An experiment in integrating large biomedical knowledge resources with RDF: Application to associating genotype and phenotype information

Satya S. Sahoo[1], Olivier Bodenreider[2], Kelly Zeng[2], Amit Sheth[1]
[1]The Kno.e.sis Center, Wright State University, Dayton, OH USA
[2]U.S. National Library of Medicine, NIH, Bethesda, Maryland, USA
{sahoo.2, amit.sheth}@wright.edu, {olivier, zeng@nlm.nih.gov}

ABSTRACT

Bridging between genotype and phenotype is generally achieved through the integration of knowledge sources such as Entrez Gene (EG), Online Mendelian Inheritance in Man (OMIM) and the Gene Ontology (GO). Traditionally, such integration implies manual effort or the development of customized software. In this paper, we demonstrate how the Resource Description Framework (RDF) can be used to represent and integrate these resources and support complex queries over the unified resource. We illustrate the effectiveness of our approach by answering a real-world biomedical query linking a specific molecular function, glycosyltransferase, to the disorder congenital muscular dystrophy, which potentially forms a new hypothesis. Some challenges encountered along the way are discussed, namely issues with the identification of biomedical entities and the lack of a reference ontology of relationships.

Categories and Subject Descriptors

H.3.3 [Information Systems] Information Search and Retrieval, H.1.m [Miscellaneous]

General Terms

Experimentation, Standardization

Keywords

Data integration, Semantic Web, Resource Description Framework, Entrez Gene, Gene Ontology, SPARQL, path queries.

1. INTRODUCTION

Integrating multiple heterogeneous knowledge sources has become a necessity in many domains, yet still represents a major challenge to both domain experts (e.g., biologists) and computer scientists.

The interpretation of experimental data generally requires physicians and biologists to compare their clinical and biological data to already existing data sets and to reference knowledge bases. However, most biomedical systems have been developed independently of each other, and, as a result, they do not have a common vocabulary or structure that would facilitate navigation across resources [1]. The integration of biomedical resources has been proposed as a solution to facilitate access to multiple, heterogeneous resources [2].

Information integration is also one of the most challenging area of research in Computer Science [3]. The use of heterogeneous schemas designed primarily to ensure optimization of storage space makes it extremely difficult for users to query data sources in an integrated manner. However, recent research in Semantic Web technologies has delivered promising results to enable information integration across heterogeneous knowledge sources. In effect, the Semantic Web provides a common framework that enables the integration, sharing and reuse of data from multiple sources. Additionally, the use of a representation formalism based on a formal language enables software applications to reason over information.

In this paper, we discuss the use the Semantic Web technology RDF (Resource Description Framework) for integrating two heterogeneous data sources frequently used in genomic studies: Entrez Gene and the Gene Ontology. We describe an experiment in integrating and querying these resources and present an application to hypothesis formulation in biomedicine. The objective of this experiment is not to match the schemas of the two resources, but rather to leverage the presence of entities common to both resources. The resulting integrated resource is shown to support complex queries (e.g., representing hypotheses) that could not be answered by any of the resources taken separately.

Finally, we discuss some of the challenges encountered along the way, namely the need for formalizing the predicates that relate entities in order to support reasoning, as well as the need for unique identifiers for entities across data sources and standard identification schemes for biomedical entities.

2. BACKGROUND

The Resource Description Framework (RDF)

The Resource Description Framework (RDF) is a W3C-recommended framework for representing data in a common format that captures the logical structure of the data [4]. This is in contrast to pure storage aspects addressed by traditional relational database schema. The RDF representational model uses a single schema in contrast to multiple heterogeneous schemas or Data Type Definitions (DTD) used by different sources to represent

WWW 2007, May 8--12, 2007, Banff, Canada.

data in XML. All data represented in RDF form a single knowledge repository that can be queried as one knowledge resource. An RDF repository consists of a set of assertions or triples. Each triple consists of three entities namely, the *subject* – the triple pertains to this entity, the *object* – the entity that states something about the object and the *predicate* – the relationship between the *subject* and the *object*.

Integrating biomedical data sources

Three main approaches to integrating heterogeneous, distributed data sources in the biomedical domain have been proposed [2] warehouse integration, navigational integration and mediator-based integration.

In the *warehouse integration* approach, data are imported from various sources and stored locally in a unique format. Data are transformed as necessary in order to make the various sources compatible with each other. Queries are made directly to the warehouse. GUS [5] is an example of warehouse for genomic sources such as Swiss_Prot and GenBank.

With *mediator-based integration*, data sources are queried remotely rather than stored locally. Queries, not data, are adapted to the specific characteristics of each source. A mapping is established between the schema of each source and a global schema representing the integrated resource. Examples of mediator-based integration systems include TAMBIS [6] and BioMediator [7].

The *navigational integration* approach focuses on links between sources, provided by the sources (e.g., cross-references) or specifically generated (e.g., BLAST similarity). The resulting system is, in effect, a graph in which the various entities are linked by paths, making it possible for users to navigate between resources. Example of navigational systems include Entrez [8].

In this paper, we propose an integrative approach to querying across knowledge sources based on the Resource Description Framework (RDF) [4]. Our approach shares some features with the traditional approaches presented earlier. Like warehouse integration, it requires the various data sources to be converted into a common format, here RDF. Our current implementation also integrates the sources into a unique store. The RDF store constitutes a large graph and is therefore similar to the underlying structure of the systems based on navigational integration. Finally, our approach relies on ontologies to support inference, which of also a feature of many mediator-based systems.

