Pacific Symposium on Biocomputing 2006: World Scientific; 2006. p. 112-115.

SEMANTIC WEBS FOR LIFE SCIENCES

ROBERT STEVENS

School of Computer Science, University of Manchester
Oxford Road, Manchester, M13 9PL, UK
E-mail: Robert.stevens@manchester.ac.uk

OLIVIER BODENREIDER

U.S. National Library of Medicine
8600 Rockville Pike, MS 43, Bethesda, Maryland, 20894, USA
E-mail: olivier@nlm.nih.gov

YVES A. LUSSIER

Departments of Biomedical Informatics and Medicine
Columbia University, New York, NY 10032, USA
E-mail: yves.lussier@dbmi.columbia.edu

The Semantic Web is a vision for the next generation of the Web [1]. The Web is a huge interlinked information resource, but is largely restricted to human use because the information is represented only in natural language. The goal of the Semantic Web is to make these data – the facts on the Web – amenable to computational processing. To date, the Semantic Web has largely been pushed by technology development, but the life sciences are seen to be a huge area for potential application development. Indeed, a recent workshop on this subject saw over 100 attendees, indicating a great interest in the community[1].

The Semantic Web's broad goal parallels that of many bioinformaticians. There are vast quantities of biological data and associated annotations, or knowledge, now available on the Web. These resources are highly distributed and heterogeneous. This heterogeneity exists at many levels, the most pernicious of which are the semantic heterogeneities in the schema and the values placed in those schema. Semantic Web technologies and the vision itself offer a solution to this long-standing problem in creating an integrated view of bioinformatics.

Lincoln Stein describes this situation as being akin to the rival city states in medieval Italy and talks of the need to create a "bioinformatics nation" [6]. This vision is at one level of heterogeneity – the programmatic access to bioinformatics resources. Stein describes the use of Web Services, a Semantic Web technology, to provide a common form of access to distributed resources with heterogeneous

[1] http://www.w3.org/2004/07/swls-ws.html

platforms and access paradigms. We already see well over a thousand Web Services offering access to bioinformatics resources[2] not seen before in bioinformatics.

Semantic Web technology also offers solutions for the problems of semantic heterogeneity and these technologies have a growing influence. The aim of the Semantic Web is to make facts amenable to machine processing. The Resource Description Framework[3] (RDF) provides a common data model of triples for this purpose. An RDF triple, a subject, predicate (verb) and object, enables any statement to be represented in a simple, flexible, common framework.

Each part of a triple names a resource using either a Uniform Resource Identifier (URI) or a literal. As many resources are transformed to this data model, the common naming scheme will mean that facts can be aggregated, forming a vast graph of descriptions of resources [8]. The Life Science Identifier (LSID) is a form of URI that can be used to uniquely identify and version bioinformatics entries [3]. *Uniprot* is already available in RDF[4] and shows this aggregation happening using LSID. Similarly, *YeastHub* [2] is a system that has transformed many yeast resources into RDF and allows querying, using an RDF query language, to provide access to these aggregated data.

One advantage of the RDF model is the open world assumption. In an open world, only that which is explicitly stated is known – we cannot assume that something does not exist simply because it has not been stated. This means that new statements can be added without fear of breaking the data model, which happens all too easily with existing schema mechanisms such as XML schema [8].

Even with this flexible model, the description of the resources themselves with relationships (predicates) and the objects that provide values for these descriptions are still highly heterogeneous. In RDF, the collection of names formed by the URI provides a vocabulary. True semantic integration requires a common, shared vocabulary. This is the role of ontologies and Semantic Web technology provides languages for this purpose. RDF Schema is an RDF vocabulary for ontologies. It enables classes and their relationships to be defined and used in an RDF graph. The Web Ontology Language[5] (OWL) offers a variety of dialects for building and maintaining ontologies, with strict and precise semantics not offered by RDFS. OWL ontologies can be delivered as RDFS for Semantic Web use.

Again, we see OWL and RDFS being used within the life sciences. The *Gene Ontology*™ [4] is available in RDFS and can be used to annotate, for instance, the RDF version of *Uniprot*. OWL is used in the *BioPAX* ontology [5], which is used to exchange data between pathway resources. OWL is used by the MGED Society to provide an ontology for marking up microarray experiments [7].

