The FungalWeb Ontology

The Core of a Semantic Web Application for Fungal Genomics

Arash Shaban-Nejad, Christopher J.O Baker, Greg Butler, Volker Haarslev
Department of Computer Science and Software Engineering
Centre for Structural and Functional Genomics
Concordia University, Montreal, Quebec, Canada

Abstract

A formal ontology design and implementation case study which serves as the core for a semantic web application in the area of fungal genomics is presented. Simplified semantic access to units of intersecting information from different biological databases is under development

Introduction

The Semantic Web aims to extend the existing Web with conceptual metadata that are more accessible to machines and thereby more effectively communicate the proposed meaning of Web resources [1]. Bioinformatics is widely accepted as an important research area for the Semantic Web [2]. These bioinformatics resources are rich in data and knowledge, but most of the information generated by biologists is in the contextual form e.g. natural language, image annotations, chemical pathways, annotated gene sequences, and protein structures.

Such information is obviously readable and understood by humans, but does not support analyses by computers [3]. Additionally, biological data is a mixed bag of data types from different experiments distributed in heterogeneous and incompatible databases. These databases must be bridged, normalized based on a conceptual aggregation and the content made accessible and comprehensible. Here, ontologies can serve to create a formal specification of biological knowledge to make the information computationally accessible and semantically clear.

Ontologies are accepted within the biology communityas a means of facilitating data exchange between databases. For example, Gene Ontology terms are used as values of fields in databases to describe molecular function, biological process and cellular components of gene products. Indeed ontologies have had their historical roots in the classification schemes for the tree of life first introdued by Linneaus in the 18 th centuary.

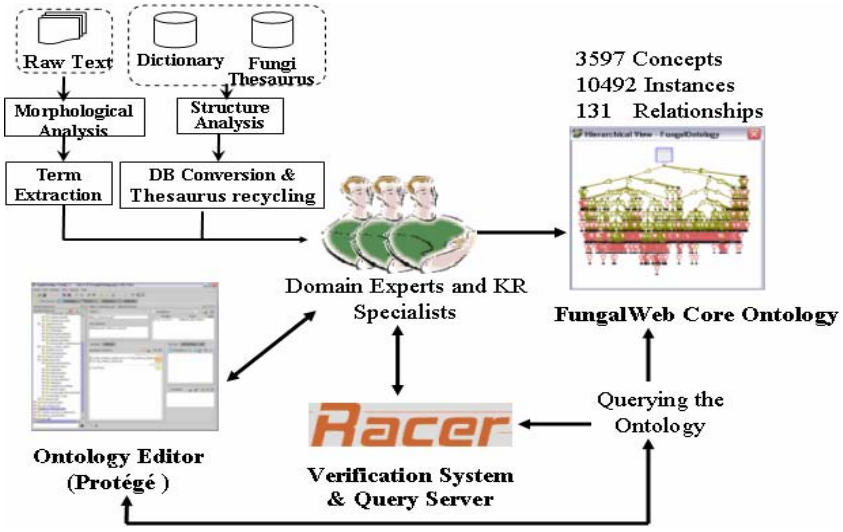

Figure 1 Ontology Development. FungalWeb: "Ontology, the Semantic Web and Intelligent Systems for Genomics" aims to represent and map fungal genomics information using ontologies**.**

In this paper we report on an ontology covering organism taxonomy, enzyme activity and properties of small chemical compounds. For the complete FungalWeb vision we also need to include mRNA expression levels as described by microarray data, metabolic pathways and regulatory networks. The main application in this paper is the use of formal ontology as represented in OWL and the query capabilities of Racer to demonstrate semantic querying of the integrated ontology.

FungalWeb Ontology Development

In the design and development of the FungalWeb Ontology we considered the following steps: specification, knowledge acquisition, implementation and semantic query.

Specification

A clear declaration of purpose, scope and degree of granularity is important for timely ontology development. The scope of the current work extended to the establishment of an instantiated knowledgebase describing fungal taxonomy and fungal enzymes and to make this semantically available to enzyme application scientists working in the field of fungal genomics.

