Using the Semantic Web for e-Science: Inspiration, Incubation, Irritation (Extended Abstract)

Carole Goble

School of Computer Science,
The University of Manchester, Manchester, M13 9PL UK
carole@cs.man.ac.uk

We are familiar with the idea of e-Commerce - the electronic trading between consumers and suppliers. In recent years there has been a commensurate paradigm shift in the way that science is conducted. e-Science is science performed through distributed global collaborations between scientists and their resources enabled by electronic means, in order to solve scientific problems. No one scientific laboratory has the resources or tools, the raw data or derived understanding or the expertise to harness the knowledge available to a scientific community. Real progress depends on pooling know-how and results. It depends on collaboration and making connections between ideas, people, and data. It depends on finding and interpreting results and knowledge generated by scientific colleagues you do not know and who do not know you, to be analysed in ways they did not anticipate, to generate new hypotheses to be pooled in their turn. The importance of e-Science has been highlighted in the UK, for example, by an investment of over £240 million pounds over the past five years to specifically address the research and development issues that have to be tacked to develop a sustainable and effective e-Science e-Infrastructure.

The Web has served scientists well. Many data sets and tools are published and accessed using web protocols and web browsers. Sharing data repositories and tool libraries has become straightforward. Widespread collaboration is possible by publishing a simple web page. However, standard web technology is now straining to meet the needs of scientists. The scale of data is one problem thanks to high throughput scientific methods – more data is about to be generated in the next five years than has been generated by mankind hitherto fore. Another problem is that communities can no longer be isolated silos – chemists must share with molecular biologists; earth scientists collaborate with physicists and so on. Yet a Web-based distributed information infrastructure is still a place where the scientists manually: *search* the web for content; *interpret and process* content by reading it and interacting with web pages; *infer* cross-links between information; *integrate* content from multiple resources and *consolidate* the heterogeneous information, while preserving the *understanding* of its *context*. Sound familiar?

It would seem self-evident that the Semantic Web should be able to make a major contribution to the fabric of e-Science [1,2]. The first W3C Semantic Web for Life Science Workshop in 2004 attracted over 100 participants with representation from all the major pharmaceutical and drug discovery players, and leading scientists (http://www.w3.org/2004/07/swls-ws.html). Scientific communities are ideal **incubators** for the Semantic Web: they are knowledge driven, fragmented, and have

Y. Gil et al. (Eds.): ISWC 2005, LNCS 3729, pp. 1 – 3, 2005.

valuable knowledge assets whose contents need to be combined and used by many applications. The content is diverse, being structured (databases, electronic lab books), semi-structured (papers, spreadsheets) and unstructured (presentations, Web blogs, images). The scale necessitates that the processing be done automatically. There are many suppliers and consumers of knowledge and a loose-coupling between suppliers and consumers – information is used in unanticipated ways by knowledge workers unknown to those who deposited it. People naturally form communities of practice, and there is a culture of sharing and knowledge curation. For a Semantic Web to flourish, the communities it would serve needs to be willing to create and maintain the semantic content. Most scientific communities embrace ontologies. The Life Science world, for example, has the desire for collaboration, a culture of annotation, and service providers that might be persuaded to generate RDF or at least annotated XML. A semantic web is expensive to set up and maintain, and thus is only likely to work for communities where the added value is worthwhile and an "open source data" philosophy prevails.

