A Life Science Semantic Web: Are We There Yet?

Eric Neumann

(Published 10 May 2005)

If you've ever driven with your kids for a long distance, you've probably heard them ask "Are we there yet?" Although it sounds like a silly question, it implies that they are expecting to be at the destination already. In this era of "-omics," many of us are wondering if we really are "there" with the necessary tools and standards for data integration and interoperability, or are we just fooling ourselves?

Let us first consider some basics. Scientific research relies on researchers sharing information and knowledge in meaningful ways, not simply on the transmission of disassociated chunks of data. Although proper computational methods and effective data-exchange protocols are key elements of modern scientific practice, they themselves cannot advance the scientific process. Critical interpretation of experimentally derived information and the consolidation of knowledge from multiple lines of reasoning are central to building and testing new hypotheses and are the essential elements that drive the scientific process of debate and rebuttal. Yet we still need information systems and tools to better support this process of converting and analyzing data, aggregating ideas, and reasoning about insights. In a recent report on drug development (*1*), the U.S. Food and Drug Administration has acknowledged the need for "a knowledge base built not just on ideas from biomedical research, but on reliable insights into the pathway to patients." Much still needs to be done to meet these goals.

A major concern in the life sciences is how to deal with the exponential growth in the amount of data, as well as the increasing variety of its forms. By some accounts, the amount of data captured yearly is approaching about 1 exabyte (10^{18} bytes). As a standard for comparison, the number of words ever spoken by humans totals about 12 exabytes (*2*). Yet the essential problem is not how to store large amounts of data (Moore's law comes to the rescue) but how best to distill and manage valuable insights from them through the application of analytical and data-mining methods. The derived information is used for higher-level reasoning and decision-making but is often represented simply as text (and usually in PowerPoint slides). However, although words can be indexed, meaning is not as easily cataloged, thereby limiting how such information can be searched and retrieved. In addition, words often have multiple meanings (polysemy) depending on the context and the audience, requiring further clarification. To overcome these limitations, contextual information must also be encoded in formal ways to allow both evidence and interpretation to be explicitly linked to each other.

Sanofi-Aventis Pharmaceuticals, 1041 Route 202-206 Bridgewater, NJ 08807, USA. E-mail: Eric.Neumann@sanofi-aventis.com

How can we ever hope to attain this? Consider how information is organized, not just within one database system but across all information resources that are accessible over an internet or intranet of connected computers. Although we use the Web for many purposes, a human must make the decision about where to navigate and how to interpret what is on a Web page or in a data table. HTML links do not specify what kind of thing they point to, they simply point. Currently, the Web is not sufficiently interpretable to allow the automatic identification of content within documents describing, for example, gene-specific information or whether a disease is known to be associated with a particular gene. Humans must first interpret the page by reading it, and then by either cutting and pasting the information or by writing short programs, they must "scrape" it from targeted pages to be indexed elsewhere. Considering all the computing power available today, it's perplexing that users still end up having to do a lot of manual work and that the meaning of text is still not available within an automated framework.

The Semantic Web (SW) (www.w3.org/2001/sw/) is a model for the Web proposed by Tim Berners-Lee, director of the World Wide Web Consortium (W3C) (www.w3.org), to create a universal mechanism for information exchange by giving meaning (that is, semantics), in a machine-interpretable way, to the content of documents and data on the Web (*3*, *4*). The W3C is helping direct SW activities through the definition of standards, markup languages, and key applications. The keystone standards are the Resource Description Framework (RDF) for describing objects and relations between them in an XML syntactical form, and Web Ontology Language (OWL, based on RDF) for specifying the ontologies (semantic systems of concepts and relations). OWL ontologies are used to define the types of objects and how they can relate to one another within an RDF document. These two Web-based language systems allow one to define multiple sets of ontologies and data models that can work in combination throughout internets and intranets, to support true information integration.

The key ingredient for these systems is the concept of a Universal Resource Identifier (URI) for each entity referenced from a document (aside from literal strings and numbers). For example, a gene would have a URI specifying it (subject) as a unique entity and where it could be found (similar to the URLs your browser uses, but with types), as would each element in a list of properties and relations for that gene (predicate-object). Together the URIs can form a system of subject-verb-object statements, or triples, that define relations between things. Many applications of SW in the life sciences have been proposed (*5*). In molecular biology, one could state, for example, that <gene A> <has product> <protein B>; or in proposing disease therapies, that <cancer ALL> <can be treated with> <drug Z> .