Related work

The creation of a permanent repository was selected over queries made "on the fly" to the resources mainly because this integrated RDF repository constitutes the central knowledge resource of a larger project. In fact, this work is a pilot contribution to the *Biomedical Knowledge Repository* under development at the U.S National Library of Medicine as part of the *Advanced Library Services* project [14]. This repository integrates knowledge not only from structured resources (database and knowledge bases), but also from the biomedical literature (e.g., MEDLINE), in order to support various applications, including knowledge discovery.

Integrating biomedical knowledge resources through RDF is also one of the goals of the BioRDF task force of the W3C Semantic Web Health Care and Life Sciences Interest Group (http://www.w3.org/2001/sw/hcls/). In a recent paper, they described how RDF-integrated resources can support translational research in the domain of neurosciences [9]. Example of applications using RDF to integrate biomedical knowledge sources include YeastHub [10] and LinkHub [11], Taverna [12] and BioDASH [13].

Biological example

A common scenario in biomedical research involves the correlation of genomic data with disease information, in other words, associating genotype and phenotype information. In the particular scenario illustrated in this paper, a researcher is interested in glycosylation and its implications for one disorder: congenital muscular dystrophy. The biological process of glycosylation results in the post-translational addition of glycosyl groups (saccharides) to proteins (and lipids). Various enzymes, namely glycosyltransferases, catalyze glycosylation reactions.

From the functional annotation of gene products with terms from the Gene Ontology (GO), a researcher can identify the genes having the molecular function of catalyzing the transfer of specific glycosyl groups (e.g., *hexosyltransferase*, for hexosyl groups). Known associations between these genes and diseases can then be mined from resources such as NCBI's Entrez Gene (EG), where phenotypic information is recorded as pointers to the Online Mendelian Inheritance in Man (OMIM) knowledge base [15].

In order to validate the hypothesis of possible association between the molecular function *glycosyltransferase* and the disease *congenital muscular dystrophy*, a researcher could simply search EG for the term *glycosyltransferase*, and all records containing the string "glycosyltransferase" in GO annotations would be returned. This approach, however, is suboptimal for at least two reasons. First, the term *glycosyltransferase* might appear as a substring in other GO terms (e.g., in *UDP-glycosyltransferase*), possibly leading to false positives. Conversely, not all GO terms related to *glycosyltransferase* actually contain the string "glycosyltransferase" (e.g., *acetylglucosaminyltransferase*, a kind of *glycosyltransferase*), possibly leading to false negatives.

To avoid false positives and false negatives, a careful researcher would likely start exploring the Gene Ontology database to create a list of *glycosyltransferase*-related terms by selecting the term *glycosyltransferase* itself (GO:0016757) and all its descendants, including specialized types of *glycosyltransferase*, such as *acetylglucosaminyltransferase*. This researcher would then look for the genes annotated with any of the *glycosyltransferase*-related terms. Resources such as the web browser AmiGO [16] support such searches and can retrieve the genes associated with any descendant of a given GO term. Finally, each of the genes found associated with any of the *glycosyltransferase*-related terms must be searched individually in EG, looking for mentions of the disease *congenital muscular dystrophy* (as an OMIM phenotype) in the corresponding records.

The procedure described above is evidently inefficient, time consuming and error prone as several web interfaces need to be utilized (AmiGO and Entrez), and as the results of the search in one resource need to be copied and pasted as search terms in the other. The main reason for such inefficiency is that high quality resources such as GO and EG have been designed primarily for consultation by humans, not for automated processing by agents or integration in applications. Moreover, these resources have been developed by different groups, independently of each other and are therefore not interoperable. No system currently supports

complex queries such as: *Find all the genes annotated with glycosyltransferase-related terms in GO and associated with the disease congenital muscular dystrophy in OMIM*. Typically, querying across the different knowledge sources is accomplished manually through meticulous work or requires the development of complex and customized software applications.

3. MATERIALS

Gene Ontology

The Gene Ontology (GO) seeks to provide a consistent description of gene products [17]. GO consists of three controlled vocabularies for biological processes (9,234 terms), molecular functions (7,456 terms) and cellular components (1,804 terms). The GO monthly releases are made available on the GO website in various formats, including RDF. The version of GO used in this study is dated of September 2006.

Entrez Gene

The Entrez Gene (EG) database records gene-related information from sequenced genomes and of model organisms that are focus of active research [18], totaling about two million genes. EG contains gene information about genomic maps, sequences, homology, and protein expression among others [18]. In contrast to GO, EG is not available in RDF, but in XML (converted from ASN1 by the program *gene2xml* provided by NCBI), and can be downloaded from the NCBI website. The version of EG used in this study is dated of July 2006.

4. EXPERIMENTAL METHODS

Our integration method can be summarized as follows and is illustrated in Figure 1. First, we extract manageable subsets from the two resources to be integrated. We then have to convert the EG subset from XML to RDF. Finally, we load both RDF resources in a common store, apply inference rules, and issue queries against it.