The Scientific Community has been **inspired** by the results of the Semantic Web initiative already. The inferencing capabilities of OWL have been shown to aid the building of large and sophisticated ontologies such as The Gene Ontology (http://www.geneontology.org) and BioPAX (http://www.biopax.org/). The self-describing nature of RDF and OWL models enables flexible descriptions for data collections, suiting those whose schemas may evolve and change, or whose data types are hard to fix, like knowledge bases of scientific hypotheses, provenance records of *in silico* experiments or publication collections [3]. These are examples where the semantic technologies have been adopted by scientific application. Genuine "Semantic WEB" examples, with the emphasis on Web, are also starting to appear. SciFOAF builds a FOAF community mined from the analysis of authors and publications over PubMed (http://www.urbigene.com/foaf/). Scientific publishers like the Institute of Physics (http://syndication.iop.org/), publish RSS feeds in RDF using standard RSS, Dublin Core and PRISM RDF vocabularies. The Uniprot protein sequence database has an experimental publication of results in RDF (http://www.isb-sib.ch/~ejain/rdf/). YeastHub [4] converts the outputs of a variety of databases into RDF and combines them in a warehouse built over a native RDF data store. BioDASH (http://www.w3.org/2005/04/swls/BioDash/Demo/) is an experimental Drug Development Dashboard that uses RDF and OWL to associate disease, compounds, drug progression stages, molecular biology, and pathway knowledge for a team of users. Correspondences are not necessarily obvious to detect, requiring specific rules. Semantic technologies are being used to assist in the configuration and operation of e-Science middleware such as the Grid [6]. These examples should be an inspiration to the Semantic Web community.

However, there is also **irritation**. There are some problems with the expressivity of OWL for Life Science, Chemical and Clinical ontologies. The mechanisms for trust, security, and context are important for intellectual property, provenance tracing, accountability and security, as well as untangling contradictions or weighting support for an assertion; yet these are immature or missing. Performance over medium-large RDF datasets is disappointing – the CombeChem combinatorial chemistry project generated 80 million triples trivially and broke most of the triple stores it tried (http://www.combechem.org). There is poor support for grouping RDF statements,

yet this is fundamental. Semantic web purists claim that the Life Science Identifier [5], for example, is unnecessary, although these critics seem not to have actually developed any applications for life scientists. Sometimes there is irritation that the wrong emphasis is being placed on what is important and what is not by the technologists, leading to a communication failure between those for whom the Semantic Web is a means to an end and those for whom it *is* the end [7].

The Web was developed to serve a highly motivated community with an application and a generous spirit–High Energy Physics. The Semantic Web would also benefit from the nursery of e-Science. In my talk I explore this opportunity, the mutual benefits, give some pioneering examples, and highlight some current problems and concerns: inspiration, incubation, and irritation.

References

[1] James Hendler *Science and the Semantic Web* Science 299: 520-521, 2003

[2] Eric Neumann *A Life Science Semantic Web: Are We There Yet?* Sci. STKE, Vol. 2005, Issue 283, 10 May 2005

[3] Jun Zhao, Chris Wroe, Carole Goble, Robert Stevens, Dennis Quan, Mark Greenwood, *Using Semantic Web Technologies for Representing e-Science Provenance* in Proc 3rd International Semantic Web Conference ISWC2004, Hiroshima, Japan, 9-11 Nov 2004, Springer LNCS 3298

[4] Cheung K.H., Yip K.Y., Smith A., deKnikker R., Masiar A., Gerstein M. *YeastHub: a semantic web use case for integrating data in the life sciences domain* (2005) *Bioinformatics* 21 Suppl 1: i85-i96.

[5] Clark T., Martin S., Liefeld T. *Globally Distributed Object Identification for Biological Knowledgebases* Briefings in Bioinformatics 5.1:59-70, March 1, 2004.

[6] Goble CA, De Roure D, Shadbolt NR and Fernandes AAA *Enhancing Services and Applications with Knowledge and Semantics* in The Grid: Blueprint for a New Computing Infrastructure Second Edition (eds. I Foster and C Kesselman), Morgan Kaufman 2003

[7] Phillip Lord, Sean Bechhofer, Mark Wilkinson, Gary Schiltz, Damian Gessler, Carole Goble, Lincoln Stein, Duncan Hull. *Applying semantic web services to bioinformatics: Experiences gained, lessons learnt.* in Proc 3rd International Semantic Web Conference ISWC2004, Hiroshima, Japan, 9-11 Nov 2004 , Springer LNCS 3298