The FungalWeb Ontology: Semantic Web Challenges in Bioinformatics and Genomics

Arash Shaban-Nejad, Christopher J.O. Baker, Volker Haarslev, and Greg Butler

Dept. of Comp. Sci. and Software Eng., Concordia University, H3G1M8 Montreal, Canada
{arash_sh, baker, haarslev, gregb}@cs.concordia.ca

1 Introduction

Bioinformatics and genomics cover a wide range of different data formats (i.e. annotations, pathways, structures, sequences) derived from experimental and in-silico biological analysis which are stored, used, and manipulated by scientists and machines. The volume of this data is huge and usually distributed in different locations, and often frequently being updated.

FungalWeb is the first project of its kind in Canada to focus on bringing semantic web technology to genomics. It aimed to bring together available expertise in ontologies, multi-agent systems, machine learning and natural language processing to build a tailored knowledgebase and semantic systems of direct use to the scientific discovery process in the domain of fungal genomics [1].

We describe the FungalWeb Ontology which is a large-scale integrated bio-ontology in the domain of fungal genomics using state-of-the-art semantic technologies. The ontology provides simplified access to units of intersecting information from different biological databases and existing bio-ontologies. In particular, the FungalWeb ontology is being used as a core for a semantic web system. This system can be used by human, bioinformatics applications or some intelligent systems for ontology-based information retrieval to provide extended interpretations and annotations. [2]

2 The FungalWeb Ontology Design and Evaluation

The FungalWeb Ontology [2] is the result of integrating numerous biological database schemas, web accessible textual resources and interviews with domain experts and reusing some existing bio-ontologies. The Ontology is designed with a high level of granularity and implemented in OWL-DL language to take advantage of the combination of a frame representation of OWL framework and expressive Description Logics (DL). The majority of the terms in the FungalWeb Ontology come from following sources:

- NCBI taxonomy database [3]: contains the names of all organisms including fungi.
- NEWT: is the taxonomy database maintained by the Swiss-Prot [4].
- BRENDA [5]: a database of enzymes which provides a representative overview of enzyme nomenclature, enzyme features and actual properties.
- SwissProt [6]: a protein sequence database providing highly curated annotations, a minimal level of redundancy and a high level of integration with other databases.
- Commercial Enzyme Vendors: Companies that retail enzymes provide detailed descriptions of the properties and benefits of their products on their websites.

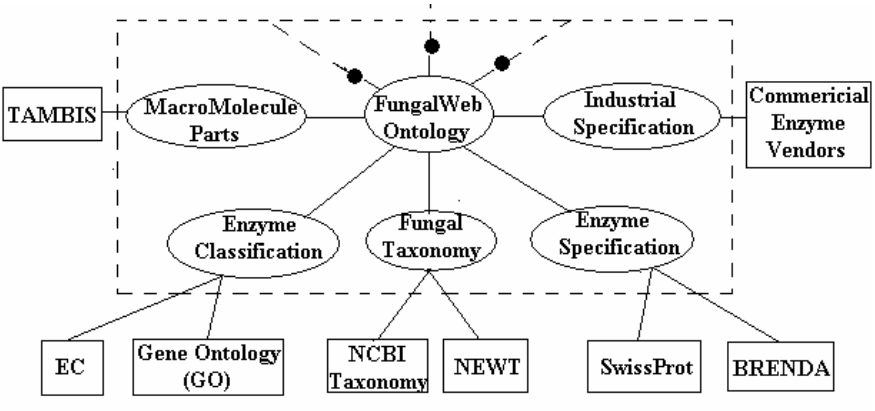
Y. Gil et al. (Eds.): ISWC 2005, LNCS 3729 , pp. 1063 – 1066, 2005.

The FungalWeb Ontology also reuses existing domain specific bio-ontologies such as Gene Ontology (GO) [7] and TAMBIS [8]. This is done by merging, mapping, sharing common concepts and partially importing instances. By reusing concepts from other generic ontologies, a set of well defined concepts is obtained.

The integration is done at two levels: Data and Semantic Integration. Data integration is done by normalizing extracted data into a consistent representation. In order to perform semantic integration these we manually identified the relevant data items and the semantic commonality to bring them in a unified frame of reference.

Currently the Ontology contains 3667 concepts, 12686 instances and 157 Properties. Efforts to expand the conceptualization are continuing. Inclusion of more instance data in the knowledgebase allows us to pose richer and more complex queries.

Different associative properties were defined to relate individuals of concepts. For example the property "has been reported to be found in" relates an enzyme individual to a corresponding fungal species.



Source name	Instances	Concepts
BRENDA & SwissProt	105	35
EC & GO	1296	198
NCBI taxonomy database	10870	3340
TAMBIS	0	22
Commercial Enzyme Vendors	401	6

Fig. 1. The major resources included within the FungalWeb Ontology

As shown in Fig.1 most of terminology in the ontology is fungal organisms and fungal enzymes. The classification of fungi presented in the ontology is based on phylum, class, order, family, genus and species. Species are considered as the fungi instances. Enzymes are classified based on catalyzed reactions recommended by the International Union of Biochemistry and Molecular Biology (IUBMB). Enzyme names are defined as the enzyme instances.

