

The Multiple Roles of Ontologies in the BioMediator Data Integration System

Peter Mork^{1,2}, Ron Shaker³, and Peter Tarczy-Hornoch^{2,3}

¹ Computer Science & Engineering

² Biomedical & Health Informatics

³ Pediatrics, University of Washington,
Seattle, WA 98011, USA

`pmork@cs.washington.edu`

`{rshaker, pth}@u.washington.edu`

Abstract. BioMediator is a data integration system that provides a common interface to multiple Internet-accessible databases containing information about genetics and molecular biology. Ontologies play several important roles in the BioMediator system: First, ontologies of genetics and molecular biology can serve as *data sources*. In this role concepts from the ontologies are returned as results of queries. Second, queries are posed against a *mediated schema*, which is an ontology describing the domain of discourse. User queries are expressed using the concepts in the mediated schema to indicate which results to retrieve. Third, each data source is an instance of the *system ontology*. This ontology describes information about the data sources including how often the source is updated and by whom. Finally, we are exploring the use of ontologies as a mechanism for *mapping* data sources to the mediated schema. This will facilitate extending BioMediator from a centralized integration platform to a distributed network of peers.

1 Introduction

Biologists seeking to understand the molecular basis of human health and disease are struggling with large volumes of diverse data (mutation, expression array, proteomic) that need to be integrated and analyzed in order to develop and test hypotheses about disease mechanisms and normal physiology. These data reside in multiple public and private databases maintained by biologists in their laboratories. For example, a set of experiments may generate both gene and protein expression data, which are queried in aggregate to find a set of expression products of potential interest. Each of these products is, in turn, queried against public domain databases such as Entrez [1], SwissProt [2], and the Gene Ontology [3]. Given the dynamic nature of the datasets federated database approaches provide advantages over warehousing approaches in terms of data currency. Federated approaches with flexible mediated schemata representing the entities of interest and their mappings to particular sources are well-suited to handle the diverse schemata necessary, particularly for the laboratory specific private data sets. The BioMediator data integration system [4, 5] takes an ontology driven federated approach to data integration for these reasons.

In this paper we present an overview of the BioMediator system emphasizing the various roles that ontologies (a term we use loosely to refer to vocabularies such as the Gene Ontology, a database schema, or a terminology expressed in a description logic such as OWL) play in the system. At the source level, the schemata of sources focused on data (e.g., Entrez) and those focused on concepts (e.g., the Gene Ontology) are treated identically by our system and knowledge about the structure and organization of both types of sources can be represented as ontologies (3.1). At mediated level, the schemata used to query across these sources are also represented as ontologies (3.2). We permit multiple mediated schemata customized to different users/query tasks, pieces of which can be shared and reused. At a meta-level the BioMediator system uses a system ontology (3.3) to describe meta-information about the sources (such as information about validation and curation). Finally, we are developing techniques for translating data from specific source schemata into a mediated schema using knowledge stored in a mapping ontology (3.4).

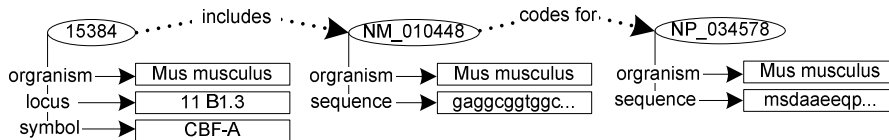


Fig. 1. Sample data viewed as a network of resources and properties; solid lines indicate datatype properties (DTP) and dotted lines, object properties (OP)

2 Background

In BioMediator, the data contained in online public databases are viewed as a network of interconnected records. For example, Online Mendelian Inheritance in Man (OMIM) [6] contains records describing genes and genetic diseases. Entrez publishes records that describe proteins and nucleotide sequences. Entrez also cross-references its protein records with related OMIM records.

2.1 Semantic Web Data Model

The data sources thus constitute a semantic web for the life sciences. In this web, each record corresponds to a node with a collection of attribute/value pairs. This is illustrated in Figure 1. The Entrez node NM_010448 has two solid edges leading from it: the organism edge indicates it pertains to the house mouse, and the sequence edge indicates the nucleic acid sequence. Expressed using RDF [7] terminology, this record is a *resource* with two *datatype properties* that link the resource to values.