Creating subsets

The entire Entrez Gene data file (in XML format) is very large (50 GB) and unnecessarily difficult to manipulate. In order to obtain a manageable subset from EG, we restricted the gene records to two species: *Homo sapiens* (human) and *Mus musculus* (mouse). The resulting EG subset contains a total of 99,861 complete gene records (excluding obsolete records).

Converting XML format Entrez Gene data to RDF

A key element of our integration approach is the conversion of Entrez Gene from XML to RDF. There are many issues involved in the conversion of XML data into RDF format, including modeling the original semantics of the data, filtering redundant XML element tags, linking data entities using meaningful named relationships and identifying entities consistently within and across resources. Unlike traditional XML to XML conversion, XML to RDF conversion should exploit the advantages of the RDF model in representing the logical structure of the information.

A naïve approach to converting XML resources into RDF would transform each XML element tag into a predicate, mechanically. The resulting RDF representation, although syntactically correct, would be semantically limited, as the naming of the element tags is not necessarily reflective of their underlying semantics and may not be consistent. Instead, we manually examined the XML element tags and converted them into meaningful and standardized relationship names that convey explicitly the semantics of the connection between the *subject* and the *object*. For example, the element *<Org-ref_taxname>* was mapped to the more meaningful relationship named *has_source_organism_-taxonomic_name*.

In fact, the main objective of the conversion of EG to RDF is not to make it syntactically compatible with other RDF resources, but mostly to add expressive semantics to the EG data through the use of named relationships connecting EG entities. In other words, the conversion process realizes limited semantic enrichment in addition to syntactic transformation.

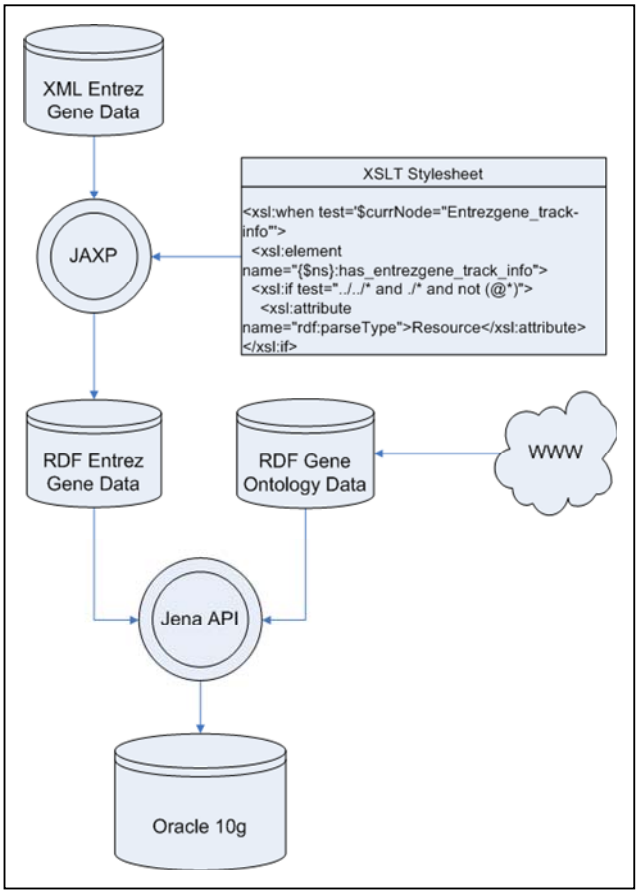

Figure 1. Overview of the integration method

We selected the eXtensible Stylesheet Language Transformation (XSLT) [19] for converting the EG XML information into RDF, because this approach allows for a clean separation between the application (using Java API for XML Processing (JAXP)) and the conversion logic (using XSLT stylesheet). Once the stylesheet is created, it can serve as an auxiliary file for existing programs realizing the XML to RDF conversion. In other words, the major interest of this approach is that no specific code is required for the conversion, because the transformation logic resides entirely in the stylesheet.

Loading the two resources into a single data store

Some of the requirements for our RDF store include native support for the RDF graph data model, support for persistence and indexing of the RDF triples, support for extensive collections of triples, and availability of a query language for the RDF graph. After surveying available RDF storage solutions, we decided to use Oracle Spatial 10g [20] as the RDF storage system. However, since we do not use any features specific to this product, we believe other RDF storage systems could be easily substituted.

The RDF file resulting from the XSLT conversion of the original XML file for EG and the RDF version of GO downloaded from the GO website are both loaded into a single RDF store. More precisely, the RDF resources are first converted to the NTriple format using the Jena API [21] and loaded into the RDF database using a loader provided by Oracle.

Applying inference rules

Unlike the Web Ontology language OWL, RDF provides no direct support for inference. However, inference rules can be implemented in the RDF store to make explicit the semantics of some predicates. For example, the relationships *is_a* and *part_of* used in GO are partial order relations, thus being reflexive, antisymmetric and transitive. The inference rules we created for implementing the transitivity and combination of these two relationships are shown in Table 1. The inference rules are stored in a rule base created in Oracle 10g.