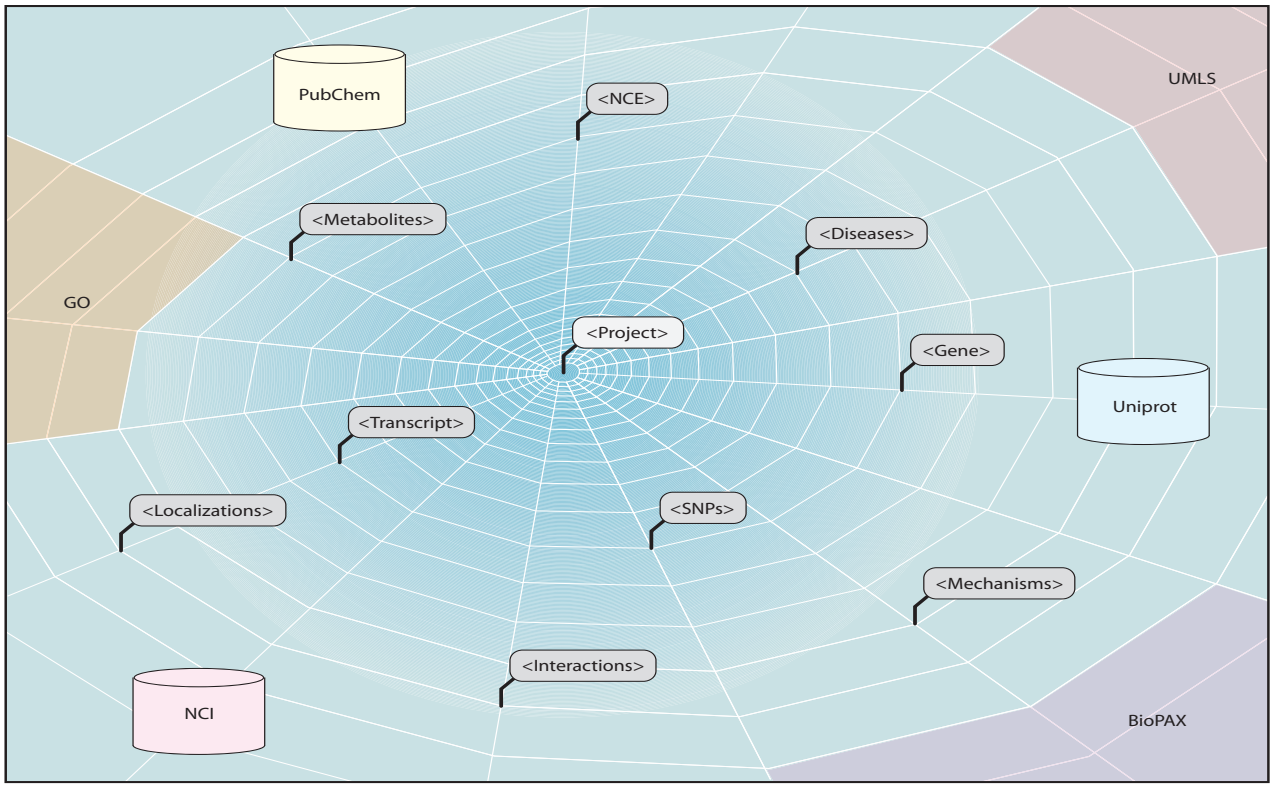

Fig. 1. Topical areas and data sources within the life sciences that are accessible to a Semantic Web.

Because URIs are unique, the ensemble of related statements forms a network of relations (specifically, a directed graph of object-relationship-object triples) that a computer program can automatically parse and understand. URIs can have different name spaces, each prefixed by a colon to the resource or relation (myDefinition:target versus yourDefinition:target), making each one globally unique so that polysemy can be handled. A special case of URI is the Universal Named Resource (URN), of which the Life Science Identifier (LSID) is a proposed life sciences data identifier standard (*6*). LSIDs have the syntactical form <urn:lsid:authority:database: identifier[:versions]>, and represent unique, immutable data objects with links to metadata and dynamic relations that are created and versioned by different authorities. Documents referring to the same LSID share a common URI. Consequently, any application with knowledge of one document can find the other document by simply "walking along the graph" across shared LSIDs, thereby aggregating all data and documents, a process that today is being reinvented thousands of times by software developers in nonstandard ways.

A semantic model can have a profound affect on how research data can be organized in meaningful bundles. Relations that exist in the minds of scientists can be explicitly defined and used to aggregate genomic, proteomic, cellular, physiological, and chemical data, even if these data exist in different databases with different schemas. Ontologies defined with the new OWL standard can automatically take advantage of this aggregating property today. The emerging BioPAX ontology for exchanging molecular pathways (www.biopax.org) is based on OWL and can use this feature to convert pathway information about molecules and reactions to a common format. Consequently, any pathway database using OWL standards (BioCYC and BIND are planning to use the BioPAX-OWL standard, and GeneOntology and UMLS will be available in OWL soon, see Fig. 1) can be accessed and its results returned in RDF using the BioPAX ontology.

