OMICS A Journal of Integrative Biology
Volume 7, Number 1, 2003
© Mary Ann Liebert, Inc.

Sharing Biomedical Data with Impunity and Ease

SUSAN B. DAVIDSON

BIOLOGICAL DATA is increasingly being shared between databases using a variety of different data formats. While some of these formats have been developed almost exclusively for communicating between viewers and databases (e.g., AGAVE and GAME) and others for sharing annotation as well as viewing (e.g., DAS), many formats are also being used to exchange data between databases (e.g., EMBL, ASN.1, and specialized XML DTDs).

The appeal of a data exchange format is that it is a way of serving data in a uniform, flexible, and easily parseable form. Data exchange formats are also largely self-describing and hence easy to understand. However, agreeing to use a specific exchange format does not solve the data exchange problem by itself. Exporters must *map* their data into the exchange format, and importers of data must again map from the exchange format into their local format (or more generally, model). Thus data exchange is inextricably tied up with writing mappings (or transformations) between data formats. Several problems are associated with writing mappings between data formats. First, it is an inherently difficult problem. The writer of the mapping must understand both how the data is being represented in the exchange format and how they are representing it in their own model. Second, semantic information is frequently not captured in a data exchange format. For example, information about keys, foreign keys and constraints is often omitted. Clearly, constructing a mapping must be guided by an understanding of the semantics of the data since otherwise the mapping may cause run-time constraint violations.

As an example of these problems, we having recently been conducting an experiment involving exchanging microarray gene expression data using a developing standard called MAGE-OM/ML (Spellman et al., 2002). The semantics of MAGE is specified using UML modeling tools (MAGE-OM), however the exchange is effected using an XML representation of the standard. Prior to this standardization effort, a relational database called RAD (On-line data) had been developed at the Penn Center for Bioinformatics to store gene expression data as well as its associated sample annotation data. When the MAGE-ML standard is finalized, data will be imported from collaborators and exported from RAD using this format. However, each of these data representations—RAD and MAGE—has been developed independently. In our experiment, the following problems emerged:

1. The data exported by RAD into MAGE-ML through some transformation may fail to validate against the constraints of MAGE-OM.

Example: Gene expression annotation in MAGE-ML requires information about the process of sample preparation. The annotation interface in RADv2 required only information about the end result of the sample preparation, with an (optional) free-text description of the process. Since in MAGE-ML the biomaterial can only exist if the biosource is present (and so on down the process), the sample information in RADv2 was inconsistent with MAGE-ML. RAD was therefore modified so that RADv3 is consistent with MAGE-ML, and a new annotation interface is under development to force the process to be captured.

2. The data imported by RAD through some transformation from MAGE-ML may violate integrity constraints in RAD. If the MAGE-ML data is consistent with respect to the constraints of MAGE-OM, then there must be some inconsistency between MAGE-OM and the constraints expressed in RAD.

Center for Bioinformatics, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, Pennsylvania.

Example: In the Experiment package in MAGE-ML, an Experiment has a unique ExperimentDesign, which can have many associated types (e.g., "time course" and "normal vs. diseased"). In RADv2, there is a single relation Groups(Group_ID, Group_Type, Description, Name), which corresponds to the Experiment class. However, this is incorrect since there could be many different types associated with an Experiment rather than the single one implied in the relational design. RADv2 is therefore being re-designed to correct this inconsistency.

The examples above identify two situations that have caused a re-design of RAD. But are these the only problems that will be encountered? Rather than recognizing inconsistencies through an ad-hoc process and laboriously going through successive redesigns of RAD to deal with them, it would be extremely helpful to have a framework in which, given a desired mapping of data and given existing constraints, all ensuing inconsistencies could be automatically exposed and corrections suggested.

In performing this experiment, we have also re-affirmed the problems of performing mappings between data sources. The MAGE standard is specified in a 125-page document, of which roughly 86 pages are necessary for understanding the model. There are 17 packages, each of which has between 3 and 20 classes. The RAD schema has roughly 6 high-level divisions and 112 tables. Understanding both of these representations and specifying the mapping on the Experiment package of MAGE took the student working on the project several months, and this mapping will have to be re-adjusted as both MAGE and RAD are still evolving (a common problem in bioinformatics). Furthermore, many of the mappings involved functions on data fields rather than simple correspondences; for example, a string in RAD must be parsed to capture individual elements that are mapped to different classes in MAGE. Techniques for automatically inferring potential connections between classes in the models and improved mapping techniques would be extremely helpful. Some work in this area has already begun (Yan et al., 2001; Doan et al., 2003).

REFERENCES

SPELLMAN, P.T., MILLER, M., STEWART, J., et al. (2002). Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biology **3(9),** research 0046.1–0046.9.

On-line data. Available: www.cbil.upenn.edu/RAD2.

DOAN, A.H., DOMINGOS, P., and LEVY, A. (2003). Learning to match the schemas of databases: A multistrategy approach. Machine Learning Journal, **50 (3),** 279–301.

YAN, L.L., MILLER, R.J., HAAS, L.M., and FAGIN, R. (2001). Data-Driven Understanding and Refinement of Schema Mappings. *Proceedings of the ACM SIGMOD International Conference on Management of Data.* (Santa Barbara, CA: Computing Machinery), 485–496.

Address reprint requests to:
Dr. Susan B. Davidson
Center for Bioinformatics
Department of Computer and Information Science
University of Pennsylvania
200 S. 33rd Street
Philadelphia, PA 19104-6389

E-mail: susan@cis.upenn.edu