---

[2] See, for example, http://www.mygrid.org.uk.
[3] http://www.w3.org/TR/2003/PR-rdf-concepts-20031215/
[4] http://www.isb-sib.ch/~ejain/rdf/
[5] http://www.w3.org/TR/owl-ref/

There are an increasing number of bioinformatics applications using Semantic Web technologies. There are, however, few real Semantic Webs of Life Sciences, where large quantities of diverse data are aggregated with RDF, described with RDFS vocabularies and then exposed for querying and automatic reasoning. *YeastHub* and *BioDash*, working over *BioPAX* data, come the closest to this vision, as will be seen in the Semantic Webs for Life Sciences session.

This session reflects the early adoption of the Semantic Web by the bioinformatics community. While most papers still focus on foundational issues such as namespaces, ontology creation, mapping and adaptation to Semantic Web formalisms, some contributions present pioneering applications implementing the vision of the Semantic Web.

For instance, two papers address human computer interactions and demonstrate how to use Semantic Web technologies to facilitate the access to otherwise heterogeneous semantics of human interfaces of computerized bioinformatics resources by scientists: *BioDash* from Neuman and Quan provides an integrated web-based dashboard for drug development, and *BioGuide* from Cohen-Bulakia et al. is a user-centric framework to help scientists to choose tools according to their preferences and strategies.

Beyond the interface, organizing heterogeneous information sources can also be approached with Semantic Web technologies. Indeed, as shown by the work of Mukherjea and Sahay, the semantic relationships presented by different applications over the Web can be mined through search engines using Semantic Web technologies and these relations can be elicited explicitly. At a more automated level of communications and automated machine processing, Yip et al. present a Semantic Web approach, *SemBiosphere*, to build a matchmaking system that automatically provides recommendations to users about microarray clustering algorithms by reasoning over the Semantic Web service descriptions of these methods.

Another group of papers focus on ontologies, a necessary technology for Semantic Web development. Zhang et al. verified the consistency of the *Foundational Model of Anatomy* by first transforming its representation in OWL and then using the best "reasoner" to identify unclassifiable classes. Good et al. propose an important and necessary improvement over the development and maintenance of ontologies: a protocol to attain affordability. Indeed, affordability issues remain one of the big challenges for sustainability of the Semantic Web. Kushida et al. describe the design of a new biomedical ontology for annotating biological pathways component. Finally, Kazic reviews the fundamental assumptions of current Semantic Web technologies and proceeds systematically to demonstrate their potential and structural limitations.

As the field matures and as a critical mass of Semantic Web resources (e.g., ontologies, Web Services) becomes available, the number of Semantic Web applications is expected to grow dramatically in the next few years, illustrating the

fact that “the combined effect of global naming, universal data structure and open world assumption is that resources exist independently but can be readily linked with little, if any, precoordination.” [8].

References

1. Berners-Lee, T., J. Hendler, et al. (2001). "The Semantic Web." Scientific American **284**(5): 34-43.
2. Cheung, K. H., K. Y. Yip, et al. (2005). "YeastHub: a semantic web use case for integrating data in the life sciences domain." Bioinformatics **21**(1): i85-i96.
3. Clark, T. and S. M. Liefeld (2004). "Globally Distributed Object Identification for Biological Knowledgebases." Briefings in Bioinformatics **5**(1): 59-70.
4. Gene Ontology Consortium (2000). "Gene Ontology: Tool for the Unification of Biology." Nature Genetics 2**5**(1): 25-29.
5. Luciano, J. (2005). "PAX of mind for pathway researchers." Drug Discov Today **10**: 937-942.
6. Stein, L. (2002). "Creating a bioinformatics nation." Nature **417**(9): 119-120.
7. Stoeckert, C. J. and H. Parkinson (2003). "The MGED ontology: a framework for describing functional genomics experiments." Comparative and Functional Genomics **4**(1): 127-132.
8. Wang, X., R. Gorlitsky, et al. (2005). "From XML to RDF: how semantic web technologies will change the design of 'omic' standards." Nature Biotechnology **23**(9): 1099-1103.