POLICY FORUM

Science and the Semantic Web

James Hendler

Scientists have become increasingly reliant on the World Wide Web for supporting their research endeavors. The Web is used for finding preprints and papers in online repositories, for participating in online discussions at sites such as *Science Online*, for accessing databases through specialized Web interfaces, and even for ordering scientific supplies. When searching for a specific Web site or a paper on a particular topic, engines like Google can do a phenomenal job of sorting through billions of possibilities and identifying potentially useful candidates, often within the first few search results. On specialized Web sites, domain-specific search engines can do even better, for example, enabling the mathematician to easily find papers on "symplectic geometry" or the physicist to see preprints relating to "mesoscopic systems and the quantum hall effect." In fact, the Web has become indispensible for supporting the traditional communications within our disciplines and the needs of scientists within their disciplinary boundaries.

However, as modern science continues its exponential growth in complexity and scope, the need for more collaboration among scientists at different institutions, in different subareas, and across scientific disciplines is becoming increasingly important. Researchers working at one level of analysis may need to find and explore results from another level, from another part of the field, or from a completely dif-

ferent scientific field. On the Web, however, scientists looking for results in sites developed for different scientific communities are often at a loss. For example, a scientist searching for a technique to analyze some image-based data may not know to look for papers on Laplacean invariants (found under the symplectic geometry category in many math sites). A general search on image analysis will find thousands of possibilities but will provide little or no guidance as to which sites can explain how to use the techniques, as opposed to finding papers formalizing the mathematical background, sites for instructors teaching the topics, or reports describing a case where the technique was used. In addi-

tion, the Web is even more limited when it comes to the integration of information from multiple sites or for looking for nontextual information. Current Web technology is clearly insufficient for the needs of interdisciplinary science and comes up short when it comes to supporting the needs of the collaborative and interdisciplinary "e-Science." Fortunately, new Web technologies are emerging with the potential to revolutionize the ability of scientists to do collaborative work. However, to realize this potential, scientists and
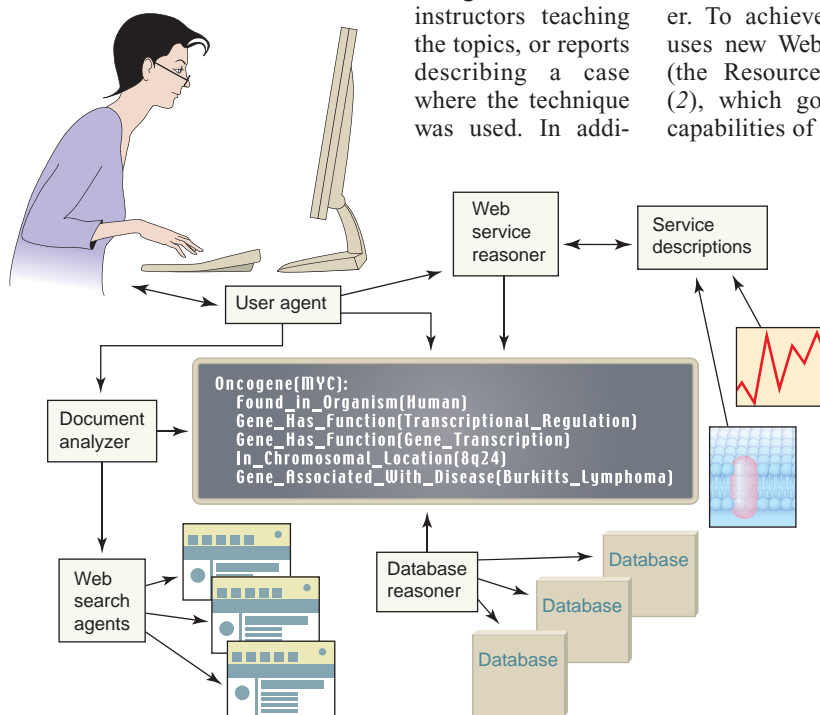
information technologists must forge new models of cooperation, and new thinking must go into the funding and dissemination of this next generation of scientific tools on the Web.