The evaluation of the ontology is done pragmatically, by assessing the ontology to satisfy the requirements of our application, including determining the logical and semantic consistency. Logical consistency is checked automatically by KR editors and semantic consistency is assisted by the DL reasoner in the identification of correct

or miss-classification. Although we sought to validate the biological data and relations by citing their origin (database or literature) or by checking consistency, validation by the domain expert was also necessary.

We use RACER [15] as a DL reasoning system with support for T-Box (axioms about class definitions) and A-Box (assertions about individuals) for reasoning on the FungalWeb Ontology and Checking the A-Box and T-Box consistency. On average, Racer solves the posed subsumption problems within fraction of a second. The performance of Racer is highly dependent on the number of individuals and response time grows with the number of individuals. The number of properties does not have an important affect on the response time, but, the number of property fillers has strongest influence on the performance with respect to instance retrieval.

3 Application Scenarios and Semantic Querying

We describe real world application scenarios to demonstrate what a bioinformatics application can gain from using ontology-based technologies. We argue for the commercial usage and business feasibility of the ontology by presenting scenarios that show how the diverse needs of the fungal biotechnology manager can be accommodated by semantic querying of an integrated set of data in the ontology. FungalWeb Ontology currently accommodates the application scenarios below [10].

- Identification of enzymes acting on substrates
- Identification of enzyme provenance and common taxonomic lineage
- Identification of commercial enzyme products for enzyme benchmark testing
- Identifying enzymes with unique properties suited for industrial application

These scenarios are illustrated [10] by posing semantic queries to the FungalWeb knowledgebase using a description logics based query language called nRQL (new Racer Query Language) [11]. nRQL is implemented in Racer with its applicability to OWL Semantic Web repositories to retrieve A-box individuals under specific conditions. nRQL is more expressive than traditional concept-based retrieval languages offered by previous DLs reasoning systems.

An example query made to the Ontology: This query retrieves the individuals of vendor name for vendors that sell products containing xylanase enzymes.

```
(RETRIEVE (?x) (AND (?x ?y |http://a.com/ontology#Sells|)
   (?y ?z  |http://a.com/ontology#Contains|) (?z |http://a.com/ontology#Xylanase|) ))
```

Also we use nRQL to retrieve values of annotation properties used to annotate ontological resources. These annotations represent metadata (i.e. comments, creator, date, identifier, source name, source URL, version, etc.) but can not be used for reasoning. This capability can be very useful for the ontology maintenance, versioning and providing proof and trust in a semantic web system.

For example the following query retrieves the source(s) for “Enzyme”.

```
(RETRIEVE (|http://a.com/ontology#Enzyme|
   (TOLD-VALUE (|http://a.com/ontology#Source|  |http://a.com/ontology#Enzyme|)))
   (BIND-INDIVIDUAL |http://a.com/ontology#Enzyme|))
```

4 Challenges

In the process of employing semantic web technology to develop ontology and a large knowledgebase in the domain of fungal biotechnology, we had to deal with variety of different challenges. Some of the major challenges included; working with highly heterogeneous and volatile data, the integration of ontologies implemented in different languages, with different semantic tools and platforms, and the lack of trustable tools for this purpose.

Our ongoing research involves improvement of querying capabilities and using Natural Language Processing (NLP) techniques for ontology update and change management. The project FungalWeb: "Ontology, the Semantic Web and Intelligent Systems for Genomics" is funded by Génome Québec.

References

1. Baker C. J. O., Butler G., and Haarslev V. Ontologies, Semantic web and Intelligent Systems for Genomics. 1st Canadian Semantic Web Interest Group Meeting (SWIG'04) , Montreal, Quebec, Canada (2004).
2. Shaban-Nejad A., Baker C. J. O., Butler G. Haarslev V. The FungalWeb Ontology: Semantic Web Application for Fungal Genomic. 1st Canadian Semantic Web Interest Group Meeting (SWIG'04) , Montreal, Quebec, Canada (2004).
3. National Centre for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/).
4. NEWT, UniProt taxonomy browser (http://www.ebi.ac.uk/newt/index.html).
5. Brenda Enzyme Database (http://www.brenda.uni-koeln.de/).
6. SwissProt protein sequence database (http://ca.expasy.org/sprot/).
7. Gene Ontology documentation, (http://www.geneontology.org/doc/GO.doc.html).
8. P.G. Baker, A. Brass, S. Bechhofer, C. Goble, N. Paton, and R. Stevens. TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. An Overview.In Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology (ISMB'98), pages 25-34, California, June 1998.
9. Volker Haarslev, Ralf Möller. RACER System Description. Proceedings of International Joint Conference on Automated Reasoning, IJCAR'2001, R. Goré, A.Leitsch, T. Nipkow (Eds.), June 18-23, 2001, Siena, Italy, Springer-Verlag, Berlin,pp. 701-705.
10. Baker C. J. O., Witte R., Shaban-Nejad A., Butler G., and Haarslev V. The FungalWeb Ontology: Application Scenarios. Eighth Annual Bio-Ontologies Meeting, co-located with ISMB 2005, Detroit, Michigan, USA (2005).
11. M. Wessel, R. Möller. A High Performance Semantic Web Query Answering Engine. International Workshop on Description Logics (DL2005), Edinburgh, Scotland, UK, 2005.