BIOMEDICAL ONTOLOGIES

OLIVIER BODENREIDER

U.S. National Library of Medicine
8600 Rockville Pike, MS 43, Bethesda, Maryland, 20894, USA
E-mail: olivier@nlm.nih.gov

JOYCE A. MITCHELL

U.S. National Library of Medicine and
University of Missouri, Department of Health Management & Informatics,
Columbia, Missouri, 65211, USA
E-mail: MitchellJo@health.missouri.edu

ALEXA T. MCCRAY

U.S. National Library of Medicine
8600 Rockville Pike, MS 52, Bethesda, Maryland, 20894, USA
E-mail: mccray@nlm.nih.gov

Biomedical ontologies provide an organizational framework of the concepts involved in biological entities and processes in a system of hierarchical and associative relations that allows reasoning about biomedical knowledge. In contrast, biomedical terminologies promote a standard way of naming these concepts. Differences among various kinds of terminological systems can be briefly summarized as follows. *Controlled vocabularies* define a set of terms to be used for a given purpose (e.g., indexing the literature, annotating gene functions). *Thesauri* organize the terms in a system of relations designed to help navigate among terms as needed, for example, in information retrieval tasks. *Ontologies*, on the other hand, aim at representing what exists independently of any specific use; they also typically follow general theories (e.g., mereology) and carefully distinguish between the various kinds of relations among things that exist. Thesauri are often limited to tasks such as information retrieval, whereas ontologies support reasoning. Both can be shared, but ontologies lend themselves to reuse, sometimes in widely differing applications from the ones for which they were originally designed. Although more than sixty terminological systems exist in the biomedical domain, few actually qualify as an ontology. Interestingly, the most recent systems tend to be ontologies, developed either from the top down (e.g., GALEN[1]) or from reengineering the knowledge present in older systems (e.g., SNOMED-CT[2]).

[1] http://www.opengalen.org/
[2] http://www.snomed.org/

Proceedings of 2003 the Pacific Symposium on Biocomputing. World Scientific; 2003. p. 562-564.

Biological knowledge is evolving so rapidly that it is difficult for most scientists to assimilate and integrate the new information with their existing knowledge. One advantage of ontologies over terminological systems is to support reasoning. The formal structure and rules of inference provided by logic may be coupled with the properties of the relations among things in an ontology in order to draw inferences. The uses of bioinformatics ontologies include natural language processing, knowledge discovery, and supporting interoperability among the many knowledge resources now available. Bridging between the terminological resources of the UMLS® (Unified Medical Language System®)[3] and the biotechnology information resources is another important issue. Increasingly, natural language processing techniques are applied to massive biomedical corpora such as the MEDLINE® bibliographic database[4] in order to extract information and discover knowledge. In these tasks, while terminology is needed for identifying the concepts in the text, ontologies help identify the relationships among concepts suggested by syntactic and discourse structures.

Ontologies are not tied to any kind of particular formalism for their representation, nor are they concerned, in principle, with issues of computer tractability. In practice, however, some representations such as frames (used, for example, in Protégé[5]) or description logics (e.g., DAML+OIL[6]) represent a trade-off between expressivity (what can be represented) and tractability (what can be inferred), needed if the ontology is to be used in computational tasks. Both representations are expressed in or can be translated to some flavor of first-order logic. Recent biomedical ontologies such as GALEN and SNOMED-CT are based on description logics; others such as the Foundational Model of Anatomy[7] are frame-based.

Although the number of ontologies available for biomedicine has not yet reached that of terminological systems, it is expected that applications relying on domain knowledge will have to deal with multiple ontologies, either because no single ontology offers a broad enough coverage, or because the task is to interoperate between applications using different ontologies. For example, a repository of interconnected ontologies represented in a standard formalism is what is envisioned as a possible infrastructure for the Semantic Web[8]. Different approaches can be used to reconcile the knowledge from distinct ontologies. The

[3] http://umlsinfo.nlm.nih.gov/
[4] http://www.nlm.nih.gov/
[5] http://protege.stanford.edu/index.html
[6] http://www.daml.org/
[7] http://sig.biostr.washington.edu/projects/da/
[8] http://www.semanticweb.org/

classical approach consists of developing methods and tools for aligning or merging several ontologies. Proposed more recently, ontology negotiation would allow applications operating on multiple ontologies to cooperate in performing a task by using similarities and differences in the ontologies in order to establish communication among them.

The availability of domain ontologies capturing the knowledge of specific subdomains of biomedicine (e.g., molecular functions, subcellular localization) is important to many applications. Conversely, by providing a framework for these domain ontologies to hook to, upper level ontologies represent an important, yet less popular, aspect of ontology development. The theories represented in upper level ontologies are general theories such as the theory of parts and wholes, the theory of dependence, and the theory of boundaries. Although not sufficient in itself for representing the knowledge of a domain, an upper level ontology provides the basis for making explicit the difference between, for example, substances and processes. IEEE's Standard Upper Ontology[9] (SUO) working group is developing a standard for specifying "a structure and a set of general concepts upon which domain ontologies (e.g., medical, financial, engineering, etc.) could be constructed". Many general properties will be defined at this upper level and these can be inherited by the domain ontologies hooked underneath them.

Biological knowledge is evolving from structural genomics towards functional genomics. The tremendous amount of DNA sequence information that is now available provides the foundation for studying how the genome of an organism is functioning, and microarray technologies provide detailed information on the mRNA, protein, and metabolic components of organisms. This knowledge allows researchers to discover new metabolic pathways, to model metabolic and regulatory networks in living organisms, and ultimately to understand the pathogenesis of diseases. In this context, in the perspective of acquiring knowledge from the literature or from various and often heterogeneous databases, it is fundamental that not only biologic knowledge, but also medical knowledge be accurately represented and the appropriate linkages be made between these domains. Existing and yet to be developed biomedical ontologies play a critical role in this effort.

[9] http://suo.ieee.org/