Making Ontologies and Ontology Libraries Work

Natalya F. Noy, Daniel Rubin, and Mark A. Musen
Stanford Medical Informatics, Stanford University,
251 Campus Drive, x-215, Stanford, CA 94305, USA
{noy,rubin,musen}@smi.stanford.edu

Today, it is impossible to contemplate successful biomedical research in the absence of canonical data structures. From primary databases (such as those found in GenBank[1] and MEDLINE) to meta-data that describe the primary data (such as those in caBIO[2]) or representable in languages such as DICOM[3] and MAGE-ML (Spellman *et al.* 2002) to knowledge bases that codify biomedical concepts (such as the Gene Ontology[4] and SNOMED-CT[5]), the biomedical computation community finds itself grappling with hundreds of different knowledge bases, meta-data formats, and database schemas. Many of these data elements and knowledge bases have emerged, out of necessity, from work that has been done in isolation by scientists who are unfamiliar with data and knowledge representation standards. Many of these resources fail to follow consistent modeling conventions and thus cannot be consistently interpreted by computer programs.

Semantic Web and Semantic Web languages, such as RDF and OWL can rectify the problem somewhat by providing a common metadata and ontology language, as well as web-based tools for dealing with ontologies and knowledge structures. However, even if there are translation mechanisms 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.

Workers in basic biology, clinical medicine, and biomedical informatics are becoming increasingly overwhelmed by the sheer number of knowledge and data resources that are being promoted by different factions. As the number of online resources grows, investigators must make decisions about how to incorporate these resources into their work—often without a clear understanding of the relative merits of alternative frameworks. How can a cancer biologist compare the alternative models of mouse anatomy from Jackson Labs (representing adult anatomy, in DAG-Edit format) or from the University of Pennsylvania (incorporating both mouse and human anatomy, in relational format) or from the University of Edinburgh (emphasizing developmental anatomy, in XML)? How do any of these models relate to the Foundational Model of Anatomy developed at the University of Washington (modeled using Protégé) or that of the GALEN project in Europe (modeled in an idiosyncratic description logic)? How can an investigator understand the limitations of the Gene Ontology and predict how those limitations might affect his or her work? How can an investigator learn about the existence of versions of these models that adhere to knowledge-representation standards that would allow these models to integrate with other software?

Part of the solution to helping biologists make sense of the huge amount of unrelated information available to them, is a *distributed set of well-maintained repositories of ontologies and other knowledge sources*. However, just providing access to the resources is not nearly enough and should not be the primary function of such repositories. To be a real solution, these repositories must not only provide access to different resources but also—and more important—enable researchers to evaluate different resources, to compare them to one another, to understand how to integrate them, and to learn about others' experience with these resources. The Semantic Web technologies provide a set of common representation standards that not only for representing content of these resources, but also for representing metadata about the resources and relations between components of the resources. In this position paper, we lay out some of the desired features of such repositories and the types of metadata describing their content that will make such repositories useful for life-science researchers.

Functionalities of an Ontology Repository

Here is a slightly more detailed look at some of the capabilities that would be essential for life-science researchers to make sense of and to use ontologies and knowledge sources available on the Semantic Web. A

[1] http://www.ncbi.nlm.nih.gov/

[2] http://cabio.nci.nih.gov/

[3] http://medical.nema.org/

[4] http://www.geneontology.org/

[5] http://www.nhsia.nhs.uk/snomed/pages

researcher faced with a task that requires a knowledge resource should be able to access such a repository, evaluate its content, understand if any of the resources there are relevant to his task, and be able to align the resources to his own resources and data. The following components can enable such functionality.

Ontology summarization To decide whether to buy a book, we read the editor's blurb on a book jacket; to decide whether a paper is relevant to our work, we read its abstract. To decide whether a particular ontology fits the requirements of our application, we would like to have some "abstract" or "summary" of what this ontology covers. Such summary can include, for example, a couple of top levels in the ontology's class hierarchy; perhaps a graphical representation of these top-level concepts and links between them. We can generate these top-level snapshots automatically or allow ontology authors to include them as meta-data for an ontology. The summary can also include "hub" concepts in the ontology. A "hub" concept can be a concept with the largest number of links in and out of it. More interesting, we can experiment with metrics similar to Google's PageRank: the concept is more important if other important concepts have links to it. This computation can take into account semantics of specific links (giving a subclass–superclass link a lower value than a property link, for instance) or exclude some links or properties. By experimenting with these measures, we can find out which measures yield the concepts that users deem important. The hub concepts are often much better starting points in exploring and understanding an ontology than the top level of a class hierarchy.