Knowledge Acquisition

This step involves extracting data from online databases, libraries, interviews with domain experts and free text analysis. Several databases, online resources, dictionaries and raw texts are currently being used as resources from which to extract the related vocabularies and concepts. By organizing the vocabularies in a hierarchical structure with (is-a) relationships a full subsumption taxonomy is built. The major resources for fungal terminologies and concepts come from following sources:

- NCBI taxonomy database [4]: contains the names of all organisms including fungi that are represented in the genetic databases by at least one nucleotide or protein sequence.
- NEWT: is the taxonomy database maintained by the Swiss-Prot [5]. It is used in conjunction with the NCBI for extracting data on fungal concepts and instances
- BRENDA [8]: a database of enzymes providing a representative overview of enzyme nomenclature, enzyme features and properties
- Saccharomyces Genome Database [6]: contains genomic information specific to the genera Saccharomyces.
- Neurospora crassa Database [7]: contains genomic information specific to the genera Neurospora informatin.

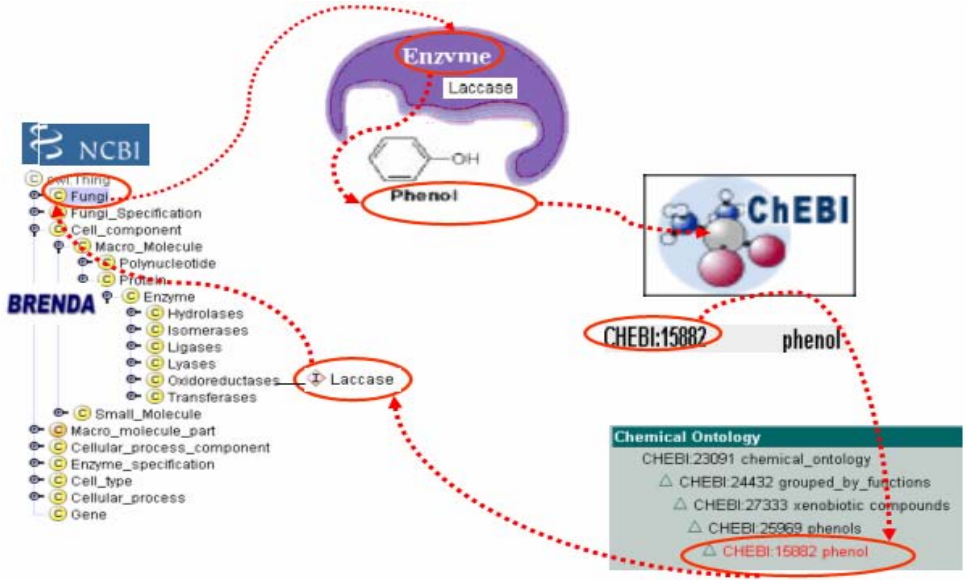

Figure 2: Semantic query for the laccase enzyme substrate, phenol, described in the context of the chemical dictionary of small molecule entities (ChEBI) and other bioinformatic resources (NCBI, BRENDA)

Implementation

This step involves conceptualization, integration and encoding. During the implementation process it is frequently necessary to alternate between these three activities.

Conceptualization

Here we define and describe concepts, instances, relations and attributes. At the end of conceptualization stage the taxonomy can be built by organizing all the defined concepts in a hierarchical tree structure. The taxonomy is based on the subsumption relationship between concept names. We use description logics (DLs) to define concepts and relationships. DLs describe knowledge in terms of concepts and relations that are used to automatically derive classification taxonomies. By using DLs, concepts are defined in terms of descriptions using other roles and other concepts. For example the concept Basidiomycota is defined as follows:

$$\text{Basidiomaycota} \equiv \text{Fungi} \cap \exists\ \text{has\_body.Fruity\_body}$$

At the conceptualization stage we define relationships, constraints and axioms. We are looking for mechanisms to query and retrieve complex relationships between concepts. The basic units in our taxonomy are the instances of fungal species. Inclusion of instances establishes a knowledgebase, however a clear distinction between ontology and knowledgebase has yet to be well defined.

Integration

The FungalWeb Ontology (FWOnt) aims to reuse and integrate existing bio-ontologies and knowledgebases by merging, mapping and sharing common concepts using Logics. The FungalWeb ontology is an integrated ontology which used components of Gene ontology (GO) [REF], TAMBIS. [REF]

Encoding

The aim of encoding is to represent the conceptualization in a formal ontology language. To achieve maximum expressiveness, without losing computational completeness and decidability of reasoning systems, we used OWL-DL. This is a sublanguage of Ontology Web language (OWL) with correspondence to description logics (DL). Protégé 2000 was used (with Owl plug-in) as a knowledge representation editor.

