INTEGRATING HIERARCHICAL CONTROLLED VOCABULARIES WITH OWL ONTOLOGY: A CASE STUDY FROM THE DOMAIN OF MOLECULAR INTERACTIONS[*]

MELISSA J. DAVIS

Institute for Molecular Biosciences and ARC Center of Excellence in Bioinformatics,
University of Queensland, St. Lucia
Brisbane, QLD 4076, Australia

ANDREW NEWMAN

School of Information Technology and Electrical Engineering,
University of Queensland, St. Lucia
Brisbane, QLD 4076, Australia

IMRAN KHAN

School of Information Technology and Electrical Engineering,
University of Queensland, St. Lucia
Brisbane, QLD 4076, Australia

JANE HUNTER

School of Information Technology and Electrical Engineering,
University of Queensland, St. Lucia
Brisbane, QLD 4076, Australia

MARK A. RAGAN

Institute for Molecular Biosciences and ARC Center of Excellence in Bioinformatics,
University of Queensland, St. Lucia
Brisbane, QLD 4076, Australia

Many efforts at standardising terminologies within the biological domain have resulted in the construction of hierarchical controlled vocabularies that capture domain knowledge. Vocabularies, such as the PSI-MI vocabulary, capture both deep and extensive domain knowledge, in the OBO (Open Biomedical Ontologies) format. However hierarchical vocabularies, such as PSI-MI which are represented in OBO, only represent simple

[*] This work is partially supported by ARC grant CEO348221, and has been performed as part of a collaborative research and development project with Pfizer Global R&D.

parent-child relationships between terms. By contrast, ontologies constructed using the Web Ontology Language (OWL), such as BioPax, define many richer types of relationships between terms. OWL provides a semantically rich structured language for expressing classes and sub-classes of entities and properties, relationships between them and domain-specific rules or axioms that can be applied to extract new information through semantic inferencing. In order to fully exploit the domain knowledge inherent in domain-specific controlled vocabularies, they need to be represented as OWL-DL ontologies, rather than in formats such as OBO. In this paper, we describe a method for converting OBO vocabularies into OWL and class instances represented as OWL-RDF triples. This approach preserves the hierarchical arrangement of the domain knowledge whilst also making the underlying parent-child relationships available to inferencing engines. This approach also has clear advantages over existing methods which incorporate terms from external controlled vocabularies as literals stripped of the context associated with their place in the hierarchy. By preserving this context, we enable machine inferencing over the ordered domain knowledge captured in OBO controlled vocabularies.

1. Introduction

Molecular biology as a field encompasses several dynamic sub-domains undergoing rapid expansion with resultant rapid discovery and growth in acquired data. High-throughput techniques and large scale biological research, such as genome and transcriptome sequencing projects and expression studies, generate abundant data. However, a significant gap exists between data acquisition and knowledge discovery. These massive quantities of data are frequently produced through a distributed effort, and need to be integrated for analysis and final presentation [1, 2]. Likewise, techniques such as microarray expression profiling produce large quantities of raw data which must be recorded and described [3]. In addition to data exchange and integration issues, many projects in computational and systems biology focus on the analysis of high-level properties of biological systems. Such projects might, for example, compare the distribution of protein functional classes between genomes [4], or analyse the genetic regulatory network of an organism [5]. Such analysis requires the integration of heterogeneous information produced from multiple sources at varying levels of resolution and described using highly variable terminologies. Solutions to these exchange and integration challenges include provision of the data in delimited text files, databases, and XML documents conformant with a given XML schema. The semantic meaning of the data is not however explicit within the documents, and relies on some external definition of concepts and relationships in the data.

Analysis at the level of biological systems also requires reasoning over large and complex data sets is beyond the ability of humans. Machine reasoning has the ability to uncover implicit relationships in the data, rather than simply retrieving explicitly represented data, as is the case of querying a database. However, machine reasoning over large and complex data sets requires the use of appropriate and meaningful knowledge representations of the domain area combined with presentation of the data in machine-readable format [6]. One technique for implementing knowledge representation that has been readily adopted in biology has been the construction of bio-ontologies to establish a precisely (if not formally) defined way to model and express the knowledge of a domain in terms of defined concepts: the classes of “things”, the relationships that exist between these classes, and the rules or axioms that apply to these concepts in the domain. The need for ontology development was recognized almost a decade ago in the creation of perhaps the most widely adopted biological ontology, the Gene Ontology (GO; http://www.geneontology.org; [7]).

