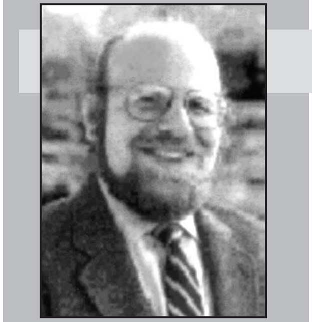

Editor: James Hendler
University of Maryland
hendler@cs.umd.edu

Making Biomedical Ontologies and Ontology Repositories Work

Natalya F. Noy, Daniel L. Rubin, and Mark A. Musen, *Stanford University*

It is becoming impossible to contemplate successful biomedical research without canonical data structures. The biomedical computation community finds itself grappling with hundreds of different knowledge bases, metadata formats, and database schemas. These include primary databases, such as those in GenBank and MEDLINE; metadata that describe the primary data, such as those in caBIO (*ca*ncer *B*ioinformatics *I*nfrastructure *O*bjects); and knowledge bases that codify biomedical concepts, such as the Gene Ontology and SNOMED-CT (*S*ystematized *No*menclature of *Med*icine, *C*linical *T*erms). These data structures are representable in languages such as DICOM (*D*igital *I*maging and *C*ommunications for *M*edicine) and MAGE-ML (*M*icro*a*rray *G*ene *E*xpression-*M*arkup *L*anguage).[1] Many of these data elements and knowledge bases have emerged out of necessity from work that scientists, unfamiliar with data and knowledge representation standards, have done in isolation. Many of these resources fail to follow consistent modeling conventions, so computer programs cannot consistently interpret them.

Workers in biology, clinical medicine, and biomedical informatics are increasingly overwhelmed by the sheer number of knowledge and data resources that different factions promote. As the number of online resources grows, investigators must decide how to incorporate these resources into their work, often without clearly understanding the alternative frameworks' relative merits. How can a cancer biologist compare models of mouse anatomy from Jackson Labs (representing adult anatomy in DAG-Edit format), the University of Pennsylvania (incorporating both mouse and human anatomy in relational format), and the University of Edinburgh (emphasizing developmental anatomy in XML)? How do any of these models relate to the Foundational Model of Anatomy (FMA) developed at the University of Washington and modeled using Protégé or to that of the GALEN (*G*eneralized *A*rchitecture for *L*anguages, *E*ncyclopedias and *N*omenclatures in Medicine) project in Europe modeled in an idiosyncratic description logic? How can an investigator understand the Gene Ontology's limitations and predict how those limitations might affect his or her work, or learn about versions of these models that adhere to knowledge-representation standards and would therefore let these models integrate with other software?

The solution: Virtual ontology repositories

Semantic Web technology and languages such as RDF and OWL can rectify the problem somewhat by providing a common metadata and ontology language and Web-based tools for dealing with ontologies and knowledge structures. However, even if translation mechanisms exist between various biomedical resources and Semantic Web languages (which, by itself, is unlikely to happen for all resources), this translation is only part of the solution.

A distributed set of well-maintained repositories of ontologies and other knowledge sources would help biologists make sense of the huge amount of unrelated information available. We do not envision these as physically containing the ontologies, but rather as virtual repositories, serving as directories and providing a unified view of ontologies distributed across the Semantic Web.

However, just providing access to the resources is not enough and should not be such repositories' primary function. To be a real solution, these repositories must not only provide access to different resources but also, and more importantly, let researchers evaluate different resources, compare them, understand how to integrate them, and learn about others' experiences with them. In an earlier column in this magazine, we discussed some of the capabilities that such repositories should have so researchers could make sense of and use ontologies and other knowledge sources available on the Semantic Web.[2]

Functionalities of an ontology repository

A researcher faced with a task that requires a knowledge resource should be able to access a virtual repository, evaluate its content, understand if any of the resources are relevant to the task, and align the resources to his or her own resources and data. Several components would help to enable such functionality.

1541-1672/04/$20.00 © 2004 IEEE
Published by the IEEE Computer Society

Ontology summarization

To decide whether to buy a book, we read the editor's blurb on its jacket; to decide whether a paper is relevant to our work, we read its abstract. To decide whether a particular ontology fits our application's requirements, we would like to have a summary of what it covers. Such a summary could include, for example, several top levels in the ontology's class hierarchy, perhaps a graphical representation of these concepts, and links between them. We could generate these top-level snapshots automatically or let the author include them as the ontology's metadata.