Table 1. Inference rules for is_a and part_of in GO

Relation	*is_a*	*part_of*
is_a	IF <**x** *is_a* **y**> & <**y** *is_a* **z**> THEN <**x** *is_a* **z**>	IF <**x** *is_a* **y**> & <**y** *part_of* **z**> THEN <**x** *part_of* **z**>
part_of	IF <**x** *part_of* **y**> & <**y** *is_a* **z**> THEN <**x** *part_of* **z**>	IF <**x** *part_of* **y**> & <**y** *part_of* **z**> THEN <**x** *part_of* **z**>

Querying the RDF graph with SPARQL

SPARQL [22] is a query language for RDF graphs, equivalent to SQL, the Structured Query Language, for relational databases. Unlike SQL, SPARQL does not require users to be familiar with the data model (e.g., tables, foreign keys), but simply to indicate how entities of interest relate to each other. For example, the structure of the query: *Find all the genes annotated with the GO molecular function glycosyltransferase (*GO:0016757) *or any of its descendants and associated with any form of congenital muscular dystrophy* is represented in Figure 2.

The query can be understood as finding a path in the RDF graph using a predetermined set of semantic relationships and would be formulated as follows. Because of the inference rules implementing the transitivity and reflexivity of the *is_a* relationship, the condition on the GO annotation "*glycosyltransferase (*GO:0016757) *or any of its descendants*" is easily expressed by '?t is_a GO:0016757'. The link between genes and GO terms is expressed by '?g has_molecular_function ?t'. Similarly, the link between genes and OMIM diseases is expressed by '?g has_associated_phenotype ?b2' (OMIM ID) and '?b2 has_textual_description ?d' (disease name). Finally, direct constraints are put on the GO term on the one hand ('?t is_a GO:0016757', to select *glycosyltransferase* (GO:0016757)) and on disease names on the other (where a regular expression is used to select disease names containing the strings "congenital", "muscular" and "dystrophy"). The simplified SPARQL query is shown in Figure 3. The actual SPARQL query used in this study is displayed in Figure 7, along with thee output it produces.

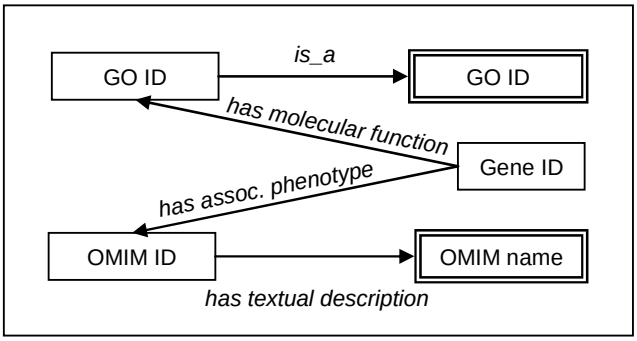

Figure 2. RDF graph corresponding to the query above

```
SELECT distinct t,g,d
FROM TABLE(SDO_RDF_MATCH(
'(?t is_a GO:0016757)
 (?g has molecular function ?t)
 (?g has_associated_phenotype ?b2)
 (?b2 has_textual_description ?d)',
SDO_RDF_Models('entrez_gene'),
SDO_RDF_Rulebases('entrez_gene_rb'),
SDO_RDF_Aliases(SDO_RDF_Alias('',''')), null) )
where (
    REGEXP_LIKE(LOWER(d), '((.)*(congenital)(.)*)')
AND REGEXP_LIKE(LOWER(d), '((.)*(muscular)(.)*)')
AND REGEXP_LIKE(LOWER(d), '((.)*(dystrophy)(.)*)'));
```

Figure 3. Example of SPARQL query (simplified)

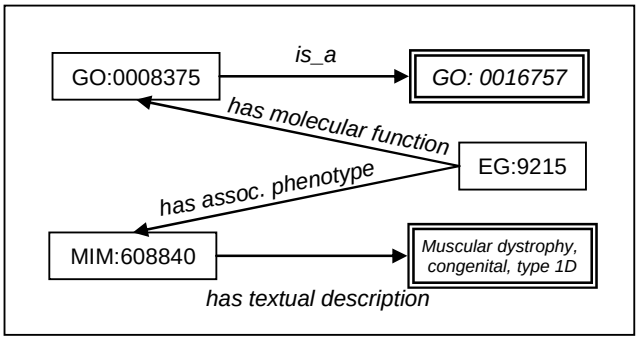

Figure 4. Instantiated RDF graph

5. RESULTS

One integrated RDF repository for Entrez Gene and GO

The subset of Entrez Gene restricted to *Homo sapiens* (human) and *Mus musculus* (mouse) as biological sources comprises 99,861 gene records. Once converted to RDF, it consists of 772,530 triples. The RDF version of GO contains 293,798 triples. Overall, there are over one million triples in the store created for this experiment, which is relatively small in comparison to the

411 million triples resulting from the conversion of the entire EG to RDF [23].

Biological query result

The SPARQL query presented above returned one result, corresponding to one path in the graph between the GO term *glycosyltransferase (*GO:0016757) and OMIM disease names containing (variants of) the string "*congenital muscular dystrophy*".

This path involved the human gene *LARGE like-glycosyltransferase* (EG:9215), annotated with the GO term *acetylglucosaminyltransferase* (GO:0008375), a descendant of *glycosyltransferase (*GO:0016757). Also involved in this path is the OMIM disease identified by MIM:608840. The name (textual description) of this disease is *Muscular dystrophy, congenital, type 1D* and contains the required substrings "congenital", "muscular" and "dystrophy". The instantiated RDF graph with path between *glycosyltransferase* (GO:0016757) and *Muscular dystrophy, congenital, type 1D* is shown in Figure 4.