An important consideration is why choose to create or use an ontology over traditional database schema, which are widely adopted knowledge representations in the field of molecular biology [8]. While there have been few successful demonstrations of automated inference using bio-ontologies, there are significant reasons for their adoption [9]. The key reason is that ontologies are designed to evolve over time and to facilitate integration of data, while database schemas are not [10]. Database schemas are typically considered an internal design decision for a given application and rarely, if ever, are schema from other databases reused. A specific ontology, on the other hand, is an external, global resource that is meant to be reused, extended and integrated with other ontologies. An ontology is also more expressive than a database [11]. Finally, databases rarely allow the preservation of data. It is still common simply to add attributes to an existing schema rather than splitting it logically due of the extent of the data migration [12]. Ontologies provide a separation between the actual data and the metadata or descriptions of the datasets and their relationships. This allows the data to be migrated independently from changes within the ontology [10].

Many types of knowledge representation exist, and there are many views of what constitutes machine reasoning [13]. The majority of ontologies developed in the biological domain to date do not take advantage of this background [14]. Technologies developed by the World Wide Web Consortium (W3C; http://www.w3.org/) to support machine reasoning include the Resource Description Framework (RDF; http://www.w3.org/RDF/) and Web Ontology Language (OWL; http://www.w3.org/2004/OWL/). Both RDF and OWL support machine inference across resources on the web.

While some bio-ontologies have been constructed using these standards, (see http://www.obofoundry.org) or have been converted into a form compliant with these standards [15], most do not take advantage of the W3C recommendations [14]. Many are presented as controlled vocabularies where concepts are represented taxonomically, and relationships are predominantly "is-a" or "part-of" relationships that establish the tree-like structure of the vocabulary. While these structures create well ordered catalogues of concepts relevant to a domain, they do not typically allow for the expression of rules defining other relationships between concepts. The end result is a simplified, flattened model of the domain that lacks the semantic depth or logical support to enable a reasoner to infer new relationships or new information.

A significant challenge in molecular biology is to understand the molecular interactions that occur within cells; research in cell and structural biology shows that many proteins rely on a complex network of interacting partners to achieve their correct localization and functional state in the cell. High-quality molecular interaction data are largely described in journal articles using natural language. Because of the unstructured nature of the observations, the discipline of molecular interactions is covered by several overlapping ontologies. Some of these, such as BioPax (http://www.biopax.org) and the Protein Standards Initiative Molecular Interaction vocabulary (PSI-MI; http://www.psidev.info/) provide significant coverage over concepts relevant to the domain, while others, such as GO, Sequence Ontology (http://www.sequenceontology.org/) and the NCBI-Taxonomy (http://www.ncbi.nlm.nih.gov/) intersect with the field. This diverse set of overlapping ontologies, and the diversity of formats in which they are presented, make molecular interactions a useful test domain for strategies to integrate bio-ontologies and reuse domain knowledge.

2. Results

The field of bio-ontology development is active, and already populated with a number of general and domain specific ontologies that have been developed, or are under development. We reviewed ontologies listed by the Open Biomedical Ontology Foundry (OBO Foundary; http://www.obofoundry.org/) and the National Center for Biomedical Ontology (NCBO; http://www.bioontology.org/). Of the ~70 ontologies listed at these sites, around three quarters are written using the OBO format [16], with the remainder using other formats, including OWL, Protégé files and plain text.

OWL is specifically designed to construct ontologies that support machine reasoning [11]. For this reason, we choose to use OWL description logic (DL) to construct a high-level ontology to integrate concepts from relevant biological ontologies and vocabularies not expressed in OWL. Given that knowledge acquisition is one of the most time consuming and necessarily manual parts of ontology construction, the knowledge captured in non-OWL ontologies constitutes a valuable resource. One approach suggested in such cases is to construct a new ontology using OWL [14], however this under estimates the value of knowledge represented in other formats. Practical strategies to rescue domain knowledge captured in non-OWL ontologies will have obvious applicability in a domain such as molecular biology where the majority of vocabularies are not expressed in OWL.

