

Two Layer Cakes

Syed Hamid Tirmizi^o and Daniel Miranker^{*}

^{*}Institute for Cell and Molecular Biology

^o* Department of Computer Sciences

The University of Texas at Austin

{hamid, miranker}@cs.utexas.edu

Extended Abstract

The penetration of the Semantic Web into bioinformatics is not like other disciplines. Given critical data resources such as the Gene Ontology (GO) and the challenges associated with the explosion in the number biological databases, biologists, as a community, are already familiar with the advantages of ontologies and the challenges of information integration. Rather than starting from the pedagogical beginnings, penetration will depend on a migration path from the Open Biomedical Ontologies language (OBO) to the Semantic Web. OBO emerged from GO, and is now host to over 60 different ontologies.

We have methodically examined each of the constructs of OBO and mapped them to constructs in the Semantic Web stack. We find that most of OBO can be decomposed into layers with direct correspondence to the Semantic Web layer cake. In the process we have enumerated constructs in each system that do not have a simple syntactic equivalent in the other. Elements of OBO “missing” in the semantic web are few, and can still be expressed in OWL. Thus, OBO ontologies may be translated to the Semantic Web. Further, we believe if certain ancillary information is retained during translation, the Semantic Web representation may be translated back to OBO, and the cycle repeated without any loss of knowledge. It is our expectation that tools to automate this process will enable important legacy ontologies onto the Semantic Web.

OBO Layer Cake

To create a transformation mechanism between OBO flat files and Semantic Web technologies, we find it useful to create a layer cake for OBO, similar to that of the Semantic Web. OBO tags can be partitioned into three layers – OBO Core, OBO Vocabulary, and OBO Ontology Extensions. See Figure 1. The correspondence between the layers of the two systems is direct. In addition, we describe the expressivity of OBO format in terms of various versions of OWL, i.e. OWL Lite, OWL DL and OWL Full.

- In OBO terminology, a concept can either be a term (class) or a typedef (relationship type). OBO Core deals with assigning IDs and namespaces to concepts, and representing some knowledge about those concepts using relationships; essentially triples (a connection worthy of additional investigation).
- OBO Vocabulary allows annotating concepts with information like names, definitions and comments. In addition, it supports describing sub-class and sub-property relationships, as well as the domains and ranges of typedefs.
- In contrast to the previous two layers, which define tags with concept-level scope only, OBO Ontology Extensions (OBO-OE) layer defines tags for expressing metadata on the entire ontology as well. It also allows defining synonyms, equivalences and deprecation of OBO concepts. Using OBO Ontology Extensions, we can also express specific properties of OBO terms (e.g. set combinations, disjoints etc.),

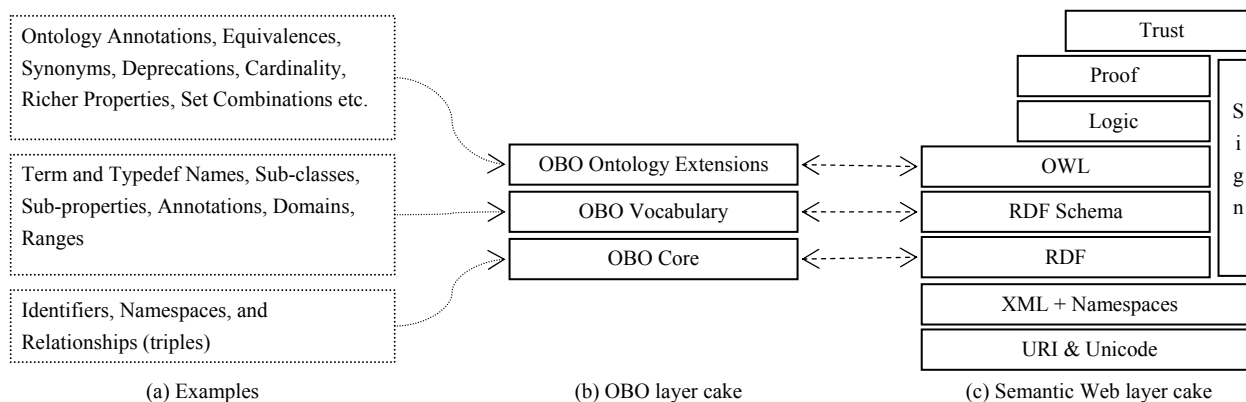


Figure 1: A layer cake for OBO, with some examples and a comparison with Semantic Web layers.

and typedefs (e.g. transitivity, uniqueness, symmetry, cardinalities etc.).