Evaluation

Although there is no uniform pattern for evaluating all ontologies we check the appropriateness of an ontology for its intended application. We evaluate the ontology for completeness, consistency and conciseness [12]. The scientific integrity of the ontology is assessed by posing queries using description logic. DLs provide reasoning capability and we use RACER [15] as a description logic reasoning system with support for T-Box (concepts) and A-Box (instances).

Semantic Query

The Semantic Web is expected to include many kinds of query-answering services with access to numerous types of information represented in widely different formats. To permit complex queries to integrated data, mapping of these database entities to ontological concepts is necessary. In the domain of interest the different biological data sources these entities are enzymes, organisms (fungal species) and enzyme properties and enzyme substrates.

To conduct such queries we are currently use OWL-QL [16] and nRQL [17] (New RACER Query Language) as query languages. nRQL is a pragmatic A-Box query language with formal semantics. It inherits from the semantics of standard DL the A-Box retrieval functions. nRQL unifies in a declarative way the majority of RACER's A-Box querying functions. Using RACER with nRQL as a Semantic Web repository enables the construction of more flexible queries in domain of interest. Figures 3 and 4 describe further sample queries written in nRQL syntax with the answers from the FungalWeb Ontology returned by Racer. These hand-written queries have value to the scientific investigator yet also demonstrate efficient query-answering with RACER with a much improved response time, beyond that of the average biologist looking up the same information. A further scenario of realistic scientific query from the FungalWeb Ontology will be included in the complete paper.

1-Give me all instances of Neolectaceae?

```
retrieve (?x) (?x |http://a.com/ontology#Neolectaceae| )
```

2-is there any instance for Pezizomycotina at all?

```
retrieve () (?x |http://a.com/ontology#Pezizomycotina|)
```

3-is Amorphotheca resinae a Basidiomycota?

```
retrieve () (|http://a.com/ontology#Amorphotheca_resinae| |http://a.com/ontology#Basidiomycota| )
```

4-All Agaricales has been reported to have enzyme Laccase?

```
retrieve (?x) (AND (?x |http://a.com/ontology#Agaricales|)(?x |http://a.com/ontology#Laccase|
|http://a.com/ontology#Has_been_reported_to_have_enzyme|))
```

5-All Enzyme has been reported to be found in Neurospora crassa?

```
retrieve (?x) (AND (?x |http://a.com/ontology#Enzyme|)(?x |http://a.com/ontology#Neurospora_crassa|
|http://a.com/ontology#Has_been_reported_to_be_found_in|))
```

6- I am looking for all Fungi which has been reported to have both enzyme Laccase and Cellulase.

```
retrieve (?x) (AND (AND (?x |http://a.com/ontology#Fungi|)(?x |http://a.com/ontology#Laccase|
|http://a.com/ontology#Has_been_reported_to_have_enzyme|))(?x |http://a.com/ontology#Cellulase|
|http://a.com/ontology#Has_been_reported_to_have_enzyme|) ) ))
```

7- I am looking for all Fungi which has been reported to have enzyme Laccase or Cellulase.

```
retrieve (?x) (AND (?x |http://a.com/ontology#Fungi|)(OR (AND (?x |http://a.com/ontology#Laccase|
|http://a.com/ontology#Has_been_reported_to_have_enzyme|)(?x |http://a.com/ontology#Cellulase|
|http://a.com/ontology#Has_been_reported_to_have_enzyme|))(OR (?x |http://a.com/ontology#Laccase|
|http://a.com/ontology#Has_been_reported_to_have_enzyme|)(?x |http://a.com/ontology#Cellulase|
|http://a.com/ontology#Has_been_reported_to_have_enzyme|)))
```

8.Which Enzymes are being used in baking and brewing?

```
retrieve (?x) (AND (AND (?x |http://a.com/ontology#Enzyme|)(?x |http://a.com/ontology#Baking|
|http://a.com/ontology#Is_being_used_in|))(?x |http://a.com/ontology#Brewing|
|http://a.com/ontology#Is_being_used_in|) )
```

Figure 3: nRQL Sample Queries

Conclusion

We have used semantic web technology to create a ontology and a large knowledgebase in the domain of fungal genomics from trusted biological sources to provide unified semantic access to these heterogeneous sources. Ongoing research involves improvement of querying capability through the provision of:

1. an integrated global view and schema of all concepts / instances;
2. tools for user formulated querying using ontologically defined terms,
3. system reformulation of queries to produce sub-queries for component resources.

