatable

design requirements, and scientific analysis to understand and predict the behaviours that result. Understanding the connections between the micro and the macro demands a third type of vital research, which we call *Web Science*.

Web Science is not just modelling the Web. It includes engineering infrastructure protocols using tools from many disciplines which may involve radical thinking about technology and society, but must also respect the Web's invariants: decentralisation to avoid bottlenecks and allow scalability; serendipitous reuse of information; fairness, openness and trust. In this way the Web will remain a technology to enhance human society, and support human aspiration – at a global scale.

The Linked Data Web relies on a series of formalisms and technologies. For instance, *Universal Resource Identifiers* (URIs) provide a global naming convention for resources, allowing anyone to link to them, or retrieve a representation of them. The *Resource Description Framework* (RDF) is a knowledge representation language based on XML that assigns URIs to the subjects, predicates and objects that it links. *OWL* is a series of languages for expressing ontologies with different trade-offs between computability and expressive power. *SPARQL* (pronounced 'Sparkle') is a query language for RDF to allow interrogation of the linked data Web.

Adopting these technologies and more can appear daunting, but at Southampton University we have shown that the technology can be transferred from the lab to real-world corporate and governmental environments without large investments in ascending a learning curve. Semantic technologies provide quick wins for users, including the ability to share information across large organisations, to integrate it with information brought in from outside, and to create useful applications rapidly.

The Linked Data Web is a work in progress. To make the vision a reality, we need: tools and interfaces to increase usability for all, some of which will come from the M4L community; Web Science to understand the feedback loops between the Web and society; methods for information accountability to promote trust and preserve privacy; tools and techniques for ontology development, maintenance and mapping; and technologies such as Natural Language Processing will be required to make the legacy free text Web machine readable. But most of all we need to publish more data, preferably using or converting to RDF, to make it available to wider communities.