Our preliminary investigation shows that a major portion of OBO-OE maps to OWL Lite and provides the same level of expressiveness. Overall, OBO-OE matches well with OWL DL. In OBO the definition of a term, or a typedef, is rigid and not as expressive as OWL Full. Recall, the primary concern is the migration of legacy OBO ontologies and their constituencies to the Semantic Web. Thus, that OBO is less expressive than OWL Full is the convenient direction of containment. It does mean that round trips can not be supported unless the editing of an OBO ontology while in OWL representation is restricted.

Transformations for Lossless Roundtrips

In our work, we define a set of transformation rules for converting OBO files to OWL. Since we have an exact mapping of layers between the two formats, deciding which constructs to use for each kind of transformation becomes a lot easier. In other words, OBO Core tags can be transformed using RDF, OBO Vocabulary tags require using RDF Schema constructs, and OBO-OE requires us to use constructs defined in OWL.

A substantial portion of OBO tags can be transformed directly to corresponding OWL elements (e.g. names, comments, relationship cardinalities). Another significant subset of OBO tags can be transformed by defining some annotation property elements using RDF Schema.

In certain cases, however, the mapping is not very clear due to lack of well-defined or documented semantics of the OBO tags. For example, the semantics and use cases of multiple kinds of synonym tags and Dbxrefs are not very clear. Another related example is the transformation of non-URI based namespaces and identifiers into RDF/XML supported form, URIs. Defining transformation rules for such cases is significantly more complicated.

Biology and the Semantic Web

Building ontologies is not a new idea for the biology community. However, the utility of ontologies has not been fully realized due to lack of language support for querying ontologies in formats like OBO and performing rule-based inferences on them. Also, due to lack of global naming schemes, it has been very hard to obtain a global view by merging different ontologies.

Semantic Web is the idea that once fully realized, can solve these problems. The use of URIs for assigning globally unique identifiers to concepts is one of the foundations of Semantic Web. Querying languages like SPARQL for ontologies expressed in RDF, and work is in progress on defining languages for rules and inference on the ontologies. We believe our work can draw an easy

picture for the biology community and expose the world of Semantic Web to them.

Our application context is the NSF's "Assembling the Tree of Life" (AToL) grand challenge. The grand challenge faced is in describing 5 to 10 million extant species, and computing and analyzing a unified phylogenetic tree. The effort spans organisms as far ranging as bacteria, plants and mammals. Numerous projects, organized around a particular group, e.g. fish, are organizing the terminology of their corpuses as ontologies and working to exploit Internet technology to tie this information together and make it highly available. A goal of our project, Morphster, includes image annotation. In our context, images are used to document the precise meaning of a biological concept. Thus, in our ontologies an image, or parts of an image will become part of a concept definition. We anticipate both drawing concept labels from existing ontologies, (Mabee 2006), and adding new concepts to ontologies through image-based definitions.

Acknowledgements

This research is supported by NSF grants IIS-0531767 and IIS-0325116.

References

- Antoniou, G., and van-Harmelen, F. eds. 2004. *A Semantic Web Primer*. MIT Press.
- Dublin Core Metadata Initiative. Website (accessed June 15, 2006). <http://dublincore.org/documents/dces/>
- Koivunen, M., and Miller, E. 2002. W3C Semantic Web Activity. In *Proceedings of Semantic Web Kick-off in Finland*, 27-33, Helsinki, Finland.
- Lee, T. 2002. *The Semantic Web*. Academic Talk. W3C World Wide Web Consortium.
- Mabee, P.M., et al. 2006. *ZFIN Anatomy Working Group: Skeletal System*. Manually curated data.
- Mungall, C. 2005. Mapping OBO to OWL. Web page and XSLT transformations (accessed June 15, 2006). www.godatabase.org/dev/doc/mapping-obo-to-owl.html
- Open Biomedical Ontologies web site (accessed June 15, 2006). <http://obo.sourceforge.net/>
- SPARQL Query Language for RDF. 2006. W3C Candidate Recommendation. Web page (accessed June 15, 2006). <http://www.w3.org/TR/rdf-sparql-query/>
- The Gene Ontology. Website (accessed June 15, 2006). <http://geneontology.org/>
- The OBO Flat File Format, specifications (accessed June 15, 2006). <http://www.geneontology.org/GO.format.shtml>