Ontology ratings

Ontologies and other knowledge sources vary widely in quality, coverage, level of detail, and so on. Furthermore, in general, there are few (if any) objective and computable measures to determine an ontology's quality. Deciding whether an ontology is appropriate for a particular use is a subjective task. Although we cannot often rely on authors to provide information on ontology quality, we can collect this information from people who use ontologies and knowledge sources. This mechanism works well for choosing consumer products on Amazon.com or movies in the Internet Movie Database (www.imdb.com).

Letting ontology users provide ratings and reviews can greatly help life-science researchers find out whether there is an existing ontology suitable for their projects. The reviews should include not only a qualitative assessment of an ontology (is it well developed? does it have major holes? is it correct?) but also, and perhaps more importantly, experience reports. In fact, some communities are beginning to organize such portals already. Open Biological Ontologies, developed in part by the Gene Ontology Consortium, is a prominent example (see the sidebar for links to this and related projects).

Online graphical browsing

If you go to most existing repositories of ontologies, you will usually find flat-file source files. These flat files contain representations of ontologies in OWL, RDF Schema, or an editor-specific format, such as formats for DAG-Edit or Protégé. The flat-file format is never designed for human consumption. So, to view such an ontology and to evaluate whether it is appropriate, you must download the files, install the appropriate graphical tools, and learn how to use them to browse and open the desired ontology. Performing all these steps only to find out that the ontology isn't appropriate for your needs seems like a daunting proposition. Therefore, a repository must be able to provide graphical, searchable, and browsable views of ontologies.

Just as our Web browser can render a flat-HTML file into a user-friendly representation when we click on it in a Google search's list of results, we should be able to click on an ontology file and browse a user-friendly representation of the ontology. Some Web-based browsers for ontologies already exist. Examples include the Protégé Web Browser, which lets you browse and search Protégé and OWL ontologies in a standard Web browser, and Object Viewer, which lets you type in a URL for an RDF file and presents a graph corresponding to the model in the file. However, these browsers must be seamlessly linked with a repository to let you explore their content easily.

Multiple-ontology search

Many ontology-development tools provide query interfaces to ontologies. A number of ontology query languages exist in the context of the Semantic Web, such as TRIPLE[3] and RQL (RDF Query Language).[4] However, these mechanisms traditionally provide a query interface to retrieve concepts in a single ontology. The user can find out if a particular ontology deals with concepts of patients and diseases, but cannot pose this question to the whole ontology library. To the best of our knowledge, virtually no ontology libraries provide a comprehensive cross-ontology search capability. This type of capability would include a keyword search across multiple ontologies, a form-based search, a search for know-ledge-base patterns, and so on.

For instance, a form-based search would let you not only specify the terms in the ontology, but also provide specifics of where these terms should appear. You might specify that the term "cancer" should be a class name, that the ontology should be originated at a particular institution, and so on. In a search based on patterns, you would specify not only a list of terms, but also a high-level view of how the terms should be linked in the ontology. For instance, you might search for all ontologies that link postmenopausal women with specific therapies for breast cancer. Such search capability across multiple ontologies and knowledge sources is crucial on the Semantic Web.

Ontology mapping and alignment

In environments and domains such as biomedicine, many specialists work on developing ontologies. They inevitably create ontologies with overlapping content, with content elements that cannot gracefully connect, and sometimes with components that simply contradict one another. Different ontologies impose different semantic, structural, and syntactic views and expectations on knowledge and data. For example, one ontology that deals with hospital admission records might need to represent time as the exact day and time of a patient's admission, but might need to consider only a simple code for the reason for the admission ("Admission on 05-01-2003 at 14h25min with reason-code 23"). However, a bed-

