A Unique Opportunity in Biological Information Object Standards

C.F. Dewey, Jr.[1,2], Aidan Downes[3], Howard Chu[3] and Shixin Zhang[1]

[1]Department of Mechanical Engineering
[2]Division of Biological Engineering
[3]Department of Electrical Engineering and Computer Science

Massachusetts Institute of Technology, Cambridge MA USA

Introduction

Over the past several years, the explosive growth of biological data generated by new high-throughput instruments has literally begun to drown the biological community. There is no established infrastructure to deal with these data in a consistent and successful fashion. This paper discusses the opportunity to develop a new informatics platform to handle a large subsection of the experimental protocols that currently exist. A consistent data definition strategy is demonstrated that handles gel electrophoresis, microarrays, fluorescence activated cell sorting, mass spectrometry, and microscopy within a single coherent set of information object definitions. Other experimental methods can be added with relative ease because the object model used to describe the data is easily extended.

The next step in the development of this platform is to enable simple access and use over the World Wide Web. The two keys to this deployment are: (a) establishing consistent ontologies accepted by the biological community and made available from repositories using OWL technology; and (b) using the rich descriptive capabilities of RDF to exchange data between repositories and users. These two technologies will enable broad data sharing and interoperability within the biological community.

Methods

Several important experimental techniques in contemporary biology have been used to create a single composite schema. The results bear a striking relationship to the DICOM standard of 1993 that provides information object definitions of all of the major medical imaging modalities (MR, CT, US, XA, NM, VL, CR, and Waveforms). The *de novae* information object definitions developed for gel electrophoresis by the authors of this paper were found to be very similar to the existing MAGE-OM information model for microarrays. Further investigation revealed that similar object definitions characterized other experimental biology methods as well. These were generalized and a full object-relational data schema was developed. The appended references cite a number of the proposed standards that were used to develop the object model.

Results

A first implementation of this work is called *ExperiBase*. It can store and query data generated by the leading experimental protocols used in biology within a single database. ExperiBase also has provisions to store derived data from analysis as a part of an expanded

definition of the information object. Transport of the raw data and analytical results between ExperiBase and external analysis packages currently uses web-based network technologies and XML representation of the data itself. The information object model is used to define the form of the XML data document. Import and export of data in spreadsheet format is also supported. ExperiBase has been ported to three leading database platforms: Oracle, DB2 and Informix. There are no platform-specific dependencies other than the necessity to support object models and large bindary data types in efficient native format. From an implementation point of view, the database in which ExperiBase is implemented should support sparse data matricies with no significant storage penalties.

Figure 1 is a high-level view of the organization of the data and metadata that, together, comprise a single experiment. Figure 2 is a condensed view of the expanded object model containing the elements used to describe each of the top-level concepts shown in Fig. 1.

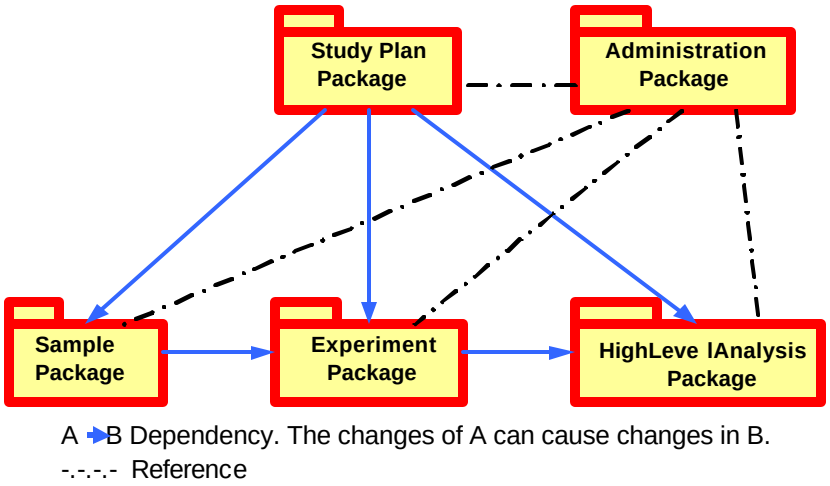

Figure 1 Biological experimental data can be grouped into five packages: study plan (also called as project), sample, experiment, high level analysis, and administration package.

