Management of Data, Knowledge, and Metadata on the Semantic Web: Experience with a Pharmacogenetics Knowledge Base

Diane E. Oliver, Micheal Hewett, Daniel L. Rubin,
Joshua M. Stuart, Teri E. Klein, and Russ B. Altman
Stanford Medical Informatics, Stanford University, Stanford, CA
oliver@smi.stanford.edu, hewett@smi.stanford, rubin@smi.stanford.edu,
stuart@smi.stanford.edu, klein@smi.stanford.edu, altman@smi.stanford.edu

Biomedical researchers are decoding the human genome with astonishing speed, but the clinical significance of the massive volumes of data collected remains largely undiscovered. Progress requires communication and data sharing among scientists. These data may be in the form of (1) raw data, derived data, and inferences that result from computational analyses, or (2) text documents published by experts who present their conclusions in natural language. The World Wide Web provides a valuable infrastructure for enabling researchers to share the rapidly growing knowledge about biology and medicine, and a fully functional Semantic Web is necessary to support data submission and retrieval, the sharing of knowledge, and interoperation of related resources.

We are working with scientists from multiple medical research centers to build a pharmacogenetics knowledge base that is publicly accessible on the Web, called PharmGKB (http://www.pharmgkb.org). The collaborative nature of our work demands building consensus data models for data submission and data retrieval. We need techniques that make it possible for both humans and machines to make sense of the data, methods for storing the data in a carefully designed knowledge base that contains up-to-date domain knowledge, and approaches for using metadata standards to express requirements for submitted data and inferred data.

The distinction between data and knowledge is often blurred, but for our purposes, **data** are the elements of information collected by researchers for experimental studies in a laboratory or clinical setting. Scientists submit these data to PharmGKB. PharmGKB is a knowledge base that is implemented in Protégé, a frame-based knowledge-management system (http://protégé.stanford.edu). PharmGKB contains an ontology comprised of clinical, pharmacologic, and biological knowledge. The domain concepts are arranged in a hierarchy of classes. Each class has slots that denote relations to other classes. Experimental data are represented by instances in the knowledge base. Thus, for our purposes, **knowledge** is conceptual information stored in the PharmGKB class-slot structure that depicts the domain. Finally, **metadata** is information about experimental data and derived data that may be imported to and exported from PharmGKB.

The architecture we have developed for PharmGKB involves seven components:

(1) **Data-entry forms** that allow people to visualize subsets of the data model and enter data
(2) An **XML processing system** that allows data to be submitted in a semantically-tagged machine-processable form
(3) A **knowledge base** that stores submitted data linked to the domain ontology
(4) An **ontology-editing environment** that developers use to extend the knowledge-base into new subdomains
(5) A **knowledge-base browsing and query tool** that permits users to view the contents of the knowledge base and to perform queries across sets of data

(6) **Standard interfaces** that enable applications to communicate with PharmGKB and that facilitate interoperation with other Web resources
(7) A **production system** and a **development system** that coexist and that must be merged on a periodic basis

The knowledge base is the central resource of the PharmGKB system that is accessible to users and applications over the Web. Some users prefer to submit data through Web forms because the forms guide them in entering certain data by hand, and for large datasets, they can upload tab-delimited files. After a user enters data through Web forms, the PharmGKB system automatically generates XML files from the data entered. However, other users, who know how to create XML files themselves, would rather avoid data entry through Web forms and prefer to submit XML files directly. In either case, the data requirements must be clear, and we have developed an XML schema that expresses metadata for data input.

The XML schema for data input must have a well-defined mapping to the domain ontology in the knowledge base, and it also must be consistent with the design of the Web forms. Thus, there are three levels of data representation that must be compatible: (1) the HTML form elements used in the user interface, (2) the XML schema, and (3) the ontology. We are working toward developing automated ways for keeping these three representations synchronized as the system evolves over time.

Evolution of the PharmGKB domain ontology is expected. When a standard relational database has a relatively stable database schema, evolution is primarily a matter of adding, deleting, and updating instance-level data, or field entries in the tables. In contrast, in PharmGKB, not only does instance-level data change as users submit data, but the PharmGKB ontology also evolves due to the addition of new subdomains and the growth of scientific knowledge. Thus, we maintain a production version to which users submit data, and a development version to which developers add ontological features. The ontology of the production version does not change continuously, but in steps or releases, while the ontology of the development version is more fluid. At each new release of the system, the development version and the production version must be merged. We are developing methods and tools for performing what can be a complicated merge process.

One of the goals of the PharmGKB project is to provide to the community data and knowledge that do not currently exist in other Web resources, and to facilitate interoperation with other databases in beneficial ways. For example, dbSNP is a public database that permits scientists to enter newly discovered data about single nucleotide polymorphisms (SNPs) in genes. PharmGKB users can submit SNP frequency data to PharmGKB, and if the SNP is not known to dbSNP, an application associated with PharmGKB can automatically make a SNP submission to dbSNP. PharmGKB also coordinates with Medline by storing citations for relevant published articles, and with GenBank, LocusLink, and Online Mendelian Inheritance in Man (OMIM) by storing relevant accession numbers. These links will make it possible to offer expanded services that require interaction with these other databases in the future.

Any knowledge-based system that serves as an evolving resource on the Web to which users from a particular community contribute, that interoperates with other Web resources, and whose underlying data and knowledge model is constantly changing, requires careful management of data, knowledge, and metadata. In our work on PharmGKB, we are confronting these challenges, and our experiences will be applicable to other similar resources that could be available and accessible on the Semantic Web.