Rating of ontologies In addition to reading an editorial blurb on a book jacket to decide whether we want to buy a book, we often read reviews of the book both by book critics and other readers. Similarly, when choosing a movie or a consumer product, such as a coffee maker or a pair of skis, we use the Web to find opinions of others. Many readers of this paper are probably frequent visitors to such sites as the Internet Movie Database (`www.imdb.com`) or Amazon customer reviews (`www.amazon.com`). A similar network for ontologies would help tremendously in guiding life-science researchers in finding whether there is an existing ontology that would be suitable for their projects. The reviews should include not only the qualitative assessment of an ontology (is it well-developed? does it have major holes? is it correct?), but also, and, perhaps, more important, experience reports. In fact, some communities are beginning to organize such portals already (see for example `obo.sourceforge.net`).

Multiple-ontology search Today, many ontology-development tools provide query interfaces to ontologies. A number of ontology query languages exist in the context of Semantic Web, such as TRIPLE (Sintek & Decker 2002) and RQL (Karvounarakis *et al.* 2002). 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, there are virtually no ontology libraries that provide comprehensive cross-ontology search capability. A comprehensive search capability would include keyword search across multiple ontologies, form-based search, search for knowledge-base patterns, and so on. For instance, a form-based search would not only allow users to specify the terms in the ontology, but also provide specifics of where these terms should appear. A user may specify that a term "cancer" should be a class name, that the ontology should be originated at a particular institution, and so on. In the search based on patterns, a user specifies not only a list of terms, but also a high-level view of how the terms should be linked in the ontology. For instance, a user may search for all ontologies that link post-menopausal women with specific therapies for breast cancer. Such search capability across multiple ontologies and knowledge sources is crucial on the Semantic Web.

Views and customization To evaluate an ontology properly, a user may need to see a *view* of an ontology that takes into account his expertise, his perspective, the required level of granularity, or a subset of the domain covered by the ontology that he is interested in. For instance, if we are developing an application that studies breast cancer, we may want to use a standard anatomy ontology, such as the Foundational Model of Anatomy (FMA)(Rosse & Mejino 2004). However, the FMA is huge and very complex (67,000 distinct concepts at the time of this writing). We may choose to use only a subset of it that includes the breast and related organs. Similarly, while FMA takes a structure-based view of anatomy and is developed as a general reference model, a radiologist or a person developing biological simulations may use different terms or view some relationships differently. If we can enable ontology developers to annotate concepts and relations with information describing the perspectives in which these terms and relations should appear, and how the terms should be presented or named in each perspective, we will be able to present these different perspectives automatically. Similarly, an ontology developer may want to indicate that certain concepts or relations should be displayed only to the users who identify themselves as experts (presenting a simpler, trimmed-down view for novices). For a life-science researcher who wants to use an ontology, it is often much easier to evaluate a smaller ontology that has only the concepts related to his concepts of interest than to evaluate a large general reference resource.

Ontology mapping and alignment In environments and domains, such as biomedicine, where many specialists work on developing ontologies, they inevitably create ontologies with overlapping content, with content elements that cannot gracefully connect with one another, 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 may need to represent time as the exact day and time of a patients admission but may only need to consider a simple code for the reason of the admission (e.g., Admission on 05-01-2003 at 14h25min with reason-code 23), whereas a bed-planning ontology may only need the approximate hospital-stay period (e.g., From Monday May 1st 2003 afternoon to Sunday May 7th morning plus or minus 1 day), and a patient-record ontology may need the detailed reason for which the patient was admitted at the hospital (severe asthma crisis, patient still conscious). Such conceptual and representational mismatches between all ontologies involved need to be resolved at the ontological level in order to enable the integration and exchange of data and knowledge elements between these ontologies.

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

While the field of biomedical knowledge has made far greater strides than other fields in developing standard terminologies and vocabularies (e.g., UMLS, SMOMED-RT, etc.), current experience shows that application developers often develop custom-tailored and smaller ontologies and link them to the standard terminologies by recording a corresponding UMLS COD for each term, for example, rather than reuse any of the resource wholesale.

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

Protégé Tools

In our group, for many years, we have been developing ontology tools that enable domain experts to develop ontologies and populate them with knowledge. Protégé[6] represents the most widely used freely available, platform-independent, open-source technology for developing and managing large terminologies, ontologies, and knowledge bases. It has more than 20,000 registered users at the time of this writing. The Protégé system was designed from the beginning as an open, modular platform upon which developers can build custom-tailored functionality (Gennari *et al.* 2003).

Protégé has been used as the primary development environment for several projects in Life Sciences. These projects include the Foundational Model of Anatomy (FMA)—a declarative representation of anatomy developed by our colleagues in the Digital Anatomist project at the University of Washington (Rosse & Mejino 2004), Cerner's Clinical Bioinformatics Ontology,[7] the DICE TS (de Keizer *et al.* 1999; Abu-Hanna *et al.* 2004), MGED Ontology,[8] and verification and identification of errors and inconsistencies in the Gene Ontology (Yeh *et al.* 2003).

