Semantic Integration in Biomedicine

Olivier Bodenreider and Songmao Zhang

U.S. National Library of Medicine, Bethesda, Maryland
National Institutes of Health, Department of Health & Human Services
{olivier|szhang}@nlm.nih.gov

Semantic integration research at NLM

In 1986, the National Library of Medicine (NLM) initiated a terminology integration project – the Unified Medical Language System® (UMLS®) – as "an effort to overcome two significant barriers to effective retrieval of machine-readable information": the variety of names used to express the same concept and the absence of a standard format for distributing terminologies. By integrating more than 60 families of biomedical vocabularies, the UMLS Metathesaurus® currently provides not only an extensive list of names (2.5 million) for its 900,551 concepts, but also over 12 million relations among these concepts. Its scope is broader and its granularity finer than that of any of its source vocabularies.

The major component of the UMLS is the Metathesaurus, a repository of inter-related biomedical concepts. The two other knowledge sources in the UMLS are the Semantic Network, providing high-level categories used to categorize every Metathesaurus concept, and lexical resources including the SPECIALIST lexicon and programs for generating the lexical variants of biomedical terms. The lexical resources play an important role in semantic integration by identifying lexically similar concepts. The potentially synonymous terms are reviewed by the Metathesaurus editors prior to being integrated into the UMLS.

As illustrated in Figure 1, by integrating the vocabulary of several subdomains of biomedicine, the Metathesaurus can be used for the integration of the various information systems and databases existing for these subdomains. For example, recently integrated terminologies include the NCBI taxonomy, used for identifying organisms, and Gene Ontology™, used for the annotation of gene products across various model organisms. The Metathesaurus also covers the biomedical literature with the Medical Subject Headings (MeSH), the controlled vocabulary used to index MEDLINE, a large bibliographic database. Core subdomains such as anatomy, used across the spectrum of biomedical applications, are also represented in the Metathesaurus with the Digital Anatomist Symbolic Knowledge Base. Finally, the subdomain represented best is probably the clinical component of biomedicine, with general terminologies such as SNOMED® International (and soon SNOMED-CT®), and the International Classification of Diseases, to name a few.

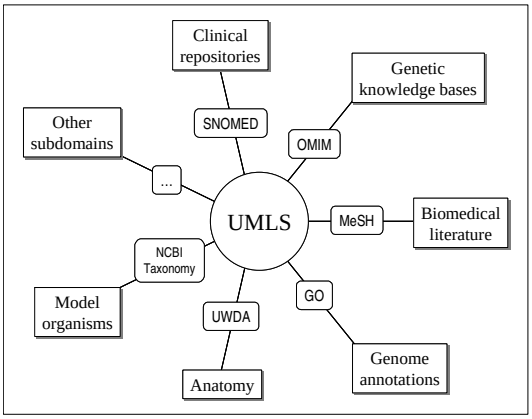

Figure 1. *The various subdomains integrated in the UMLS.*

More recently, the Medical Ontology Research project was initiated at NLM. The objective of this project is not to build an ontology of the biomedical

domain, but rather to develop methods whereby ontologies could be acquired from existing resources (including the UMLS Metathesaurus), as well as validated against other knowledge sources. Toward this endeavor, we have developed methods for aligning the UMLS with general ontologies (e.g., Cyc, WordNet) or specialized ones (e.g., the Gene Ontology). Additionally, methods have been developed for aligning UMLS knowledge sources (the Metathesaurus with the Semantic Network) and several biomedical ontologies outside the UMLS (the Foundational Model of Anatomy and GALEN). Related work developed as part of the Medical Ontology Research project also includes studying consistency and redundancy in biomedical terminologies and ontologies.

In the last eighteen months, we have been particularly interested in comparing two representations of anatomy: the Foundational Model of Anatomy and GALEN. Although the ultimate goal of this study is to compare the reasoning potential of these two ontologies, we have devoted most of the effort so far to aligning the two ontologies using a combination of lexical and structural techniques. We have also studied from both a quantitative and a qualitative perspective the contribution to the alignment of the different techniques used to obtain relationships from each ontology (knowledge augmentation, inference, etc).

Challenges and solutions

The challenging issues in semantic integration are many. In the biomedical domain, polysemy is one of them. For example, in molecular biology, a gene, the protein it produces, and the disease resulting from a mutation of this gene often have the same name. While geneticists and biologists usually have no problem identifying what is referred to by a particular name, this may not be the case for computer programs performing tasks such as information extraction or semantic interpretation.

While there are relatively few biomedical ontologies, there are, in contrast, many terminology systems developed for various purposes. Instead of building a medical ontology from the top-down (e.g., GALEN), the UMLS has attempted to integrate these terminology systems. Although the resulting Metathesaurus does not claim to be an ontology, we believe it can be used as the basis for building one. The biggest issue here is that the relations useful for organizing biomedical concepts for a given purpose (e.g., information retrieval) may not always be principled or consistent across terminological systems.

This approach to integrating many terminologies results in a semantic structure that may contain inconsistencies. On the other hand, redundancy is another feature of such systems that can be beneficial to semantic integration. The assumption here is that relations that appear in several sources are more likely to be semantically valid than relations asserted by one source only.

We also believe that domain knowledge can largely benefit semantic integration. Instead of using generic systems such as schema matching, we usually prefer to take advantage of the specific features of a given domain. For example, as illustrated in our paper, linguistic clues can be used reliably for extracting relations from anatomical concept names.

About the authors

Olivier Bodenerider is a Staff Scientist at the Lister Hill National Center for Biomedical Communications, US National Library of Medicine. He obtained a M.D. degree from the University of Strasbourg, France, in 1990 and a Ph.D. in Medical Informatics from the University of Nancy, France, in 1993. His research interests include terminology, knowledge representation, and ontology in the biomedical domain, both from a theoretical perspective and in their application to natural language understanding, reasoning, information visualization, and interoperability.

Songmao Zhang is currently a guest researcher at the Lister Hill National Center for Biomedical Communications, US National Library of Medicine. She obtained her PhD degree in computer science in 1992 at the Institute of Mathematics, Chinese Academy of Sciences where she is now an associate professor. Her research interests include ontology matching, knowledge representation, data mining, AI-based automatic animation, and natural language understanding.