We have reviewed two ontologies used to describe molecular interactions: the OWL ontology BioPax, and the OBO ontology PSI-MI. However it is not our intention to comparatively evaluate these ontologies, as has been done recently [17, 18]. Briefly, BioPax is designed to describe pathway rather than specific molecular interaction data. However of the ~40 classes and ~70 properties that BioPax defines, many are key concepts and relationships necessary to describe molecular interactions. The PSI-MI vocabulary, on the other hand, is specifically designed to describe molecular interaction data and captures >800 concepts from the domain. However it is represented in OBO and expresses only hierarchical relationships between these classes. While BioPax lacks the descriptive power of PSI-MI, it is more suited for machine reasoning because it is represented in OWL. BioPax recognizes the value of external controlled vocabularies such as PSI-MI and GO, by providing a facility to exploit these external vocabularies through the inclusion of a class openControlledVocabulary. This class stores a term from an external vocabulary along with a cross reference holding the identifier of that term and the name of the vocabulary (as literal strings).

However, it is not sufficient merely to store the data from external controlled vocabularies as literals - the term becomes devoid of meaning if taken out of the context of the concept tree within the originating vocabulary. In order to preserve the meaning of the term, its relationship to other terms and its place in the hierarchy must also be preserved. To illustrate this point, consider the natural language expression, “The protein Emerin is localized to the nuclear inner membrane” (Figure 1).

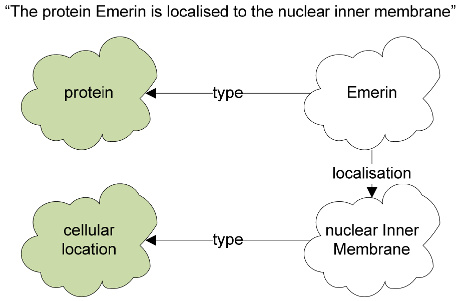

Figure 1 A natural language assertion decomposed into generic concepts, in green, and specific instances of those concepts in white. The specific instance “Emerin” is of the generic type *protein*, while the specific instance “Nuclear inner membrane” is of the generic type *cellular location*. The relationship between “Emerin” and its localisation is represented by a labeled arrow, *localisation*.

This statement is composed of generic and specific concepts and relationships: two generic concepts, protein and cellular location, and two specific instances of these concepts, “Emerin”, a protein, and “nuclear inner membrane”, a cellular location. A relationship also exists between these two instances, namely, that “Emerin” has a property localisation, the value of which is “nuclear inner membrane”.

The same statement could also be expressed using the BioPax ontology, as illustrated in Figure 2, by using the openControlledVocabulary class to include a term from an external vocabulary like the GO Cellular Component hierarchy.

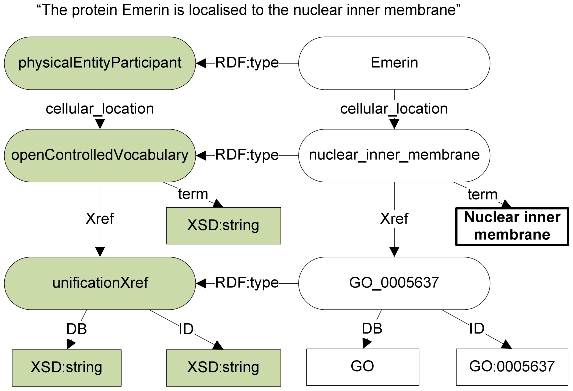

Figure 2 Expression of an assertion of subcellular location in BioPax. Classes and properties from the BioPax ontology are depicted in green on the left of the diagram, while the instance data describing the localisation of Emerin is on the right. Elipses represent classes, or instances of classes, while rectangles represent typed literals, in this case, strings. The term in bold, “nuclear inner

membrane" has been imported from the GO Cellular Component hierarchy, and is included here as a string.

However, as the references to the term from the external vocabulary are all simple text strings, they lack context or meaning A search for proteins where the value of cellular_location is the string "organelle membrane" would not retrieve proteins where this value was "nuclear inner membrane" unless the query application was hard-coded with additional information about the relationship between these two strings. External terms used in this fashion lack meaning. Because of the limitation of the BioPax approach, we developed a different approach to include domain knowledge captured in external controlled vocabularies (see Figure 3).