This simple SPARQL provides an easy way of testing the biological hypothesis under investigation, i.e., the existence of a possible link between glycosylation and *congenital muscular dystrophy*. On manual inspection of the Entrez Gene record shown in Figure 6, we also note that the given gene may be involved in the development and progression of meningioma through modification of ganglioside composition and other glycosylated molecules in tumor cells.

6. DISCUSSION

Significance

In this study, we demonstrated the feasibility of integrating two biomedical knowledge resources through RDF. We also provided anecdotal evidence for the benefits of such integration by showing how glycosyltransferase can be linked to congenital muscular dystrophy. The integrated resource is greater than the sum of its parts as it supports complex queries that could typically not be handled otherwise without tedious manual intervention or customized software applications.

Integrated resources based on a graph model are particularly important in an exploratory context where researchers need to "connect the dots" in order to validate a hypothesis. This approach also facilitates intuitive hypothesis formulation and refinement. For example, after verifying that glycosyltransferase is linked to congenital muscular dystrophy, our researchers may narrow the focus of their wet lab experiments to only *hexosyltransferase* out of the potential seven glycosyltransferases. Analogously, they can focus their research on *Muscular dystrophy, congenital, type 1D*, out of several other diseases.

Arguably, the graph data model of RDF resources is more intuitive than the database schemas. In fact, the RDF data model enables us to model the inherent logical relations between entities that mirror the human cognitive model of the real world. In fact, users familiar with the conceptual structure of EG and GO should be able to query the RDF graph integrating these two resources. For example, users are only required to know that genes have molecular functions and are associated with diseases. This is why an important element of the RDF conversion is to create explicit relationship names reflecting the semantics of the links.

Finally, from a technical perspective, the RDF data model offers more flexibility than database schemas for accommodating changes to the underlying model.

Generalization

The particular biological example presented here was suggested by colleagues from the Complex Carbohydrate Research Center at the University of Georgia, not involved with the development of our RDF integration project. Moreover, this example was identified *after* creating the integrated resource. In other words, the subset of EG was not tailored to support this particular query, which suggests it could support queries in many other biological subdomains. In fact, the only reason why EG was restricted to a subset is to limit the size of the store, for practical reasons.

In this feasibility study, we were primarily interested in demonstrating how one particular hypothesis, i.e., the existence of an association between *glycosyltransferase* and *congenital muscular dystrophy*, could be refined through the existence of paths in the RDF graph. Another use of the graph would be to mine hypotheses, instead of refining them. For example, researchers could create SPARQL queries to identify all classes of enzymes involved with a given disease, or with an arbitrary list of diseases, thus generating hypotheses, not only refining ad hoc hypotheses.

The resources integrated in our pilot system are currently limited to the Gene Ontology and a subset of Entrez Gene. However, as part of the *Advanced Library Services* project, we are also extracting relations from the biomedical literature. These relations, also represented in RDF, will soon be integrated in our *Biomedical Knowledge Repository*, together with information extracted from several structured knowledge resources. This repository will thus support a wider range of queries. For example, future projects include an analysis of the genes associated with tobacco smoking behavior, identified by the Genome Wide Association of Nicotine Dependence – NICSNP Project, in collaboration with the National Institute on Drug Abuse.

Why RDF?

We have used RDF as the representational format, in place of RDF/S or OWL, even though RDF/S and OWL allows more expressivity in capturing domain knowledge. For example, RDF/S supports property inheritance, and the expressive features of OWL include membership and numerical restrictions on concepts or relations. Importantly, these features are currently not present in the original data model of EG or GO. Adding such features to a resource is generally labor intensive and expensive, and therefore hardly suitable for the automated integration of existing data sources sought in this project.

In contrast to OWL, the RDF data model is well suited for the representation of the original EG data along with (automated) incorporation of named relationships. Moreover, RDF allows us to use a powerful and scalable rule base to reason over the integrated knowledge source. Additionally, the reasoning services currently available for OWL-DL are limited in the volume and complexity of the data they can handle, and would typically be overwhelmed by the hundreds of thousands entities involved in our repository.

The SPARQL query language for RDF is currently not customized for optimal performance in scenarios involving

multiple traversals of named relationships to answer a query. Future work will involve the collection of empirical data to evaluate the performance of the SPARQL with respect to different categories of queries.