Our goal is to provide a user with answers without requiring them to have knowledge of query methodology Further additions to the ontology to cover microarray data will draw on the MAGE object model MGED. For coverage of metabolic pathways and regulatory networks we will use the model of BioCyc [18] as a starting point. The capabilities of Racer and nRQL query language fit naturally as a back end to conventional Natural Language Processing (NLP) question-answer front ends. Such natural language querying capabilities have been effective in focused domains like FungalWeb.

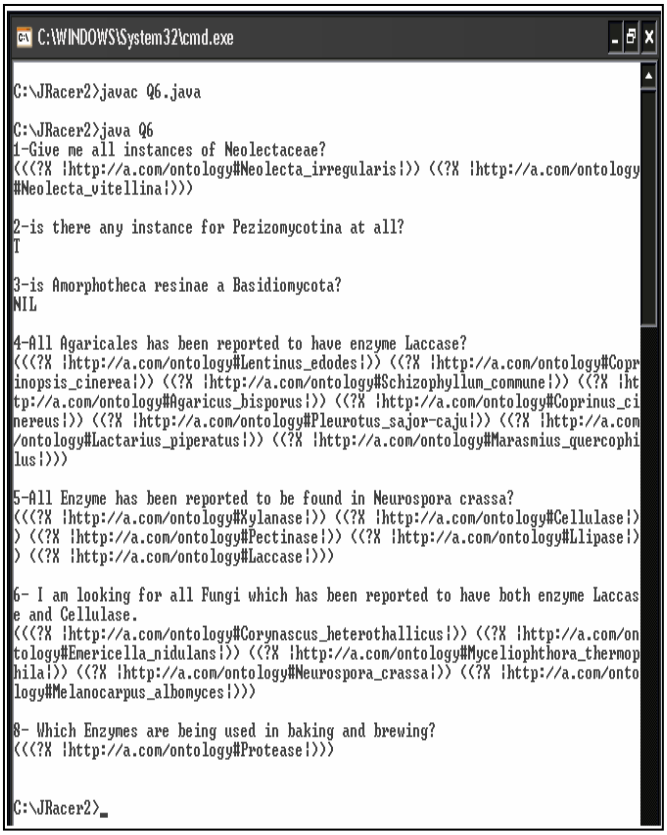

Figure 4: nRQL Sample Answers

Acknowledgements

The project FungalWeb: "Ontology, the Semantic Web and Intelligent Systems for Genomics" is funded by Génome Québec.

References:

1. Quan D., Martin S., Grossman D, Applying Semantic Web Techniques to Bioinformatics
2. Berners-Lee T., Hendler J., Lassila O., The Semantic Web, 2001
3. Stevens R. A Semantic Web of Bioinformatics Resources, Fifth Annual Bio-Ontologies Meeting, 2002
4. NCBI (http://www.ncbi.nlm.nih.gov/)
5. NEWT (http://www.ebi.ac.uk/newt/index.html)
6. Saccharomyces Genome Database (http://www.yeastgenome.org/)
7. Neurospora crassa Database (http://www.broad.mit.edu/annotation/ fungi/ neurospora/)

8. Brenda (http://www.brenda.uni-koeln.de/)
9. SwissProt (http://ca.expasy.org/sprot/)
10. ChEBI (http://www.ebi.ac.uk/chebi/)
11. Gómez-Pérez, A. A Framework to Verify Knowledge Sharing Technology, 1996
12. The Protégé Ontology Editor and Knowledge Acquisition System (http://protege.stanford.edu/)
13. OWL Web Ontology Language Guide
14. (http://www.w3.org/TR/owl-guide/)
15. RACER (http://www.cs.concordia.ca/~haarslev/racer/)
16. OWL-QL Project for the Stanford Knowledge Systems Lab. (http://ksl.stanford.edu/projects/owl-ql/)
17. nRQL user Guide (http://www.cse.concordia.ca/%7Ehaarslev/racer/racer-queries.pdf)
18. Karp P.D., Arnaud M., Collado-Vides J., Ingraham J., Paulsen I.T., Saier M.H. Jr. (2004). "The *E. coli* EcoCyc Database: No Longer Just a Metabolic Pathway Database." ASM News 70(1): 25-30.