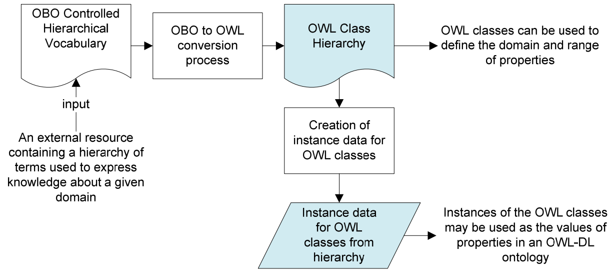
OBO Controlled Hierarchical Vocabulary
OBO to OWL conversion process
OWL Class Hierarchy
OWL classes can be used to define the domain and range of properties
input
An external resource containing a hierarchy of terms used to express knowledge about a given domain
Creation of instance data for OWL classes
Instance data for OWL classes from hierarchy
Instances of the OWL classes may be used as the values of properties in an OWL-DL ontology

Figure 3 Conversion process for an OBO Controlled Vocabulary. The OBO vocabulary is converted into OWL-DL. This results in an OWL class hierarchy where terms from the original vocabulary become classes in the OWL ontology. Instance data is then created for the ontology so each class has a single instance comprising the vocabulary term that can then be used as an object or subject of triples. Outputs from this process, an OWL class hierarchy and associated instance data, are represented in blue.

An external controlled vocabulary that contains relevant descriptive terms is converted into OWL-DL. Most frequently, the external vocabulary is in OBO format, so we currently use the OBO to OWL conversion application [19], however this process may be more generally applied to any hierarchical controlled vocabulary. It is important that both the class hierarchy and the instance data for this hierarchy are created. OWL-DL classes are used to define restrictions for properties within the ontology, through the specification of allowable domain and range values [11]. However, a class cannot also be an instance, and only instances may be used as the values of properties. For this reason, a single instance of each class is created, taking the form of the original term from the hierarchical vocabulary. At the end of this process, both an OWL ontology representing the terms from the controlled vocabulary and a set of instance data are available to use in conjunction with an OWL ontology, as shown for an excerpt from GO Slim [20] in Figure 4.

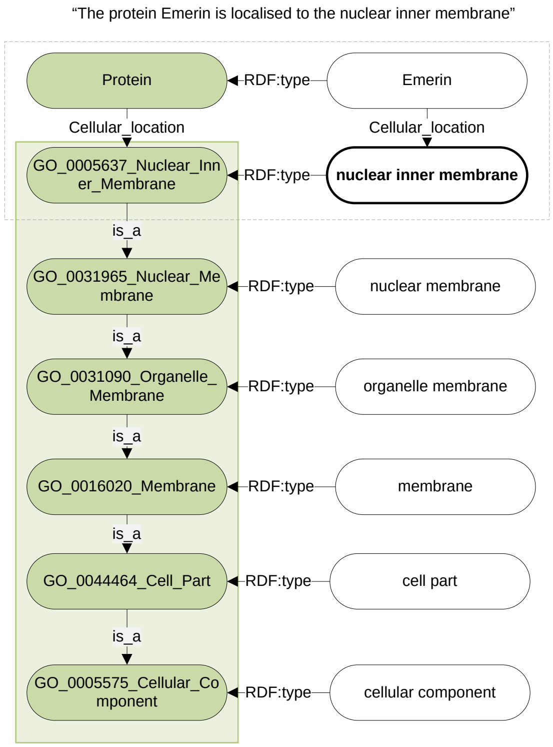

Figure 4 Application of the conversion process. Classes and instances used to express the statement are boxed in a grey dashed line. The term, “nuclear inner membrane” in bold is an instance of the class GO_0005637_Nuclear_Inner_Membrane, which is related to other classes in the hierarchy boxed in green. Domain knowledge is explicitly captured in these hierarchical relationships, so that the relationship of the protein “Emerin” to other cellular locations can be inferred.

In this example, the instance “nuclear inner membrane” is used as the value of the *cellular_location* property. Not only is this value meaningful to a human reasoner who understands what is meant by the words, it is also meaningful to a machine reasoner that has access to the underlying ontology. This machine reasoner, when presented with the assertion that Emerin is located in the “nuclear inner membrane”, may correctly infer that Emerin is also located in an “organelle membrane”, and located in a “membrane”. By using this process to incorporate components of external vocabularies under a high-level extensible ontology, external terms become more than text labels, and enable implicit relationships to be extracted from explicit data.