LocusLink (LL) [8] provides other information related to this nucleotide sequence. LL resource 15384 describes the CDF-A gene. Also, LL publishes an *object property* that links the LL resource to the Entrez resource. This establishes that one possible sequence for the CDF-A gene is described by the indicated Entrez resource.

We distinguish between datatype properties (DTP) and object properties (OP) for two reasons. First, DTPs indicate the actual content of a resource. DTPs capture what

can be thought of as the information represented by the resource. OP are correspondences between resources; they are typically displayed (in a web browser) as hyperlinks. The second distinction pertains to ownership. In BioMediator, each resource is owned by a single data source and only that source can provide DTPs for that resource. OPs, on the other hand, can be provided by any data source. For example, not only does LL provide a property linking CDF-A to a sample sequence (NM_010448), but LL also links this nucleotide sequence to a corresponding protein (Entrez record NP_034578) using RefSeq [9].

Viewing the data sources as a semantic network distinguishes BioMediator from other data integration projects (such as Kleisli [10] or OPM [11]). The semantic network paradigm facilitates organizing the resources with an ontology. This approach was pioneered (for biologic domains) by TAMBIS [12] and, as we describe in this paper, extended by BioMediator. In this context, the ontology organizes the resources (and properties) into a hierarchy of concepts, against which users can query.

2.2 System Interface

BioMediator allows client programs to interact with this semantic web in a number of ways. The most basic interaction, *seed*, retrieves a specific resource and its associated DTPs. The client program provides the resource's accession number, and the database in which the resource can be found. For example, a program can request resource NM_010448 from Entrez, and BioMediator will retrieve the associated attribute/value pairs (e.g., organism/Mus musculus). Microarray researchers with chips annotated using accession numbers use this operation extensively [13].

Resources can also be retrieved using a *query*. In this case, the client program selects one of the classes in the mediated schema (see below) and one or more attribute/value restrictions. BioMediator retrieves all of the resources that are instances of the given class and that include all of the indicated attribute/value pairs. For example, a program can request all phenotype resources whose name is narcolepsy, or genes whose locus is 11 B1.3 and whose organism is the house mouse.

These first two interactions produce DTPs only. OPs can be retrieved using *expand*. Given a resource (or set of resources), this operation retrieves all related OPs (either leading from or pointing to the indicated resource). Both the mediated schema and the system ontology (see below) can be used to restrict which OPs will be retrieved. For example, a client program might be interested in the 'codes-for' property for a sequence, but not the more general 'related-to' property.

Finally, BioMediator can recursively *grow* the network, which expands each new resource it encounters. In this case, it is often useful to limit the OPs using the system ontology (e.g. limiting the growth to include only externally validated properties).

2.3 Architectural Overview

To support these operations, BioMediator relies on a series of components as illustrated in Figure 2. The system relies heavily on the source knowledge base (SKB), which is represented using Protégé-2000 [14], and accessed via the Protégé API. The SKB (Fig. 2A) contains the mediated schema and the system ontology, both of which are described in the following section.

The *query processor* (Fig. 2B) provides an API for launching and managing queries posed using the mediated schema. The *metawrapper* (Fig. 2C) translates these mediated schema queries into source specific queries [15]. *Wrappers* (Fig. 2E) pass the remapped queries through to the *data sources* (Fig. 2F). Data sources return results in native format (e.g., HTML, ASN1), which are converted to XML syntax with native semantics by the wrappers. The metawrapper applies mapping rules in translating the XML results from native semantics to mediated schema semantics.

The query processor then retrieves data from the metawrapper, organizes that data and generates events that can be used to synthesize a navigable representation of the result set. Once a result set has been constructed, it may be repeatedly queried, expanded or grown using the query processor's API.

3 Multiple Roles of Ontologies

As described in the previous section, BioMediator uses ontologies in several roles. The SKB contains two ontologies: The mediated schema provides a hierarchical vocabulary for organizing resources published by the underlying data sources and the system ontology describes how the data sources are maintained. In addition, BioMediator can access external ontologies as data sources.