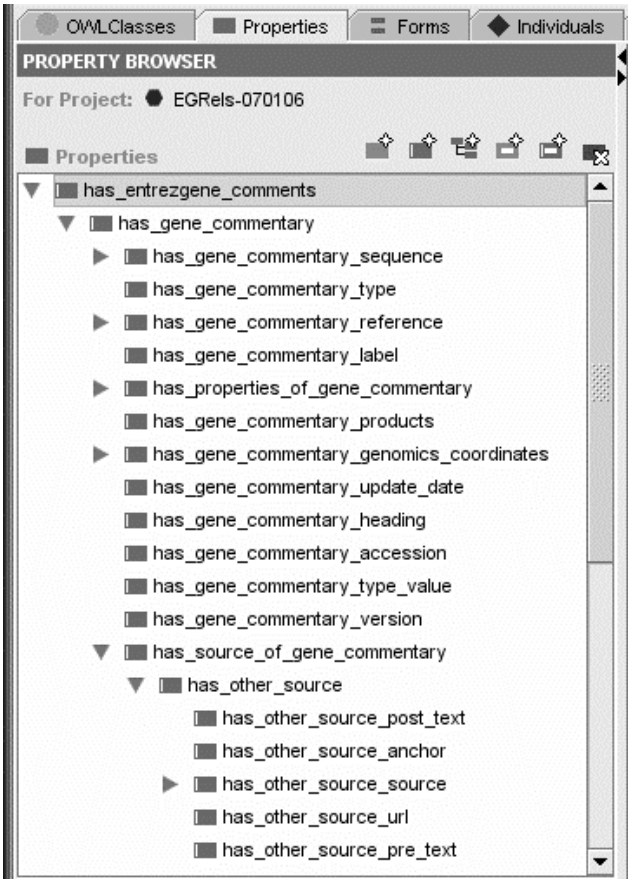

Figure 5. A hierarchy of relationships among entities in Entrez Gene (partial representation)

A formal model of relationships

Meaningful relationships connecting biological entities play a critical role in the successful integration of data sources using a Semantic Web approach. The named relationships (or predicates), i.e., the links in the RDF graph, are first class objects in the RDF data model. The relationships are not only represented explicitly in the RDF store, but they are also an integral part of the queries.

In our current effort we have tried to capture the original semantics of the EG schema when converting the original XML element tags into named relationships. The conversion of the element tags resulted on a list of over one hundred relationships. However, in order to take full advantage of the relationships in reasoning tasks, these relationships need to be organized not as a flat list, but into a formal, hierarchical model defining relationships among relationships.

To our knowledge, there exists no such formal model of relationships comprehensive and fine-grained enough to accommodate the kinds of relationships encountered in Entrez Gene, for example. A small number of top-level relationships in biomedical ontologies have been defined and could provide a formal framework for defining finer-grained relationships [24]. The Semantic Network of the Unified Medical Language System (UMLS) defines, although less formally, a larger set of 54 relationships, which could be used as a backbone for organizing the relationships in our system [25].

In addition to supporting reasoning, a detailed reference relationship ontology would also be useful in the conversion process in which XML element tags are converted into RDF relationships using an XSLT stylesheet mechanism. Attached to a given relationship in the ontology would be the list of XML element tags in the various information sources to be integrated, whose conversion should result in this relationship. Because the ontology contains both the relationships used in the RDF graph and the corresponding element tags found in the XML sources, an XSLT stylesheet generator can take advantage of the ontology to automate the generation of the XSLT, i.e., to map the XML element tags to RDF predicates.

Our relationship ontology currently comprises the XML element tags in EG along with the named relationships we created for them. In its initial stage, the organization of the relationships simply reflects the tree-like structure of the EG XML schema. A portion of this relationship ontology is shown in Figure 5. The only relationships present in the GO are *is_a* and *part_of*. As we integrate additional resources, we will reorganize the relationships into a more expressive structure, aligning them with existing relationship ontologies, such as the UMLS Semantic Network's.

Unique Identifiers for biomedical entities

Heterogeneous resources can interoperate in a RDF graph only if the entities shared by these resources are identified consistently. For example, the Gene Ontology can be used easily in conjunction with Entrez Gene, because Entrez Gene uses GO identifiers to refer to terms in the GO. This seamless integration allows us to relate genes not only to the GO terms to which these genes are annotated, but also to the ancestors of these terms. Similarly, we would like to be able to abstract away from specific diseases in OMIM and relate genes to higher-level disease categories (e.g., *muscular dystrophy*, as opposed to *Muscular dystrophy, congenital, type 1D*). However, in contrast to GO terms, OMIM diseases are not organized in a hierarchy, nor are they integrated in the hierarchical structure of the UMLS Metathesaurus, where such disease categories are represented. As a consequence, little reasoning can be performed on the side of diseases in our current RDF store.

In order to identify biomedical entities, we plan to rely as much as possible on comprehensive and already integrated terminological resources. This is the case, for example, of the UMLS Metathesaurus, integrating the names of some 1.4 million biomedical entities, including diseases, drugs and organisms. The UMLS is a stable resource that has been developed and updated regularly for 20 years by the National Library of Medicine (NLM) [25]. In order to compensate for its limited coverage of genes, we also plan to use other resources of the NLM such as Entrez Gene. Because the UMLS and Entrez Gene already integrate names from several terminological resources, they provide a broader namespace compared to individual ontologies, thus reducing the need for mapping between namespaces.

A distinct issue is that there is no universally accepted schema for identifying entities. The main contenders are the Life Science Identifier (LSID) [21] and solutions based on the HTTP protocol (i.e., URIs (Universal Resource Identifiers), URLs (Universal

Resource Locators) and URNs (Universal Resource Names)). Differences between them include support for versioning and resolution (i.e., what information can be obtained from the identifier). As shown in Figure 7, until one schema is adopted, we decided to use the EG DTD URL as the namespace to create the identifier for gene entities in the RDF store. For GO terms, we temporarily use the URL of GO. These decisions can be changed with minimal effort, simply by modifying the XSLT stylesheet when a resource is loaded into the RDF store.