Semantic definitions serve as "smart glue" to specify which objects are related to others, and the established relations can be managed in a separate system outside of the current databases. In drug development, semantics must be specified for both the domain (biology, chemistry, pharmacology, or clinical) and the business process (therapeutic strategy, compound progression, or New Drug Application submission). Because many such processes are confined to different stages of drug development and in different parts of the organization (sometimes referred to as silos), the SW model may offer a practical solution to bridge the different areas of basic research, drug development, and clinical trials. Researchers will be able to annotate findings and decisions in ways that can be understood by both humans and machines. This will be necessary in order to derive knowledge from vast numbers of analyses and to propose strategies for how to advance projects. By applying inference rules to large bodies of aggregated facts and hypotheses, new knowledge can be produced seamlessly and accountably.

The SW will not only improve server-side applications (for example, Web services and database systems) but the client side as well. This new system will usher in a new generation of browsers that not only render the semantic documents visually but aggregate select information referenced by documents and data objects. Automated rules can then be applied to filter or create new relations for the aggregated information that is to be viewed, so that users are not overwhelmed by too much data. The browser component that makes this possible is an intelligent information organizer and viewer called a "semantic lens"

that is defined to identify specific meaning within a particular chunk of information. The semantic lens contains user interface information as well as a collection of logic rules that can be applied to specific semantically defined information to pull together complex information from different sources and render the ensemble in meaningful ways. For instance, consider the following scenario:

A scientist searches for a protein in a molecular pathways database. Once found, it is returned along with a description of its properties, a result similar to what is currently returned from the UniProt database. In addition, the semantic lens used by this scientist's browser application requests pathway information for up to five reactions away from the query point. The rule selects only those reactions in a signal transduction pathway, so no metabolic reactions are rendered in the view. The lens then directs the browser to use a graph view component to show the pathway content as a graph and the protein information in a lower frame. The scientist annotates this graph directly, noting any protein features, and shares these insights with collaborators. All of the information captured is retrievable from any of the individual molecular components and in the context of the scientist's particular research project.

What is seen through a lens is also represented logically (and is therefore storable and retrievable). Because lenses are defined with RDF and usually do not require programming, users can add new features to them and share them with each other, allowing scientists to define and enhance their view of biological knowledge. An excellent example of a semantic browser is the Haystack project at the Massachusetts Institute of Technology (*7*).

Semantic lenses will also have a major impact on how we create and use portals, which are Web sites designed for integrating various Web-based resources into a common user view. However, there is no mechanism to specify relations between entities contained within the portal views, and so no tangible summary can be constructed from the resources. All that would be required is for each portal subview to define a URI for each of its content objects: text, data fields, chemical structures, tables, images, etc. This would allow any other element or service to "identify" any of the viewable objects throughout the intranet via the SW model.

In the SW paradigm, we begin to consider biological, chemical, and clinical information as part of a viewable and computable web of related facts and hypotheses, not simply as disassociated data bundles. Many of the data models currently used (such as that used within GenBank) were defined at a time when sequence data was submitted in chunks and delimited by technical constraints. More current databases such as Entrez (*8*) and Reactome (*9*) have much more intrinsic connectivity to related information of diverse forms. However, the semantics are implicit and are only seen in the eyes of scientists; these must be made explicit and accessible in order for SW applications to fully use the information. Investigations exploring semantic approaches have been discussed for medical language systems (*10*), health care management (*11*), chemistry (*12*), cancer research (*13*), and clinical trial management (*14*).

So how do we proceed? The good news is that there is already a lot of structured information that is accessible on the Web; most life science data objects and documents can be uniquely tracked with URLs. In some case, data sources such as UniProt have already been converted to RDF (www.isb-sib.ch/~ejain/rdf/). If data objects are assigned a universal identifier (either through LSIDs or another standard), then even if only some of the necessary ontologies have been built, the data can be uniquely mapped on the Web and semantically wrapped (by either the data authorities or by local research communities). This would be an example of a "bottom-up" approach, which has proven to be quite practical in the case of the Gene Ontology Consortium (*15*). Where XML standards already exist [such as at the National Center for Biotechnology Information; MAGE (MicroArray and Gene Expression) for gene expression; CML (Chemical Markup Language) for chemical entities; and CDISC (Clinical Data Interchange Standards Consortium) for clinical data], conversion into SW forms is possible using XML transform tools (see www.w3.org/2005/02/13-KEGG/). Where query (SQL) interfaces exist, RDF query-mapping tools such as SqlDB (www.w3.org/2004/10/04-pharmaFederate/) can be used to federate databases through RDF. On a larger scale, the myGrid (*16*) (www.mygrid.org.uk/) and REWERSE (http://rewerse.net/) projects are investigating the potential of the SW in the context of bioinformatics and analytical workflows. Once information is in RDF/OWL format, it can be aggregated, filtered, and managed with semantic browsers such as Haystack.