However, Protégé is not only an ontology-development and knowledge-acquisition tool, but also a platform for developing knowledge-based application, and, more specifically, Semantic Web applications. Its *knowledge model*, which is based on the Open Knowledge Base Connectivity (OKBC) protocol (Chaudhri *et al.* 1998), supports very flexible meta-modeling mechanism. This mechanism enables us to build ontology editors for different ontology languages. There are Protégé plugins to support ontology editing in both RDF[9] and OWL.[10] In fact, the Protégé OWL Plugin is arguably the most widely used editor for OWL ontologies today. Second, its *architecture* allows developers to extend the environment with a wide range of plugins, that perform various types of inference, provide visualization mechanisms, support queries, enable access to and integration with standard terminologies, such as UMLS.

One such plugin is PROMPT (Noy & Musen 2003)—a suite of tools for ontology management. This suite provides some of the functions that we have described in the previous section in the stand-alone environment of Protégé. For instance, it supports semi-automated ontology merging and mapping. The ontology-versioning support includes structural comparison of ontology versions and a mechanism to accept and reject changes. The view mechanism allows extraction of views from large ontologies. Many of these functionalities can be wrapped in Web Services (see a position statement for this workshop by Olivier Dameron and Mark Musen) and become a value-added set of services provided by ontology repositories. Then a life-science researcher can access such repository, determine ontologies that are potentially interesting to him, compare them to one another and to other standard terminologies, follow development across different versions, and so on.

[6] `http://protege.stanford.edu`

[7] `http://www.cerner.com/products/products_3a.asp?id=2940`

[8] `http://mged.sourceforge.net/ontologies/`

[9] `http://protege.stanford.edu/plugins/rdf/`

[10] `http://protege.stanford.edu/plugins/owl/index.html`

Concluding Remarks

While creating large-scale resources like this one would require significant funding (and, hopefully, funding for these projects will become available), we can make strides towards this vision by providing components tools necessary to implement various functions from the list above, developing and perhaps standardizing metadata for describing not only the content of resources but also information about their previous and potential use and relation to other resources, by providing these services in small to medium scale repositories, and by linking different repositories together.

Some efforts in creating ontology libraries for life sciences are already under way. For example, the Gene Ontology consortium recently participated in the development of Open Biological Ontologies (OBO).[11] OBO is a Web site on sourceforge that provides many different biological ontologies and vocabularies. However, OBO is little more than a repository of a diverse set of uploaded ontologies, and it is not yet able to directly tackle the standardization and integration issues raised above. We can build on resources such as OBO, using principles described in this position paper to create useful resources for the biomedical community.

References

Abu-Hanna, A.; Cornet, R.; de Keizer, N.; Crubézy, M.; and Tu, S. 2004. Protégé as a vehicle for developing medical terminological systems. *International Journal of Human-Computer Studies* to appear.

Chaudhri, V.; Farquhar, A.; Fikes, R.; Karp, P.; and Rice, J. 1998. OKBC: A programmatic foundation for knowledge base interoperability. In *Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, 600–607. Madison, Wisconsin: AAAI Press/The MIT Press.

de Keizer, N.; Abu-Hanna, A.; Cornet, R.; Zwetsloot, J. H. M. .; and Stoutenbeek, C. 1999. Design of an intensive care diagnostic classification. *Methods of information in Medicine* 38(2):102–112.

Gennari, J.; Musen, M. A.; Fergerson, R. W.; Grosso, W. E.; Crubézy, M.; Eriksson, H.; Noy, N. F.; and Tu, S. W. 2003. The evolution of Protégé: An environment for knowledge-based systems development. *International Journal of Human-Computer Interaction* 58(1).

Karvounarakis, G.; Alexaki, S.; Christophides, V.; Plexousakis, D.; and Scholl, M. 2002. RQL: A declarative query language for RDF. In *Eleventh International World Wide Web Conference*, 592603.

Noy, N. F., and Musen, M. A. 2003. The PROMPT suite: Interactive tools for ontology merging and mapping. *International Journal of Human-Computer Studies* 59(6):983–1024.

Rosse, C., and Mejino, J. L. V. 2004. A reference ontology for bioinformatics: The foundational model of anatomy. *Journal of Biomedical Informatics.*

Sintek, M., and Decker, S. 2002. TRIPLE—a query, inference, and transformation language for the semantic web. In *International Semantic Web Conference (ISWC).*

Spellman, P. T.; Miller, M.; Stewart, J.; Troup, C.; Sarkans, U.; and et.al. 2002. Design and implementation of microarray gene expression markup language (mage-ml). *Genome Biology* 3.

Yeh, I.; Karp, P.; Noy, N.; and Altman, R. 2003. Knowledge acquisition, consistency checking and concurrency control for gene ontology (go). *Bioinformatics* 19:241–248.

[11] http://obo.sourceforge.net/