A new generation of Web technology, called the Semantic Web (*1*), is designed to improve communications between people using differing terminologies, to extend the interoperability of databases, to provide tools for interacting with multimedia collections, and to provide new mechanisms for the support of "agent-based" computing in which people and machines work more interactively.

Whereas the current Web provides links between pages that are designed for human consumption, the Semantic Web augments this with pages designed to contain machine-readable descriptions of Web pages and other Web resources. These documents can be linked together to provide information to the computer as to how the terms in one relate to those in another. To achieve this, the Semantic Web uses new Web languages based on RDF (the Resource Description Framework) (*2*), which go beyond the presentation capabilities of HTML (Hypertext Markup Language, which is used on most Web sites today) and the document-tagging capabilities of the Extensible Markup Language (XML), a more recent innovation being used to allow parts of documents to be more precisely delineated.

The Center for Bioinformatics of the U.S. National Cancer Institute (NCI), as part of the Metathesaurus project (*3*), is turning a large vocabulary of cancer research terms into a machine-readable "ontology"—essentially an expanded thesaurus that delineates precise relationships between the vocabulary items and that is available in the RDF-based Web Ontology Language, OWL (*4*). For example, it can provide an expanded definition for an oncogene like the one shown in the center of the figure (left), with respect to organism, function, locus, and associated diseases. Specifically, MYC is found in humans, it has the functions of gene transcription and transcriptional regulation (each of which would be defined elsewhere in the ontology), its unique location is 8q24, and it is associated

The author is in the Computer Science Department, University of Maryland, College Park, MD 20742, USA. E-mail: hendler@cs.umd.edu

with Burkitt's lymphoma. In addition, the definition contains some restrictions on these properties, for example, that there can only be a single, unique value for the chromosomal location and that at least one of the diseases associated with an oncogene must be a cancer.

These new documents provide a way to build a knowledge base that is not restricted to particular keywords and to situate this knowledge base in a distributed way among the documents and resources on the Web. Thus, the oncogene definition, by virtue of being made machine-readable can be linked to by different Web sites, databases, devices, and programs. A Web page describing an ongoing research project in Burkitt's lymphoma could be linked to the definition of MYC by the fact that the disease is associated with that gene. A biotechnologist who has been sequencing chromosome 8 might link data about location 8q24 to this same definition and also link to the loci associated with it, such as PVT1. Other online resources, such as PubMed, could also link to these ontological terms—for example, providing a link from PVT1 to a paper titled "Rearrangement of a DNA sequence homologous to a cell-virus junction fragment in several Moloney murine leukemia virus-induced rat thymomas" (*5*). Thus, the Semantic Web would contain the links needed to find this paper in response to a researcher's query for "chromosomal locus–associated Burkitt's lymphoma," even though the paper does not specifically mention Burkitt's (or MYC) and thus could not be found with a current Web keyword search.

Currently, ontologies like the ones above are just starting to come to the Web, and the links between them are just beginning to be made. In the foreseeable future, the web of links between documents, databases, and programs, using definitions like the ones above, can provide a new level of interaction among scientific communities. For example, the World Health Organization Classification of Neoplastic Diseases (*6, 7*) could become a model, with links to other diseases, databases, and clinical trials. Those trials could be linked to research on epidemiology or causative factors or to ontologies from other fields. Recent workshops have focused on the use of the Semantic Web to support the biosciences (*8*) and environmental science (*9*). Other examples include the National Virtual Observatory (*10*), which is exploring the use of Semantic Web technologies for linking numerous astronomical resources together, and the British MONET project (*11*), which is exploring the use of the Semantic Web for making mathematical algorithms Web-accessible from a variety

of software packages. Further, as these models will use the Semantic Web's common, machine-processable structure, it will become possible for computers to help us make links where relationships are currently unsuspected.

New software tools are being developed for mapping and linking the terms between different ontologies; for using ontologies in the markup of Web sites, scientific publications, and databases; and for capturing semantic metadata about images and other multimedia objects. New search technologies are under development to exploit ontological and other Semantic Web technologies, as well as to extend the capabilities of Semantic Web languages to allow more complex information to be expressed (for example, representing how a particular process might change over time, or how a set of Web-accessible programs could be automatically combined). Of particular note are some of the first demonstrations of Semantic Web "agents" that can integrate the information from Web pages and databases and can pass them to programs for analysis and query processing.