To help facilitate the coordination of activities based on real scientific needs and practical strategies, an SW for Life Sciences Interest Group has been formed at W3C to bring the requirements of scientists into close proximity with the semantic technologies community. At the Life Science Semantic Web Workshop held in October 2004 (www.w3.org/2004/10/swls-workshop-report.html), it became clear that there are many critical needs that may be met by such approaches, and that many of the resources and tools available today could be applied. We invite the life science communities to participate together and begin realizing this vision.

References

1. *Innovation or Stagnation, Challenge and Opportunity on the Critical Path for New Products* (FDA Report, U.S. Food and Drug Adminstration, March 2004).
2. J. Enriquez, The data that defines us. *CIO Magazine* **2003**, 22–24 (Fall/Winter 2003).
3. T. Berners-Lee, J. Hendler, O. Lassila, The Semantic Web. *Sci. Am.* **284**, 34–43 (May 2001).
4. J. Hendler, Science and the Semantic Web. *Science* **299**, 520–521 (2003).
5. E. Neumann, E. Miller, J. Wilbanks, What the Semantic Web could do for the life sciences. *Drug Discov. Today BioSilico* **2**, 228–236 (2004).
6. T. Clark, S. Martin, T. Liefeld, Globally distributed object identification for biological knowledgebases. *Brief Bioinform.* **5**, 59–70 (2004).
7. D. Quan, D. Karger, in *Proceedings of the 13th International Conference on World Wide Web* (Association for Computing Machinery Press, New York, 2004), pp. 255–265.
8. D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, W. Helmberg, D. L. Kenton, O. Khovayko, D. J. Lipman, T. L. Madden, D. R. Maglott, J. Ostell, J. U. Pontius, K. D. Pruitt, G. D. Schuler, L. M. Schriml, E. Sequeira, S. T. Sherry, K. Sirotkin, G. Starchenko, T. O. Suzek, R. Tatusov, T. A. Tatusova, L. Wagner, E. Yaschenko, Database resources of the National Center for Biotechnology Information: Update. *Nucleic Acids Res.* **33**, D39–D45 (2005)
9. G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu, L Matthews, S. Lewis, E. Birney, L. Stein, *Nucleic Acids Res.* **33**, D428–D32 (2005).
10. V. Kashyap, The UMLS semantic network and the Semantic Web. *Am. Med. Inform. Asoc. Annu. Symp. Proc.* **2003**, 351–355 (2003).
11. G. Goebel, K. L. Leitner, K. Pfeiffer, Use of Semantic Web technologies in medicine and health care. *Medinfo* **2004**, 1618 (2004).

12. P. Murray-Rust, H. S. Rzepa, S. M. Tyrrell, Y. Zhang, Representation and use of chemistry in the global electronic age. *Org. Biomol. Chem.* **2**, 3192–3203 (2004).
13. S. De Coronado, M. W. Haber, N. Sioutos, M. S. Tuttle, L. W. Wright, NCI thesaurus: Using science-based terminology to integrate cancer research results. *Medinfo* **2004**, 33–37 (2004).
14. M. N. Kamel Boulos, A. V. Roudsari, E. R. Carson, A dynamic problem to knowledge linking Semantic Web service based on clinical codes. *Med. Inform. Internet Med. Sep.* **27**, 127–137 (2002).
15. The Gene Ontology Consortium (M. A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, R. White), The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261 (2004).
16. R. D. Stevens, H. J. Tipney, C. J. Wroe, T. M. Oinn, M. Senger, P. W. Lord, C. A. Goble, A. Brass, M. Tassabehji, Exploring Williams-Beuren syndrome using myGrid. *Bioinformatics* **4** (suppl. 1), I303–I310 (2004).

Citation: E. Neumann, A life science Semantic Web: Are we there yet? *Sci. STKE* **2005**, pe22 (2005).

A Life Science Semantic Web: Are We There Yet?

Eric Neumann

Sci. STKE **2005** (283), pe22.
DOI: 10.1126/stke.2832005pe22