Related Links

caBIO: http://cabio.nci.nih.gov
Cerner's Clinical Bioinformatics Ontology: www.cerner.com/products/products_3a.asp?id=2940
DICOM: http://medical.nema.org
GenBank: www.ncbi.nlm.nih.gov
Gene Ontology: www.geneontology.org
MGED Ontology: http://mged.sourceforge.net/ontologies
Object Viewer: http://objectviewer.daml.org
Open Biological Ontologies: http://obo.sourceforge.net
Protégé: http://protege.stanford.edu
Protégé plugins: http://protege.stanford.edu/plugins
Protégé Web Browser: http://protege.stanford.edu/plugins/protege_browser/index.html
SNOMED-CT: www.nhsia.nhs.uk/snomed/pages

planning ontology might need only the approximate hospital-stay period ("From Monday 1 May 2003 afternoon to Sunday 7 May morning plus or minus 1 day"), and a patient-record ontology might need a detailed reason for the patient's admission ("Severe asthma crisis, patient still conscious"). Such conceptual and representational mismatches among the ontologies involved must be resolved at the ontological level to enable the integration and exchange of data and knowledge elements.

It is impractical to assume that everyone will eventually conform to a single set of standard ontologies. In fact, experiences even in a mature field such as industrial databases show that a small set of standard schemas and ontologies is still unattainable. It is not uncommon for a large enterprise to use more than a dozen database schemas for purchase orders, for example.

Biomedical knowledge has made greater strides than other fields in developing standard terminologies and vocabularies, such as UMLS (Unified Medical Language System), SNOMED-RT (Reference Terminology), and so on. However, experience shows that application developers often develop custom-tailored and smaller ontologies and link them to the standard terminologies by recording a corresponding UMLS concept identifier for each term, for example, rather than reuse any of the resource wholesale.

Hence, an ontology repository should let contributors create mappings between their ontologies and standard terminologies and vocabularies and between their ontologies and others. Ideally, automated or semiautomated tools should help in this process by identifying candidate mappings and providing infrastructure to record them, query them, and use them in mediating content knowledge.

Ontologies and Protégé tools

For many years, our group has been developing ontology tools that let domain experts develop ontologies and populate them with knowledge. With more than 20,000 registered users, Protégé represents the most widely used, freely available, platform-independent open-source technology for developing and managing large terminologies, ontologies, and knowledge bases (see related sidebar for links to some of these projects). The Protégé system was designed as an open, modular platform upon which developers can build custom-tailored functionality.[5]

Several life-sciences projects have used Protégé as their primary development environment. These projects include the FMA—a declarative representation of anatomy that our colleagues in the Digital Anatomist project developed[6]—as well as Cerner's Clinical Bioinformatics Ontology, the DICE TS (*D*iagnoses for *I*ntensive *C*are *E*valuation *T*erminological *S*ystem),[7,8] MGED Ontology, and verification and identification of errors and inconsistencies in the Gene Ontology.[9]

However, Protégé is not only an ontology-development and knowledge-acquisition tool, but also a platform for developing knowledge-based applications and, more specifically, Semantic Web applications. Its knowledge model, based on the Open Knowledge Base Connectivity protocol,[10] supports a flexible meta-modeling mechanism. This mechanism lets developers build editors for different ontology languages.[11] Protégé plugins support ontology editing in both RDF Schema and OWL. In fact, the Protégé OWL Plugin is arguably the most widely used editor for OWL ontologies.

Protégé's architecture also lets developers extend the environment with a wide range of plugins, which perform various types of inference, provide visualization mechanisms, support queries, and enable access to and integration with standard terminologies, such as UMLS. For example, one such plugin—PROMPT,[12] a suite of ontology-management tools—already provides many of the functionalities that we described in the previous section in the stand-alone environment of Protégé. PROMPT is a suite of ontology-management tools. For instance, PROMPT supports semiautomated ontology merging and mapping. The ontology-versioning support includes a structural comparison of ontology versions and a mechanism to accept and reject changes. A view mechanism lets you extract views from large ontologies. Many of these functionalities can be wrapped in Web Services, becoming a value-added set of services that the ontology repositories provide.[13] So, a life-science researcher could access a repository, determine ontologies that are potentially interesting, compare them to each other and to other standard terminologies, follow development across different versions, and so on.

Although creating such large-scale resources would require significant funding (which hopefully will become available), we can make strides toward this vision by providing the components and tools necessary to implement functions from the list we have discussed and by developing and perhaps standardizing metadata for describing the resources' content, information about their previous and potential use, and their relation to other resources. We can provide these services in small- to medium-scale repositories and link different repositories together.

