Use of the Semantic Web in e-Research

Kieron O'Hara*, Tim Berners-Lee, Wendy Hall & Nigel Shadbolt

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

Version of chapter to appear in Dutton, W. H., and Jeffreys, P. W. (eds), *World Wide Research: Reshaping the Sciences and Humanities*. Cambridge, MA: The MIT Press, 2010

A New Way of Finding Information: Basic Technologies of the Semantic Web

A vital factor in the way the World Wide Web has revolutionized research has been its radical decentralization: any page can link to any other. This decentralization is scalable and removes bottlenecks in supply. Navigation can be via associational links, maintaining relevance, or key-word search, which allows the user a measure of control that makes a suitably connected computer a virtual, near-universal library.

Yet automation of research has farther to go. Information embedded in a document may still not be easy to find, a problem exacerbated when it is distributed over several documents. For these reasons, research continues to evolve the Web from a "Web of Documents" to a "Web of Data." "Semantic Web" is the name given to a conception of a Web of linked data, underpinned by a series of technologies and standards developed under the auspices of the World Wide Web Consortium[1] since the late 1990s. Here we briefly summarize four key components of the Semantic Web (for more detail on its various layers, see, for example, Shadbolt, Hall, and Berners-Lee 2006).

1. The basis for a Web of linked data is the system of uniform resource identifiers (URIs).[2] The URIs allow widespread and consistent reference by providing a global convention for naming resources, interpreted consistently across contexts. Associating a URI with a resource allows anyone to refer to it; retrieve a representation of it if it is a document; retrieve a document about it if it is not a document; or—crucially—to link to it.

2. The Resource Description Framework (RDF)[3] is a simple knowledge representation language for the Semantic Web based on the "subject-predicate-object" form. A statement in RDF, called a "triple," links two objects (individuals, kinds of things, attribute values) and a property, relation, or two-placed predicate. Each member of the triple is assigned a specific URI. Using RDF therefore involves the use

[1] For progress on the consortium's work, see http://www.w3.org/2001/sw/.
[2] See http://tools.ietf.org/html/rfc3986.
[3] See http://www.w3.org/TR/rdf-concepts/.

of URIs to ground reference to objects and relations, which opens the door to automatic processing not only of documents, as in the current Web, but also directly of data. Linked RDF statements form a directed, labeled graphical representation.

3. "Ontologies" are common conceptualizations pinning down the vocabularies of domains. They support interoperability, information integration, and knowledge sharing by aligning vocabularies and underpinning translations between terms. We can distinguish two types of ontology. *Deep* ontologies are detailed presentations of the scientific vocabulary where great effort is put into the conceptualization of the domain and the maintenance of the ontology relative to ongoing scientific discovery; this type of ontology is often encountered in scientific or engineering contexts. *Shallow* ontologies, by contrast, are composed of a relatively small number of commonly used terms describing basic relations that tend not to change very much in the short to medium term; such shallow ontologies, though relatively simple, can be used to organize very large quantities of data.

4. Simple Protocol and RDF Query Language (SPARQL)[4] is a protocol and query language designed for Semantic Web resources. In particular, it can be used to express queries across diverse data sources, if the data is either stored in RDF or can be viewed as an RDF graph via middleware. SPARQL supports a number of querying functions, including the querying of an RDF graph for required or optional graph patterns, as well as for conjunctions and disjunctions of patterns.

These four technologies allow a research community to develop heterogeneous data repositories as a common resource—grounded out by URIs and linked and integrated by ontologies. Ontologies themselves have become important resources in science and e-science (e.g., Shadbolt, Hall, and Berners-Lee 2006:96), which is particularly evident in interdisciplinary studies such as climate change or epidemiology, where several different sets of terms are employed, and data stores are particularly diverse and large scale.

The Semantic Web's Value to Researchers

Enthusiastic early adopters of the Semantic Web approach were typically communities needing to integrate and share information. Such communities have a degree of cohesion and a perceived need for shared semantics. For example, researchers often need to query large numbers of databases. Without Semantic Web technology, this task would require either complex scripts to overcome incompatibilities or a laborious manual process of cutting and pasting between Web interfaces. Semantic Web technologies, including ontologies and annotation, have been shown to be very useful in preserving information quality (Preece et al. 2006), and SPARQL provides a network protocol that allows effective querying.