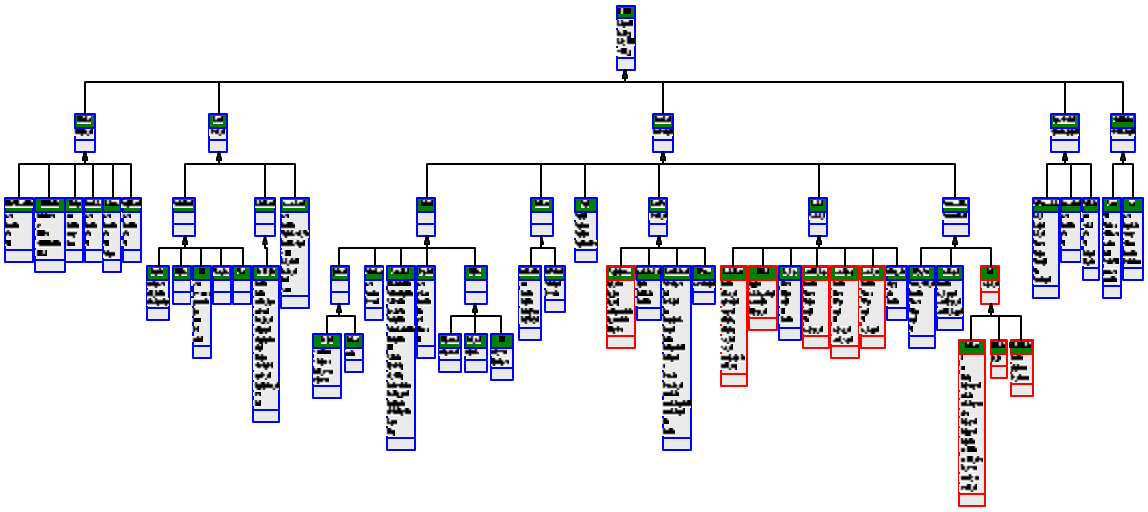

Figure 2 The information object definition of Western blot experiments incorporating the PEDRo schema. The red color represents the ideas coming from PEDRo that were not included in the original definitions. It also indicates that the IOD structure is extensible; new objects can easily be added into the schema without any changes of the structure.

Discussion

This work is being submitting to the World Wide Web Consortium (W3C) as a candidate application to assist in developing working models of the use of RDF and OWL in biological applications. Unique identifiers called LSIDs are used to tag each information entity so that the definitions can be traced to existing ontological compilations in OWL accessible using RDF-enabled applications. Community reuse of code and standards is one of the objectives of this work. RDF can be use as a rich transport mechanism between data sources and data users.

The Pacific Northwest National Laboratory and other sites have committed to using ExperiBase in their biological infrastructure. Emerging ontological standards in microarrays (MAGE-OM) and other proteomics efforts such as that proposed by the Human Proteome Organization (HUPO) either are already reflected in the code or can be easily incorporated.

Conclusion

The medical and biological communities are invited to participate in this effort to develop international standards to handle the massive data collections that are now being created in every pharmaceutical company and every academic biology laboratory. Having consistent formats for the information objects will greatly speed the development of analysis tools

Acknowledgements

This research was supported by the Defence Advanced Research Projects Agency and the Pacific Northwest National Laboratories (Department of Energy).

This paper was prepared for the World Wide Web Conference on Biological Uses of the Semantic Web, October 27-28, 2004 in Cambridge, MA, USA.

Background References

1. Chris F. Taylor, et al. A systematic approach to modeling, capturing, and disseminating proteomics experimental data. Nature biotechnology, March 2003, Volume 21, page 247-254. http://pedro.man.ac.uk. [Proposed ontology for mass spectrometry and 2D gel data that has been used as the basis of the ExperBase definition for these experimental methods.]
2. Gollub J, Ball CA, Binkley G, Demeter J, Finkelstein DB, Hebert JM, Hernandez-Boussard T, Jin H, Kaloper M, Matese JC, Schroeder M, Brown PO, Botstein D, Sherlock G. The Stanford Microarray Database: data access and quality assessment tools. Nucleic Acids Res 2003 Jan 1;31(1):94-6. http://genome-www.stanford.edu/microarray [Proposed ontology for Stanford microarray data that has been used as the basis of the ExperBase definition for this experimental method.]
3. MAGE-OM, Gene Expression RFP [Proposed ontology for microarray data that has been

used as the basis of the ExperBase definition for this experimental method.]
4. Lao H. Saal, Carl Troein, Johan Vallon-Christersson, Sofia Gruvberger, Åke Borg and Carsten Peterson. BioArray Software Environment: A Platform for Comprehensive Management and Analysis of Microarray Data. Genome Biology 2002 3(8): software0003.1-0003.6. http://base.thep.lu.se/. [Proposed ontology for BASE data that has been used as the basis of the ExperBase definition for this experimental method.]
5. CDISC laboratory data interchange standard, http://www.cdisc.org/pdf/Lab1-0-0-Specification.pdf. [Proposed ontology for user and laboratory data that has been used as the basis of the ExperBase definition for these elements.]
6. R.C. Leif, S.H. Leif, S.B. Leif, "CytometryML, An XML Format based on DICOM for Analytical Cytology Data ", accepted for publication Cytometry (2003). http://www.newportinstruments.com/cytometryml/cytometryml.html. [Proposed ontology for flow cytometry data that has been used as the basis of the ExperBase definition for this experimental method.]
7. Swedlow JR, Goldberg I, Brauner E, Sorger PK. Informatics and quantitative analysis in biological imaging. Science . 2003;300(5616):100-102. http://www.openmicroscopy.org/index.html. [Proposed ontology for optical microscope image data that has been used as the basis of the ExperBase definition for this experimental method.]
8. Web Ontology Language (OWL). See the OWL Home Page.
9. Semantic Web Activity Statement. See Semantic Web Activity Statement.