Integrating the UMLS into an RDF-Based Biomedical Knowledge Repository

Kelly Zeng, M.S., Olivier Bodenreider, M.D., PhD
U.S. National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA
{zeng,olivier}@nlm.nih.gov

Background

As part of *Advanced Library Services* project at the National Library of Medicine, we are creating a very large *Biomedical Knowledge Repository* (BKR), which serves as background knowledge for applications including knowledge discovery and multi-document summarization [1]. The BKR integrates relations extracted from the biomedical literature (e.g., Medline citations) and from structured knowledge sources (e.g., Entrez Gene). It will also host relations contributed by external collaborators.

Effective warehouse knowledge integration requires the entities and relationships from various sources to be identified in reference to a common vocabulary. In this project, the Unified Medical Language System (UMLS) serves as the basis for identifying biomedical entities and relationships. For this reason, we first need to seed the BKR with UMLS concepts. Additionally, the UMLS is a source of terminological knowledge and we also want to integrate into the BKR the large set of terminological relations, symbolic and statistical, present in the UMLS Metathesaurus.

The current pilot BKR uses the Semantic Web technology RDF (Resource Description Framework) for its representation. In practice, the BKR is a large graph whose nodes are UMLS concepts and whose edges represent the relations extracted from various sources. In this abstract, we briefly discuss several aspects of the process of importing the concepts and relations from the UMLS into an RDF-based repository. Our goal is not to represent all the features present in the UMLS.

Importing concepts and relations

An RDF graph is composed of triples in which the subject and the object are linked by a predicate (relationship). In the MRREL.RRF relational table, UMLS relations are represented as tuples associating, among other things, one subject concept (CUI1), one object concept (CUI2) and one relationship. The precise nature of the relationship (RELA) is not always specified. When absent, we use the more generic REL instead (e.g., *parent of* instead of *is a*). The three elements of the RDF triple have to be represented by URIs (Unified Resource Identifiers). We create a URI for each UMLS concept based on the UMLS identifier (CUI). The relationships are currently converted into source-specific URIs. Relations in the Metathesaurus are represented bidirectionally, which is not required in the RDF graph. Therefore, only half of all UMLS relations are imported in the BKR.

Metadata

In addition to the knowledge represented in the RDF graph, we store information about the triples (i.e., metadata), including the source and version of the relations and timestamps for versioning purposes. We purposely store this information outside the RDF graph, in order to avoid using reified nodes in the RDF graph ("blank nodes").

Implementation

We use Oracle 11g (beta version) as the storage system. In practice, we use a utility program (loader) provided by Oracle to load the RDF triples and relevant metadata into the database. Technically, metadata are stored in relational tables, whereas native RDF triples are stored and managed internally by the Oracle database system.

Issues and challenges

The choice of RDF over other formalisms such as OWL (more expressive) or SKOS (less expressive) is a trade-off. We take advantage of rule bases in Oracle to support reasoning over the RDF graph. One important limitation is the absence of an ontology of biomedical relationships. Currently, relationships are organized according to the framework provided by the UMLS Semantic Network.

References

1 Bodenreider O, Rindflesch TC. Advanced library services: Developing a biomedical knowledge repository to support advanced information management applications. Technical report. Bethesda, Maryland: Lister Hill National Center for Biomedical Communications, National Library of Medicine; September 14, 2006. http://lhncbc.nlm.nih.gov/lhc/docs/reports/2006/tr200601.pdf