BY MARK A. MUSEN, MD, PhD

It Takes a Village:
Building the Next Generation
of Biomedical Ontologies

Although the notion of ontology has been around since Aristotle, the perceived need to develop ontologies in biomedicine has accelerated in recent years as investigators attempt to make sense of the terabytes of high-throughput data that are now finding their way into public repositories. While the number of biomedical terminologies and ontologies continues to increase as new areas of biomedical content become formalized, the creation and annotation of these resources can't quite keep up. The flood of information may necessitate a new approach involving vastly more ontology developers. It may, in fact, take a village.

The construction of biomedical ontologies has long been a cottage industry, with even vast systems such as SNOMED (the Systemized Nomenclature of Medicine) initially representing the handiwork of a very small group of dedicated individuals. Venerable ontologies such as the Foundational Model of Anatomy and the NCI Thesaurus represent the work of a surprisingly small set of developers. Nevertheless, as the demand for ever larger and more granular ontologies accelerates, and as large-scale systems such as the International Classification of Disease are being reengineered, the scientific community has increasingly raised concerns about whether ontology development ultimately can be a scalable enterprise. Practical ontologies comprise tens of thousands of concepts, and a handful of individuals can never have personal knowledge of everything that needs to be represented in such a system.

To address this problem, workers in biomedicine are attempting to democratize the development of large-scale ontologies. The engineering of the Gene Ontology, for example, has been characterized by an open development process to which nearly anyone can contribute. The actual editing of the Gene Ontology content, however, is still performed by only a handful of trusted curators. The National Cancer Institute is experimenting with an open process for extensions to the current content of the NCI Thesaurus via the BiomedGT initiative. Here, nearly anyone can annotate the ontology via a Web-based wiki and suggest changes and extensions, although again, the modification of the actual ontology content will be channeled through a set of trained individuals who understand principles of knowledge representation and the use of knowledge-editing tools.

Probably the best exemplar of an open, nearly democratic ontology-development initiative is the Open Directory Project (ODP). Founded more than 10 years ago, the ODP has enlisted more than 75,000 volunteers to flesh out the extensive open-content ontology of Web pages that has been adopted by Google, Yahoo!, Netscape, and a host of other companies. The ODP has generated an enormous ontology (commonly known as dmoz) that provides standard, categorized entrée to virtually all the content on the Web. All of us use dmoz, perhaps unknowingly, every time we browse the Web by categories in Google and Yahoo!, rather than searching the Web's free text for particular terms. Embracing everything imaginable that a user could search for, dmoz is a remarkable demonstration of how scalable ontology engineering can be, particularly when volunteers step forward to provide fine-grained descriptions of their particular areas of personal interest.

At bioportal.bioontology.org, the user is presented with a list of possible ontologies to explore and visualize. This screenshot of the Cellular Components ontology shows the relationship between major components and some of the minor components. Any registered user can comment on any ontology.

guest editorial

The dmoz ontology is very simple in its structure, and lacks the rich semantics of ontologies developed in formal knowledge representation systems such as the Web Ontology Language (OWL). When the developers of dmoz make modeling errors, the consequences are unlikely ever to impede the advancement of science or to threaten lives. Nevertheless, the dmoz ontology stands as a stunning example of how legions of volunteers can be mobilized to generate an enormous and undeniably useful ontology. Imagine if the lessons of dmoz could be applied to SNOMED or to BiomedGT!

At the National Center for Biomedical Ontology (NCBO), we are experimenting with ways in which the biomedical community can take an active part in contributing to the construction of scalable ontologies and controlled terminologies. Our BioPortal system allows any registered user to comment on any ontology in our distributed repository, to comment on the comments left by other users, and to demonstrate how the elements of one ontology may relate to those of another. We have used this capability extensively in the engineering of the Biomedical Resource Ontology used to describe the online software and data resources developed by the National Centers for Biomedical Computing and by the recipients of Clinical and Translational Science Awards. BioPortal, at present, does not play a role in completely open ontology editing, however.

There are very legitimate concerns about how we can

> It is clear that the ontology-development community needs at least to experiment with new methods of ontology engineering that can scale to future biomedical requirements.

maintain the quality of ontologies if the development process is democratized. Organizations such as the Open Biomedical Ontologies (OBO) Foundry have been established under the assumption that there must always be central management of ontology development to ensure the quality of the content. And yet there continue to be too much data, too many medical records, and too many experiments for the ontology-development community to keep up with existing needs.

I don't know whether the dmoz approach will really be practical in biomedicine, but it is clear that the ontology-development community needs at least to experiment with new methods of ontology engineering that can scale to future biomedical requirements. Surely there are ways to take advantage of the expertise distributed among all biomedical investigators in a way that will overcome many of the limitations of centralized ontology curation. Workers at NCBO are extremely excited about the possibilities that new technology might provide in enabling this more open approach to ontology engineering. Experimentation with community-based ontology development not only may accelerate the engineering of badly needed ontology content, but also can provide a laboratory for the study of new mechanisms for collaboration and interaction in biomedicine. □

DETAILS

BioPortal: http://bioportal.bioontology.org

The Open Director Project: http://www.dmoz.org

Mark A. Musen, M.D., Ph.D. is Professor of Medicine (Biomedical Informatics Research) and Computer Science at Stanford University. He is Director of the Stanford Center for Biomedical Informatics Research and principal investigator of the National Center for Biomedical Ontology (NCBO).

A Note from the Managing Editor:

Thanks to all who participated in the BCR survey. Your names were entered in a drawing for an iPod shuffle which went to Alan Villalobos from DNA2.0. The survey results are helping us to plan for the future.

If you didn't get a chance to answer the survey, you can still give us feedback on the magazine by visiting http://www.biomedicalcomputationreview.org and clicking on the "Feedback" link.

Starting in our next issue, we will launch a new "Debate" column, starting with the topic selected by the survey respondents: "To Mine or Not to Mine: Are clinical data repositories useful sources of untapped discoveries awaiting data-mining algorithms or are they too noisy and messy."

Best,

Kathy Miller MANAGING EDITOR