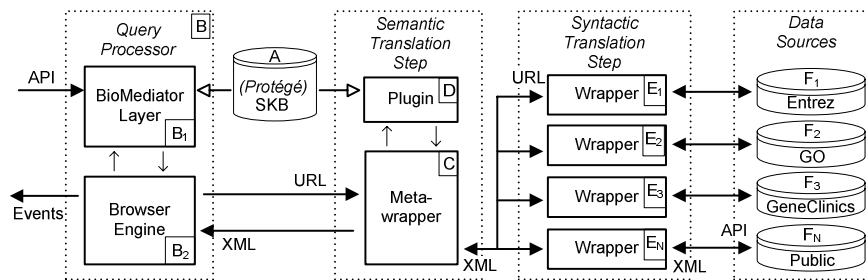


Fig. 2. Architecture pipeline of the BioMediator system

3.1 Data Source

In many cases, an ontology can be represented in the semantic web data model. In this case, resources represent named concepts and properties represent relationships among the concepts. For example, the Gene Ontology (GO) [3] includes two inter-concept properties ('is-a' and 'part-of') and one property relating external resources to concepts ('classified-as').

Properties provided by an ontology are treated no differently than other properties. This means that, for better or worse, we do not attribute any meaning to these properties. For example, given that the nuclear membrane is part of the nucleus, and the nucleus is part of the cell, we should be able to infer that the nuclear membrane is part of the cell. Instead of making this inference, BioMediator returns only those properties explicitly present in the sources.

This simplicity is advantageous because properties relating data resources and properties relating concept resources are treated uniformly. For example, given a collection of nucleotide sequences up-regulated in an experimental group (relative to a control), BioMediator can first identify the corresponding proteins (using the expand operation) and then organize these proteins based on functional classification (using the expand operation a second time). This helps a researcher answer the question, “What do these experimental results mean?”

When a simplistic view of the data is not sufficient (*e.g.*, a user needs to answer a very precise question), more machinery is needed. In this case, the mediated schema provides a common vocabulary for expressing more precise interactions (such as “A mutation of what gene results in dysprothrombinemia, haemophilia caused by an inactive protein?”).

3.2 Mediated Schema

At the heart of a data integration system is a mediated schema. The simplest mediated schema is the union of the source schemata which has two key limitations. First, application developers must understand all of the source schemata to author queries. Second, when a new source is added, each application needs to be modified to reference the new source. For example, both SwissProt [2] and Entrez [1] contain information about proteins. In the absence of a mediated schema, the only way to capture this similarity is by requiring all applications to query for the union of these sources. When another source containing information about proteins (*e.g.*, GeneTests [16]) is identified, every application program must be updated.

Given this limitation, database research has focused on formalisms for expressing the mediated schema in terms of the source schemata. In TAMBIS [12], the mediated schema is an ontology expressed using the GRAIL description logic [17]. The mediated schema is described independently of the underlying sources. The contents of the sources are then described in terms of the mediated schema, and an inference engine is used to determine how the source schemata relate to the mediated schema. For example, an OMIM record can be defined to be the union of genes and phenotypes for which the value of the organism attribute is human.

When a new source is added to the system, neither the existing definitions need to be updated, nor do existing applications. As a result, new sources can more transparently be introduced into the system. However, if the mediated schema is changed, then it becomes necessary to revisit every definition.

BioMediator uses a strategy similar to TAMBIS, but with greater emphasis placed on modularity. Instead of a single mediated schema, one of our goals is to support multiple mediated schemata simultaneously. In Figure 2, each user group can have its own SKB, independent of all other user groups.

Thus, even though the users see the same sources, they may organize these sources differently. One sample mediated schema is shown in Figure 3. This schema was developed for a statistician performing analyses on microarray data (*i.e.*, it is not intended to represent everything about microarray experiments, let alone all of molecular biology). Several concepts in Figure 3 are common to a variety of user

groups: Genes are an abstract unit of inheritance. Each gene can include a number of closely related sequences as examples of the gene. These sequences code for proteins, which produce (cause) the manifestation of a phenotype.

Some additional concepts are needed to support microarray analyses. First, we added several classes that describe microarrays. An experiment is performed using a specific chip. That chip contains several spots. Each spot is associated with a specific sequence. The statistical analyses also required functional information (from GO), which was one of the motivations for treating ontologies as data sources. Here GO is modeled as a hierarchical vocabulary, which differs from a controlled vocabulary in that inter-concept properties are allowed (as described above).