7. CONCLUSION

The integration approach demonstrated in this study takes advantage of technologies developed for the Semantic Web, such as RDF. We showed how two large biomedical knowledge resources can be integrated through RDF and we presented one application of the integrated RDF store to generating research hypotheses. At a time when biomedical knowledge is overabundant, heterogeneous and scattered, we believe that this approach can help researchers process it in a more efficient way.

8. ACKNOWLEDGEMENT

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM) and by the Integrated Technology Resource for Biomedical Glycomics (5 P41 RR18502), funded by the National Institutes of Health National Center for Research Resources.

9. REFERENCES

1. Mitchell J.A., M.A.T., Bodenreider O., *From phenotype to genotype: issues in navigating the available information resources.* Methods Inf Med, 2003. **42**(5): p. 557-63.
2. Hernandez, T. and S. Kambhampati, *Integration of biological sources: Current systems and challenges ahead.* Sigmod Record, 2004. **33**(3): p. 51-60.
3. Sheth, A.P.a.J.A.L., *Federated database systems for managing distributed, heterogeneous and autonomous databases.* ACM Computing Surveys, 1990. **22**: p. 183--236.
4. Manola, F., Miller, E.(Eds.). *RDF Primer*. W3C Recommendation 2004 10 February
   Available from: http://www.w3.org/TR/rdf-primer/.
5. Davidson, S.B., et al., *K2/Kleisli and GUS: Experiments in integrated access to genomic data sources.* Ibm Systems Journal, 2001. **40**(2): p. 512-531.
6. Stevens, R., et al., *TAMBIS: transparent access to multiple bioinformatics information sources.* Bioinformatics, 2000. **16**(2): p. 184-5.
7. Donelson, L., et al., *The BioMediator system as a data integration tool to answer diverse biologic queries.* Medinfo, 2004. **11**(Pt 2): p. 768-72.
8. Wheeler DL, B.T., Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Ostell J, Miller V, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E. , *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Research, 2007 Jan. **35**((Database issue)): p. D5-12.
9. Ruttenberg, A., et al., *Advancing translational research with the Semantic Web.* BMC Bioinformatics, 2007: p. (in press).
10. Cheung KH, Y.K., Smith A, Deknikker R, Masiar A, Gerstein M. , *YeastHub: a semantic web use case for integrating data in the life sciences domain.* Bioinformatics., 2005. **21**(Suppl 1): p. i85-96.
11. Cheung, K.-H., et al., *Semantic Web approach to database integration in the life sciences*, in *Semantic Web: Revolutionizing knowledge discovery in the life sciences*, C.J.O. Baker and K.-H. Cheung, Editors. 2007, Springer: New York. p. 11-30.
12. Goble, C., et al., *Knowledge discovery for biology with Taverna: Producing and consuming semantics in the Web of Science*, in *Semantic Web: Revolutionizing knowledge discovery in the life sciences*, C.J.O. Baker and K.-H. Cheung, Editors. 2007, Springer: New York. p. 355-395.
13. Neumann EK, Q.D. *BioDash: a Semantic Web dashboard for drug development*. in *Pac Symp Biocomput.* 2006.
14. Bodenreider O, R.T., *Advanced library services: Developing a biomedical knowledge repository to support advanced information management applications.* September 14, 2006, Lister Hill National Center for Biomedical Communications, National Library of Medicine: Bethesda, Maryland.
15. *Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins Univer-sity (Baltimore, MD) and National Center for Biotechnology In-formation, National Library of Medicine.*
   Available from: http://www.ncbi.nlm.nih.gov/omim/.
16. *AmiGO: Gene Ontology browser*.
   Available from: http://www.godatabase.org/.
17. Ashburner, M., Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet., 2000. **25**((1)): p. 25-9.
18. Maglott, D., Ostell, J., Pruitt, K.D., and Tatusova, T., *Entrez Gene: gene-centered information at NCBI.* Nucleic Acids Res., 2005 January 1. **33**((Database Issue)): p. D54–D58.
19. *XML Schema Language Transformation*.
   Available from: http://www.w3.org/TR/xslt.
20. Alexander, N., Ravada S. *RDF Object Type and Reification in Oracle"—Technical White Paper*. Available from: http://download-east.oracle.com/otndocs/tech/-semantic_web/pdf/rdf_reification.pdf.
21. McBride, B., *Jena: A Semantic Web Toolkit.* IEEE Internet Computing, Nov. 2002. **6**: p. 55-59.

22. *SPARQL Query Language for RDF*. W3C Working Draft 2006
Available from: http://www.w3.org/TR/rdf-sparql-query.
23. *BioRDF subgroup: Health Care and Life Sciences interest group* Available from http://esw.w3.org/topic/HCLSIG_BioRDF_Subgroup.
24. Smith, B., et al., *Relations in biomedical ontologies.* Genome Biol, 2005. **6**(5): p. R46.
25. Vizenor, L., et al., *Enhancing biomedical ontologies through alignment of semantic relationships: Exploratory approaches.* AMIA Annu Symp Proc, 2006: p. 804-8.

