OMICS A Journal of Integrative Biology
Volume 7, Number 1, 2003

The Need for Dictionaries, Ontologies, and Controlled Vocabularies

HELEN M. BERMAN and JOHN WESTBROOK

BIOLOGY HAS NOW SEEN an enormous increase in the amount and types of data collected. There are many types of databases ranging from archival collections containing the results of analyses from a broad field of biology to highly curated databases containing a great deal of information about a very small field of biology. In addition, laboratories routinely use LIMS to collect and organize experimental data. The need for sound informatics infrastructure is very high.

Any sound informatics infrastructure should provide for the following:

- *Acquisition:* Accurate collection/harvesting of information
- *Exchange:* Data exchange including format conversion with semantic precision
- *Dissemination:* Documentation, database schemas definition, and application program interface definition
- *Maximum software reusability*

Data dictionaries provide an important underpinning to the informatics infrastructure. The content of such a dictionary needs to contain the following:

- Precise definitions and examples are the most valuable elements
- Controlled vocabularies
- Allowed ranges and boundary conditions
- Data types
- Data relationships (parent-child/foreign key)

The dictionary content can be expressed in any/many concrete formats and is not bound to a particular technology. In creating such a dictionary the data model must be adequate to represent the particular subject matter. It must also be accessible to real users in that it must be easy to understand and extend, and it needs to be supportable with reusable software tools.

mmCIF IS AN EXAMPLE OF SUCH A DICTIONARY

There are examples of these types of dictionaries including Gene Ontology. The one used to support the Protein Data Bank (PDB, http://www.pdb.org) is the Macromolecular Information File (mmCIF, http://www.deposit.pdb.org/mmcif). The mmCIF dictionary was created as part of a community effort mandated by the International Union of Crystallographers (IUCr). It took several years to create the dictionary. The people involved had deep knowledge of the field and involved other experts in creating the dictionary definitions. The first version of the dictionary was published in 1996. It had 1700 definitions that spanned the terms used to describe the crystallographic experiment as well as the results of the experiment—the three-dimensional structure. The data model is reusable, extensible, and easily imported into standard relational database engines.

The Protein Data Bank, Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, New Jersey.

Because so much care was taken, both with the content of the dictionary content and the syntax, the data can now be expressed in alternate formats. Thus, an mmCIF can be expressed either in the legacy PDB format or in the more modern XML. In addition, the dictionary itself can be expressed as an XML schema.

This dictionary has proven to be an effective underpinning for every aspect of the PDB operation and is now serving as the basis for an Application Programming Interface and data exchange.

LESSONS LEARNED

The creation of the mmCIF dictionary was a voluntary effort. The creation of the data model and all the definitions were done by committed members of the community who were convinced that such an effort was required if we were ever to be able to cope with the large amounts of complex data that needed to be handled. Funds to support these sorts of initiatives will ensure that they go forward more efficiently and with the necessary peer review.

Address reprint requests to:
Dr. H. Berman
Department of Chemistry & Chemical Biology
Rutgers University
610 Taylor Road
Piscataway, NJ 08854

E-mail: berman@rcsb.rutgers.edu