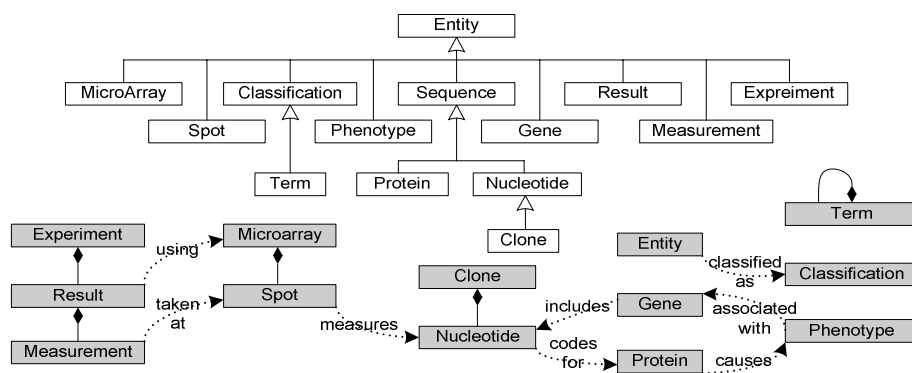


Fig. 3. Sample mediated schema for annotating microarray experiments. The top half displays the inheritance hierarchy; the bottom half displays containment relationships (diamonds) and other object properties

Once the mediated schema has been designed, rules must be written so that the metawrapper can transform source data into the mediated namespace. When multiple groups agree on portions of the mediated schema, they can also share these transformation rules. In the case of disagreement, transformation rules must be modified or removed. Finally, each source must be added into the system ontology.

3.3 System Ontology

Within the system ontology, each data source is represented as an instance of the database class. A database is a collection of resource tables and property tables. Of the resource tables, one is designated as the primary table (references into a database that do not specify a type are assumed to index into the primary table).

A resource table stores the metadata needed to retrieve a collection of resources. Each *resource table* is associated with a class from the mediated schema; all resources in the table are instances of that class. Likewise, each *property table* is associated with a property from the mediated schema. The domain and range of each property table must also be specified (*i.e.*, the resource tables connected by the property table).

Moreover, for each property table, we also record metadata describing how the property table is maintained. These metadata include descriptions of: a.) population, b.) validation, c.) update, and d.) causality (i.e., whether the correspondance indicates a causal mechanism, such as gene coding for protein vs. merely observed correlation).

Metadata can be used to constrain the property tables that will be considered when using the expand or grow operations. For example, a clinician might be interested in browsing only those property tables ‘validated’ by an external review process, whereas a researcher might choose to browse only ‘causal’ relationships (even if the relationship has not yet been proven experimentally).

Each table is also associated with rules used by the metawrapper to convert source data into the BioMediator data model. For example, a rule is used to indicate that when OMIM returns a disease record it should be converted to a resource that is an instance of the mediated class phenotype. The value of the title attribute is mapped to a name datatype property.

3.4 Mappings

We have begun exploring OWL [18] as an alternative to the current rule language for expressing relationships between the source schemata and the mediated schema. The hope is that OWL constructs will allow us greater flexibility. Not only will it be possible to translate from a source namespace to the mediated, but the inverse will also be possible. This will allow us to distribute our system in a peer-to-peer fashion.

For example, we can declare that an OMIM record describes a gene or a phenotype, *i.e.*, an OMIM record is defined to be the union of these two classes. A GeneTests record for a gene is equivalent to the class, Gene, in the mediated schema. A query requesting information about a specific gene can be *rewritten* as a query against GeneTests (because $\text{Gene} \equiv \text{GeneTests Gene Record}$).

More sophisticated rewritings are also possible. At first, it does not seem that a gene query can use OMIM because an OMIM record is more general than gene ($\text{Phenotype} \subseteq \text{Gene} \cup \text{Phenotype} \equiv \text{OMIM Record}$). However, assume the mediated schema asserts that the domain of the property, AssociatedWith, is $\text{NucleotideSequence} \cup \text{Gene}$, we can rewrite the query to request OMIM records that participate in the AssociatedWith property ($\text{OMIM Record} \cap \geq 1 \text{ AssociatedWith}$). We are exploring algorithms for efficiently generating all valid rewritings.