Unfortunately, most scientists are unaware of the Semantic Web effort, and most of the current development is going on separate from the scientific enterprise. This situation parallels that of the development of the original Web, where scientists largely served as customers and users of Web technology, rather than helping to evolve the technology toward the needs of their fields. In fact, much of the information technology research investment for science has gone into technologies that could not compete with the Web and that ended up less used than the commercially available Web technology. Scientific Web site development is often done by publishers or students in their spare time, and being good at bringing science to the Web is typically not seen as a major career enhancer.

There are many reasons for this, but one important one is that crosscutting efforts like this are hard to fund within traditional discipline-oriented review. The e-Science initiative in the UK (*12*) is a good example of how research scientists and information technologists can work together for the betterment of science, and recent efforts to unite the Semantic Web and Grid computing (*13–15*) show great promise. Scientists around the world should unite with their colleagues in the field of Computer Science and Information Technology to push similar interdisciplinary programs. In addition, scientists and information technologists need to work together to make sure Semantic Web technologies are included in programs such as the U.S. National Science Foundation's proposed

CyberInfrastructure or the National Center for Research Resources' BioInformatics Research Network in the United States and similar scientific-infrastructure support programs internationally.

There is also another issue on which information technologists and scientists must start to speak with a single voice. The success of the Semantic Web will be significantly limited if content and tools are not widely shared, at least in the early period of Semantic Web exploration. Much as the original World Wide Web grew from an open-source, open-content model, so too must the Semantic Web. Although it is possible that, in the long run, methods may be developed to blend open and restricted access to Semantic Web content, in the short run, an atmosphere of exploration and cooperation must be fostered. Research scientists must team with their computer science brethren and fight against the intellectual property policies and runaway patent madness that make free dissemination of our products impossible. The original World Wide Web revolution was enabled by open-code, free software, and the wide dissemination of low-cost computing technology. The Semantic Web requires similar openness.

References and Notes

1. T. Berners-Lee, J. Hendler, O. Lassila, *Sci. Am.* (May 2001).
2. Details of these languages and their relationship to the Semantic Web can be found at www.w3.org/2001/sw.
3. See http://ncicb.nci.nih.gov/ and http://ncimeta.nci.nih.gov/indexMetaphrase.html for information on the NCI cancer Metathesaurus project.
4. See www.w3.org/TR/owl-ref/ for the details of the OWL language.
5. G. Lemay, P. Jolicoeur, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 38 (1984); available at www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&uid=84119459&Dopt=r.
6. World Health Organization Classification of Neoplastic Diseases of the Hematopoietic and Lymphoid Tissues: Report of the Clinical Advisory Committee Meeting, Warrenton, VA, November 1997.
7. N. L. Harris *et al.*, *J. Clin. Oncol.* **17**, 3835 (1999).
8. American Medical Informatics Association Symposium on Ontologies, Terminologies, and the Semantic Web for Clinical and Bio Scientists; www.amia.org/2002online/T23.HTM
9. Science on the Semantic Web (SWS) Workshop, Oct 2002; http://cimic.rutgers.edu/semantic/.
10. A. S. Szalay, *ASP Conf. Ser.* **238**, 3 (2001).
11. See http://monet.nag.co.uk/cocoon/monet/index.html.
12. The e-Science Initiative provides infrastructure to allow scientists to access very large data collections, and to use large-scale computing resources and high-performance visualization programs. See www.research-councils.ac.uk/escience.
13. The U.S. Grid computing effort, funded by the National Institutes of Health and the National Science Foundation, provides scientists with access to large-scale computing infrastructure and services. See *The Anatomy of the Grid: Enabling Scalable Virtual Organizations* (*14*).
14. I. Foster, C. Kesselman, S. Tuecke, *Int. J. Supercomput. Appl.* **15**(3), (2001).
15. Integrating Semantic Web and Grid technology was one theme of the Euroweb 2002 Conference held in Oxford, England, on 17 to 18 December 2002; www.w3c.rl.ac.uk/Euroweb/.

521