One fruitful approach is the idea of a "Semantic Grid," whereby the data, computing resources, and services characteristic of the Grid computing model are given semantics using Semantic Web ideas and technologies (De Roure and Hendler 2004). The myGrid project,[5] for example, supports data-intensive querying in the life sciences, linking together diverse resources using Web service protocols and providing support for managing scientific work flow, sharing and reusing information, and understanding provenance.

[4] See http://www.w3.org/TR/rdf-sparql-query/.
[5] See http://www.mygrid.org.uk/.

Such "*in silico* experimentation" using a computer simulation has also been used outside the life sciences. In chemistry, the synthesis of new compounds requires the assembly, integration, and querying of large quantities of primary data. The CombeChem project[6] has employed a similar large, service-based infrastructure to create a knowledge-sharing environment, using pervasive devices to capture live metadata in the laboratory and linking data using shared URIs. This approach allows knowledge sharing across data sets created by different stakeholders, with provenance traceable back to the source.

Semantic Web technologies have also been used to support research in social science. The UK's National Centre for e–Social Science has been set up to apply e-science techniques to social science data, both quantitative and qualitative (Proctor, Batty, Birkin, et al. 2006). For instance, the PolicyGrid project[7] brought social scientists and Semantic Web technologists together to create a metadata infrastructure to support annotation, data sharing, and social simulation.

The sharing of raw data in social science raises privacy concerns. Releasing someone's zip or postal code is not a problem, but releasing his or her medical history is, and Semantic Web technology allows this distinction to be made. Contrast this capability with the traditional document Web, wherein a document containing both pieces of information has to be either released or withheld as a whole or laboriously anonymized. Research on a "policy-aware Web" (Weitzner, Hendler, Berners-Lee, et al. 2005) will enable more automated reasoning to be carried out on privacy policies in the Web's open environment.

Building on the Semantic Web Approach

Semantic technologies have proved important in automating scientific and social scientific research, enabling it to cope with the much larger quantities of data available through advanced computing techniques and better-founded datasharing practices. The automation of information processing will have many effects on research and other aspects of Web use. These effects are hard to predict in detail without study of the two-way relation between microlevel protocol development and macrolevel social change. The Web Science Trust[8] was set up to explore precisely this interrelationship in order to ensure that developments such as the Semantic Web have benign effects on society, knowledge sharing, and the performance of scientific research (Berners-Lee, Hall, Hendler, et al. 2006). The future of science *on* the Web will depend on the development of the science *of* the Web.

References

Berners-Lee, T., W. Hall, J. A. Hendler, K. O'Hara, N. Shadbolt, and D. J. Weitzner. 2006. "Creating a science of the Web." *Science* 313(5788):769–771. Available at http://eprints.ecs.soton.ac.uk/12615/.

De Roure, D., and J. A. Hendler. 2004. "E-science: The Grid and the Semantic Web." *IEEE Intelligent Systems* 19 (1):65–71.

Preece, A., B. Jin, E. Pignotti, P. Missier, S. Embury, D. Stead, and A. Brown. 2006. "Managing information quality in e-science using Semantic Web technology." In

[6] See http://www.combechem.org/.
[7] See http://www.policygrid.org/.
[8] See http://webscience.org/.

Proceedings of 3rd European Semantic Web Conference. Berlin: Springer, 472–486. Available at: http://www.csd.abdn.ac.uk/~apreece/qurator/resources/qurator_eswc2006.pdf.

Proctor, R., M. Batty, M. Birkin, R. Crouchley, W. H. Dutton, P. Edwards, M. Fraser, P. Halfpenny, Y. Lin, and T. Rodden. 2006. "The National Centre for e–Social Science." In Proceedings of the 2006 e-Science All- Hands Meeting, 542–549. Available at: http://www.allhands.org.uk/2006/proceedings/proceedings/proceedings.pdf.

Shadbolt, N., W. Hall, and T. Berners-Lee. 2006. "The Semantic Web revisited." *IEEE Intelligent Systems* 21 (3):96–101. Available at http://eprints.ecs.soton.ac.uk/12614/.

Weitzner, D. J., J. Hendler, T. Berners-Lee, and D. Connolly. 2005. "Creating a policy-aware Web: Discretionary, rule-based access for the World Wide Web." In *Web and information security*, ed. E. Ferrari and B. Thuraisingham. Hershey, PA: Idea Group. Available at: http://www.w3.org/2004/09/Policy-Aware-Web-acl.pdf.