4 Conclusions

BioMediator is a data integration system that uses ontologies in several roles. The network-based data model allows us to use an ontology such as the Gene Ontology as a data source. This is particularly useful for organizing experimental results into functional groups. To support more precise interactions, users can formulate queries in terms of a mediated schema. The role of this mediated schema is to provide a common nomenclature applicable to multiple local or remote data sources. The mediated schema also defines the object properties that can link data instances. These properties are further annotated using the system ontology, which describes how the

underlying data sources are maintained. This approach provides several benefits. First, the results returned by BioMediator are as current as the underlying sources. Second, each user group can customize its mediated schema, and the mappings that relate the data sources to that common namespace. Finally, our architecture supports both precise queries (the database standard) and more generic browsing. These advantages make BioMediator an excellent platform for supporting a variety of biomedical data needs.

Acknowledgments

We would like to thank Hao Mei (microarray mediated schema) and Scott Brockenbrough (editorial comments). Funding: R01HG02288 & T15LM07442.

References

- [1] Entrez search and retrieval system. National Center for Biotechnology Information, National Library of Medicine [Online]. Available: <http://www.ncbi.nlm.nih.gov/Entrez/>
- [2] Swiss-Prot Protein knowledgebase. Swiss Institute of Bioinformatics [Online]. Available: <http://us.expasy.org/sprot/>
- [3] Gene Ontology™ Consortium (GO). Gene Ontology (GO).[Online]. Available: <http://www.geneontology.org/>
- [4] Mork, P., Halevy, A. Y., Tarczy-Hornoch, P.: A Model for Data Integration Systems of Biomedical Data Applied to Online Genetic Databases. In: Proc. Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium (2001) 473–477
- [5] Tarczy-Hornoch, P., Halevy, A. Y., Rossini, A. J., Mork, P., Shaker, R., Donelson, L. BioMediator: A Data Integration System for Biomedical Databases. University of Washington [Online]. Available: <http://www.biomediator.org>
- [6] Online Mendelian Inheritance in Man, OMIM™. McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine [Online]. Available: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM>
- [7] Brickley, D., Guha, R. RDF Vocabulary Description Language 1.0: RDF Schema. World Wide Web Consortium (W3C®) [Online]. Available: <http://www.w3.org/TR/rdf-schema/>
- [8] LocusLink. National Center for Biotechnology Information, National Library of Medicine [Online]. Available: <http://www.ncbi.nlm.nih.gov/LocusLink/>
- [9] Pruitt, K. D., Maglott, D. R.: RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Research* 29 (2001) 137–140
- [10] Chung, S. Y., Wong, L.: Kleisli: a new tool for data integration in biology. *Trends in Biotechnology* 17 (1999) 351–355
- [11] Chen, I.-M. A., Markowitz, V.: An Overview of the Object Protocol Model (OPM) and the OPM Data Management Tools. *Information Systems* 20 (1995) 393–418
- [12] Baker, P. G., Brass, A., Bechhofer, S., Goble, C. A., Paton, N., Stevens, R.: TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. In: Proc. Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology (1998) 25–34

- [13] Mei, H., Tarczy-Hornoch, P., Mork, P., Rossini, A. J., Shaker, R., Donelson, L.: Expression array annotation using the BioMediator biologic data integration system and the Bioconductor analytic platform. In: Proc. Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium (2003) 445–449
- [14] Musen, M., Crubézy, M., Ferguson, R., Noy, N. F., Tu, S., Vendetti, J. The Protégé Ontology Editor and Knowledge Acquisition System. Stanford Medical Informatics [Online]. Available: <http://protege.stanford.edu/>
- [15] Shaker, R., Mork, P., Barclay, M., Tarczy-Hornoch, P.: A Rule Driven Bi-Directional Translation System Remapping Queries and Result Sets Between a Mediated Schema and Heterogeneous Data Sources. In: Proc. Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium (2002) 692–696
- [16] GeneTests. University of Washington [Online]. Available: <http://www.genetests.org/>
- [17] Zanstra, P. E., van der Haring, E. J., Flier, F., Rogers, J. E., Solomon, W. D.: Using the GRAIL language for Classification Management. In: Proc. Fifteenth International Congress of the European Federation for Medical Informatics (1997) 897–901
- [18] Web-Ontology (WebOnt) Working Group. OWL Web Ontology Language Overview. World Wide Web Consortium (W3C®) [Online]. Available: <http://www.w3.org/TR/owl-features/>