3. Discussion

Many of the existing bio-ontologies are written in the OBO format, and represent a rich source of biomedical domain knowledge. By using the approach described here, vocabularies covering specific aspects of a domain may be plugged into a high level extensible OWL ontology designed to facilitate these modular extensions. The parent-child relationships of the new vocabulary are maintained, and made available to a reasoner to infer implicit relationships from the explicitly represented data. The current strategy used by BioPax is inadequate for machine reasoning. To make bio-ontologies useful for machine reasoning, they need to be explicitly represented in languages such as OWL. Converting and importing relevant hierarchies of terms and associated instances is one solution to maintaining the meaning of terms in controlled vocabularies.

One concern when creating an ontology or expanding an existing ontology is the trade off between the ability of the ontology to express concepts in the domain, and to provide tractable inferencing, and the effects of this trade off are difficult to evaluate [21]. A strategy which we will evaluate to address this is to identify which branches of a given concept tree are required and only convert those branches, as meaning contained in the concept hierarchy outside of the branch would not be referenced.

Since so many biological concepts are framed in terms of hierarchically inherited properties, and the majority of biological ontologies take the form of hierarchical controlled vocabularies, the process described here is a useful generic strategy for incorporating the existing wealth of ordered knowledge into a semantically rich ontology constructed using OWL. This will help to extend the utility of bio-ontologies into the arena of machine inference.

References

1. Collins, F., M. Morgan, and A. Patrinos, The Human Genome Project: Lessons from Large-Scale Biology. Science, 2003. 300: p. 286-290.
2. Carninci, P., et al., The transcriptional landscape of the mammalian genome. Science, 2005. 309(5740): p. 1559-63.
3. Brazma, A., et al., Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet, 2001. 29(4): p. 365-71.
4. Andrade, M.A., et al., Functional classes in the three domains of life. J Mol Evol, 1999. 49(5): p. 551-7.
5. Li, S., et al., A map of the interactome network of the metazoan C. elegans. Science, 2004. 303(5657): p. 540-3.

6. Bry, F. and M. Marchiori. Reasoning on the semantic web: Beyond ontology languages and reasoners. in 2nd European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies. 2005. London, United Kingdom: IEEE.
7. Ashburner, M., et al., Gene Ontology: tool for the unification of biology. Nature Genetics, 2000. 25: p. 25-29.
8. Galperin, M.Y., The Molecular Biology Database Collection: 2007 update. Nucleic Acids Res, 2007. 35(Database issue): p. D3-4.
9. Keet, M., M. Roos, and M. Marshall, A survey of requirements for automated reasoning services for bio-ontologies in OWL, in OWLED 2007: Third International Workshop on OWL Experiences and Directions. 2007, CEUR-WS: Innsbruck, Austria.
10. Noy, N. and M. Klein, Ontology Evolution: Not the Same as Schema Evolution. Knowledge and Information Systems, 2004. 6: p. 428-440.
11. Bechhofer, S., et al., OWL Web Ontology Language Reference, in W3C Recommendations, M. Dean and G. Schreiber, Editors. 2004, World Wide Web Consortium.
12. Elmasri, R., S. Navathe, and C. Shanklin, Fundamentals of Database Systems. 2000, Boston: Addison-Wesley.
13. Davis, R., H. Shrobe, and P. Szolovits, What is a Knowledge Representation?, in AI Magazine. 1993. p. 17-33.
14. Soldatova, L. and R. King, Are the current ontologies in biology good ontologies? Nature Biotechnology, 2005. 23: p. 1095-1098.
15. Aranguren, M.E., et al., Understanding and using the meaning of statements in a bio-ontology: recasting the Gene Ontology in OWL. BMC Bioinformatics, 2007. 8: p. 57.
16. Cote, R.G., et al., The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. BMC Bioinformatics, 2006. 7: p. 97.
17. Stromback, L. and P. Lambrix, Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX. Bioinformatics, 2005. 21(24): p. 4401-7.
18. Stromback, L., et al., Representing, storing and accessing molecular interaction data: a review of models and tools. Brief Bioinform, 2006. 7(4): p. 331-8.
19. Moreira, D.A. and M.A. Musen, OBO to OWL: A Protege OWL Tab to Read/Save OBO Ontologies. Bioinformatics, 2007.
20. Martin, D., et al., GOToolBox: functional analysis of gene datasets based on Gene Ontology. Genome Biol, 2004. 5(12): p. R101.
21. Levesque, H. and R. Brachman, Expressiveness and tractability in knowledge representation and reasoning. Computational Intelligence, 1987. 3: p. 78-93.