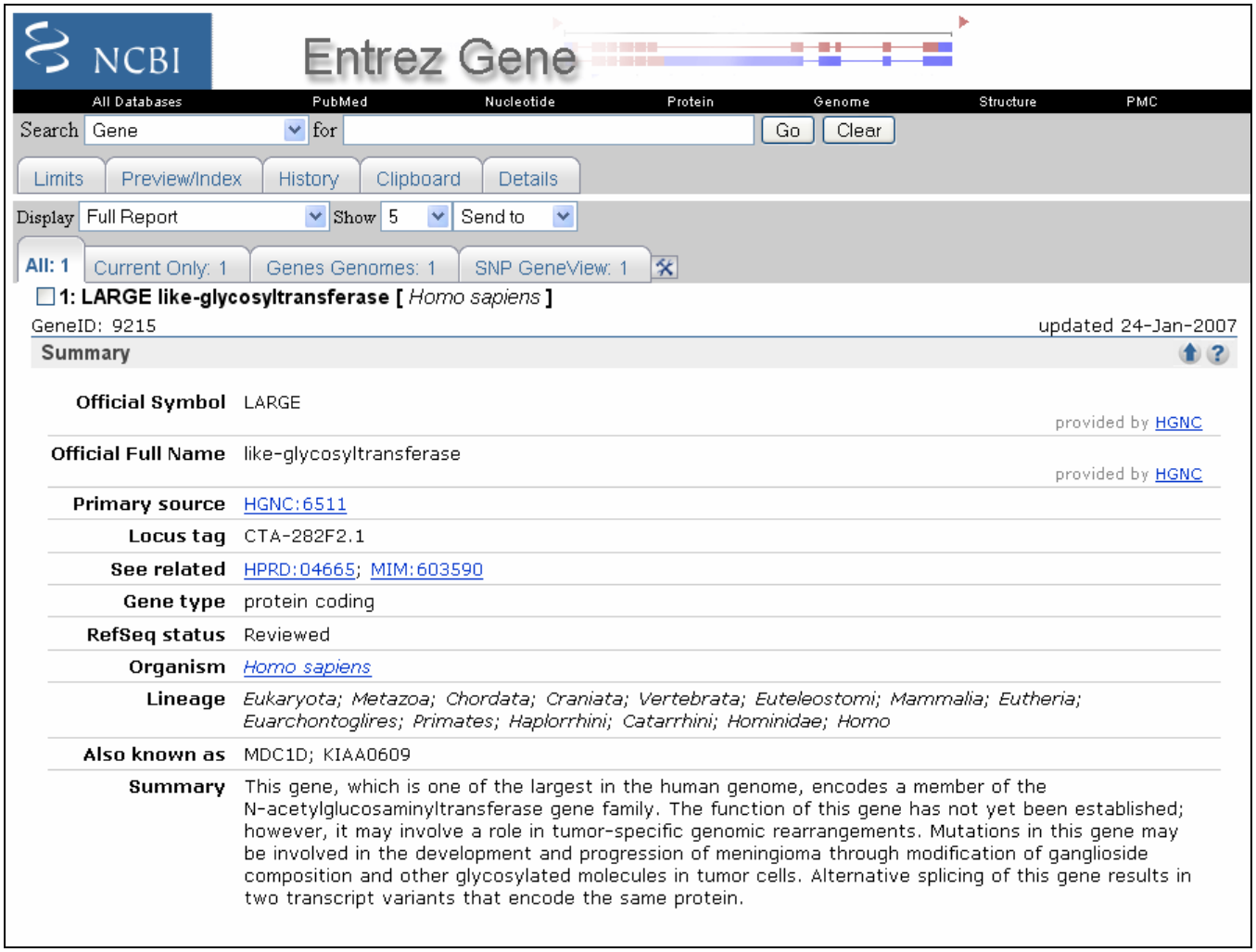

Figure 6. Screenshot of the Entrez Gene website for the gene "LARGE like-glycosyltransferase"

```
SELECT distinct t,g,d
FROM TABLE(SDO_RDF_MATCH(
'(?t <http://www.geneontology.org/dtds/go.dtd#is_a>
                                                <http://www.geneontology.org/go#GO:0016757>)
(?g <http://www.ncbi.nlm.nih.gov/dtd/NCBI_Entrezgene.dtd/has_GeneOntology_annotation> ?b1)
(?b1 <http://www.ncbi.nlm.nih.gov/dtd/NCBI_Entrezgene.dtd/has_GO_ID> ?t)
(?g <http://www.ncbi.nlm.nih.gov/dtd/NCBI_Entrezgene.dtd/has_OMIM_record> ?b2)
(?b2 <http://www.ncbi.nlm.nih.gov/dtd/NCBI_Entrezgene.dtd/has_textual_description> ?d)',
SDO_RDF_Models('entrez_gene'),
SDO_RDF_Rulebases('entrez_gene_rb'),
SDO_RDF_Aliases(SDO_RDF_Alias('','')),
null)
) where ( REGEXP_LIKE(LOWER(d),
'((.)*(muscular)(.)*)')  AND REGEXP_LIKE(LOWER(d),
'((.)*(dystrophy)(.)*)') AND REGEXP_LIKE(LOWER(d),
'((.)*(congenital)(.)*)'));
```

```
http://www.geneontology.org/go#GO:0008375
http://www.ncbi.nlm.nih.gov/dtd/NCBI_Entrezgene.dtd/9215
Muscular dystrophy, congenital, type 1D
```

Figure 7. The actual SPARQL query (top) and output (bottom) used in our extended example