Some efforts in creating ontology libraries for life sciences are already under way. For example, as we mentioned earlier, the Gene Ontology Consortium recently participated in the development of OBO, a Web site that provides many different biological ontologies and vocabularies. At the moment, OBO is little more than a repository of a diverse set of uploaded ontologies, and it is not yet able to tackle directly the standardization and integration issues that we have raised. However, we can build on resources such as OBO, using the principles that we outlined, to create useful resources for the biomedical community. ■

References

1. P.T. Spellman et al., "Design and Implementation of Microarray Gene Expression Markup Language (MAGE-ML)," *Genome Biology*, vol. 3, no. 9, 2002; http://genomebiology.com/2002/3/9/research/0046.
2. Y. Sure et al., "Why Evaluate Ontology Technologies? Because It Works!" *IEEE Intelligent Systems*, vol. 19, no. 4, 2004, pp. 74–81.
3. M. Sintek and S. Decker, "TRIPLE—A Query, Inference, and Transformation Language for the Semantic Web," *Proc. 1st Int'l Semantic Web Conf.* (ISWC 2002), LNCS 2342, I. Horrocks and J.A. Hendler, eds., Springer-Verlag, 2002, pp. 364–378.
4. G. Karvounarakis et al., "RQL: A Declarative Query Language for RDF," *Proc. 11th Int'l World Wide Web Conf.*, ACM Press, 2002, pp. 592–603.
5. J. Gennari et al., "The Evolution of Protégé: An Environment for Knowledge-Based Systems Development," *Int'l J. Human-Computer Interaction*, vol. 58, no. 1, 2002, pp. 89–123.
6. C. Rosse and J.L.V. Mejino, "A Reference Ontology for Bioinformatics: The Foundational Model of Anatomy," *J. Biomedical Informatics*, vol. 36, no. 6, 2003, pp. 478–500.
7. N. de Keizer et al., "Design of an Intensive Care Diagnostic Classification," *Methods of Information in Medicine*, vol. 38, no. 2, 1999, pp. 102–112.

8. A. Abu-Hanna et al., "Protégé as a Vehicle for Developing Medical Terminological Systems," to be published in *Int'l J. Human-Computer Studies*, 2005.

9. I. Yeh et al., "Knowledge Acquisition, Consistency Checking and Concurrency Control for Gene Ontology (GO)," *Bioinformatics*, vol. 19, Jan. 2003, pp. 241–248.

10. V. Chaudhri et al., "OKBC: A Programmatic Foundation for Knowledge Base Interoperability," *Proc. 15th Nat'l Conf. Artificial Intelligence* (AAAI-98), AAAI Press/The MIT Press, 1998, pp. 600–607.

11. N.F. Noy et al., "Creating Semantic Web Contents with Protégé," *IEEE Intelligent Systems*, vol. 16, no. 2, 2001, pp. 60–71.

12. N.F. Noy and M.A. Musen, "The PROMPT Suite: Interactive Tools for Ontology Merging and Mapping," *Int'l J. Human-Computer Studies*, vol. 59, no. 6, 2003, pp. 983–1024.

13. O. Dameron et al., "Accessing and Manipulating Ontologies Using Web Services," *The Semantic Web—ISWC 2004: Proc. 3rd Int'l Semantic Web Conf.*, LNCS 3298, S.A. McIlwraith, D. Plexousakis, and F. Van Harmelen, eds., Springer-Verlag, 2004.

Natalya F. Noy is a senior research scientist at the Stanford Medical Informatics Laboratory at Stanford University. Contact her at Stanford Medical Informatics, 251 Campus Dr., Stanford Univ., Stanford, CA 94305; noy@smi.stanford.edu.

Daniel L. Rubin is a research scientist at the Stanford Medical Informatics Laboratory at Stanford University. Contact him at Stanford Medical Informatics, 251 Campus Dr., Stanford Univ., Stanford, CA 94305; rubin@smi.stanford.edu.

Mark A. Musen is a professor of medicine (medical informatics) and computer science at Stanford University and is head of the Stanford Medical Informatics laboratory. Contact him at Stanford Medical Informatics, 251 Campus Dr., Stanford Univ., Stanford, CA 94305; musen@smi.stanford.edu.