Adapting resources to the Semantic Web: Experience with Entrez Gene

Satya S. Sahoo[1], Olivier Bodenreider[2], Kelly Zeng[2], Amit P. Sheth[1]

[1]LSDIS Lab, Department of Computer Science, University of Georgia, Athens, GA, USA, [2]U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

{sahoo, amit@cs.uga.edu},{olivier, zeng@nlm.nih.gov}

Modern biomedical research is increasingly supported by information technologies. Biologists and physicians rely not only on the biomedical literature (e.g., MEDLINE), but also on the many knowledge bases available online (e.g., through the National Center for Biotechnology Information's (NCBI) Entrez portal). While these resources are undeniably valuable to humans, most of them are text-based and heterogeneous, and cannot be easily processed by computers. For example, the information retrieved from sources like Entrez Gene (EG) [2] or the Online Mendelian Inheritance in Man (OMIM) [4] is represented in XML but follows different data type definitions (DTD). Hence, queries across the different NCBI data sources are only possible through implementation of complex linkages. Moreover, within one data source namely EG, a traditional relational database schema makes it extremely difficult to query for information using the *relationships* between the concepts. The Biomedical Knowledge Repository (BKR) under development at the National Library of Medicine addresses these limitations. It consists of an extensive collection of normalized assertions (i.e., concept-relationship-concept triples), represented in a common format, and processable by computers. Therefore, it can be understood as a specialized version of the Semantic Web. In this paper, we describe a pilot contribution to the BKR: the transformation of the EG database into the W3C Resource Description Framework (RDF) format [5].

There are many issues involved in the conversion of XML data into RDF format, including using unique identifiers, preserving of the original semantics of the data being converted, resolving bidirectional relationships and filtering redundant element tags. Unlike traditional XML to XML conversion, converting XML to RDF should take into account the advantages of the RDF model in representing the logical structure of the information and the modeling of the relationships between concepts. The objective of converting XML data into RDF is to capture the semantics of the data and to make it available for querying the RDF graph, making it possible to retrieve not only the explicit knowledge, but also additional knowledge through inference.

We selected the eXtensible Stylesheet Language Transformation (XSLT) [6] for converting the EG XML information into RDF, because this approach allows for a clean separation between the application (using Java API for XML Processing (JAXP)) and the conversion logic (using XSLT stylesheet). The conversion rules used in this workflow are specific to the EG database.

We chose not to convert the element tags of the native EG XML representation mechanically into the *predicates* of the RDF triples. Instead, we manually converted the element tags of the native EG XML representation into meaningful relationship names that convey explicitly the semantics of the connection between the *subject* and the *object*. For example, the element *<Gene-track_geneid>* was mapped to the more meaningful relationship named '*has_unique_geneid*'. This relationship also captures the uniqueness of a '*geneid*' associated with each gene record in EG, only implicit in the XML representation.

Initially, we used one EG record to prototype our approach. We converted to RDF the EG record for the gene APP in *Homo Sapiens* (ID 351). To ensure syntactic accuracy of the RDF file generated, we used the W3C web-based RDF validating application (http://www.w3.org/RDF/Validator/). Once converted to RDF, the EG record for the APP gene comprises 16,180 triples. The 50 GB XML file for the complete EG data source (conforming to the EG DTD) was converted into a 39 GB RDF file. The primary reason for this reduction in size is that many elements from the original EG XML format were ignored or represented differently in RDF, including multiple redundant elements tags (e.g., *<Seq-loc_mix>*) and elements that formed multiple-layer containers around elements with actual attributes and values (e.g., *<Date>*). We used 105 elements out of the 133 unique element tags in an EG DTD. The RDF file was then converted into n-triple format (33 GB) (using the Jena toolkit [3]) for integration in the Oracle 10g [1] database. 411 million such triples were imported in the database.

Extending this work, we are now converting the Medical Subject Headings (MeSH) thesaurus into RDF. Once both are represented in RDF, MeSH and EG can be integrated seamlessly. Hierarchies in MeSH can then be exploited to query genes, enabling researchers to formulate queries such as "Find all genes involved with neurodegenerative diseases?".

Reference:

1. Alexander, N., Ravada S., "RDF Object Type and Reification in Oracle"—Technical White Paper (http://download-east.oracle.com/otndocs/tech/semantic_web/pdf/rdf_reification.pdf)

2. Maglott, D., Ostell, J., Pruitt, K.D., and Tatusova, T., "Entrez Gene: gene-centered information at NCBI", *Nucleic Acids Res.* 2005 January 1; 33(Database Issue): D54–D58.

3. McBride, B. 2002. Jena: A Semantic Web Toolkit. *IEEE Internet Computing* 6, 6 (Nov. 2002), 55-59.

4. Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (http://www.ncbi.nlm.nih.gov/omim/)

5. Resource Description Framework (RDF), http://www.w3.org/TR/2004/REC-rdf-primer-20040210/

6. XML Schema Language Transformation (XSLT), http://www.w3.org/TR/xslt