CONFERENCE PROCEEDINGS

ICBO

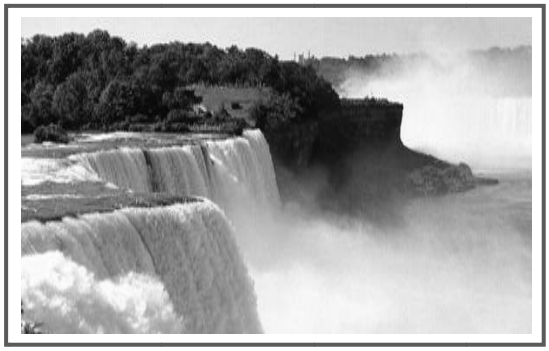
International Conference on Biomedical Ontology

July 24-26, 2009
Buffalo, New York, USA

SPONSORED BY
University at Buffalo College of Arts and Sciences
National Center for Ontological Research
National Center for Biomedical Ontology
with support from
National Human Genome Research Institute

CONFERENCE PROCEEDINGS

Edited by Barry Smith



ICBO

International Conference on Biomedical Ontology

July 24-26, 2009
Buffalo, New York, USA

University at Buffalo
The State University of New York

ncor national center for ontological research

THE NATIONAL CENTER FOR BIOMEDICAL ONTOLOGY

PREFACE

Ontologies are being used in a variety of ways by researchers in almost every life science discipline, and their use in annotation of both clinical and experimental data is now a common technique in integrative translational research. When data from different sources are described using shared, logically structured, controlled vocabularies such as the Gene Ontology (GO), this makes the data more easily retrievable and navigable, and it also enhances the degree to which they can be analyzed and combined to serve new purposes.

Two major international conferences on Standards and Ontologies for Functional Genomics (SOFG), held at the Wellcome Trust Genome Campus in Hinxton, UK in 2002 and at the University of Pennsylvania School of Medicine in Philadelphia, PA in 2004, have thus far been devoted to the topic of biomedical ontology. They served to bring together researchers involved in the development and use of ontologies in addressing the data annotation needs created by the new high-throughput experimental techniques and new applications of comparative genomics in the investigation of human biology and disease.

In the years since SOFG, the number of applications of ontologies in biomedicine has expanded greatly, and so also has the range and functionality of associated software tools. The Web Ontology Language (OWL) serves as a central enabling technology of the Semantic Web, which is itself increasingly being used to serve computational biomedical research in reflection of the growth of the Internet as a medium for data exchange. Ontologies are now being used in almost every domain of biomedical research, including translational medicine, drug discovery, clinical trials, neuroimaging, environmental studies, evolutionary biology, and many other fields. Ontologies have been developed to describe data at all scales, from molecular pathways to mammalian anatomy and from protein modifications to infectious disease.

Increasingly, it is recognized that these different ontologies need to work together in order to maximize the degree to which they can serve the needs of researchers and clinicians, and it is the recognition of this need that led to the organization of the International Conference on Biomedical Ontology (ICBO) in Buffalo on July 24-26, 2009.

ICBO brings together scientists and informaticians developing and using ontologies across the entire spectrum of biology and clinical and translational medicine. It represents a continuation not only of the SOFG series but also of multiple dissemination events organized under the auspices of the National Center for Biomedical Ontology (NCBO), an NIH Roadmap Center for Biomedical Computing. We acknowledge the financial support of the NCBO, and also of the University at Buffalo College of Arts and Sciences, the National Center for Ontological Research (NCOR), and the National Human Genome Research Institute through award number R13 HG005049-01.

Barry Smith
Buffalo, New York, USA
July 2009

ICBO ORGANIZING COMMITTEE

Chair: Barry Smith, University at Buffalo, Buffalo, NY, USA
Co-Chair: Suzanna E. Lewis, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Michael Ashburner, Cambridge University, Cambridge, UK
Judith A. Blake, The Jackson Laboratory, Bar Harbor, ME, USA
Yves A. Lussier, University of Chicago, Chicago, IL, USA
Mark A. Musen, Stanford University, Stanford, CA, USA
Alan Ruttenberg, Science Commons, Cambridge, MA, USA
Susanna-Assunta Sansone, European Bioinformatics Institute, Hinxton, UK
Christian J. Stoeckert, Jr., University of Pennsylvania, Philadelphia, PA, USA

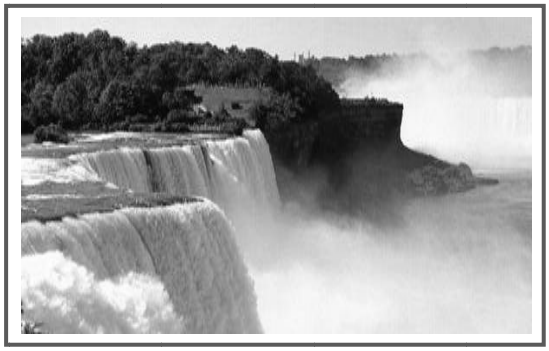
ICBO SCIENTIFIC COMMITTEE

Chair: Barry Smith, University at Buffalo, Buffalo, NY, USA
Co-Chair: Suzanna E. Lewis, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Michael Ashburner, Cambridge University, Cambridge, UK
Colin Batchelor, Royal Society of Chemistry, Cambridge, UK
Judith A. Blake, The Jackson Laboratory, Bar Harbor, ME, USA
Olivier Bodenreider, National Library of Medicine, Bethesda, MD, USA
Werner Ceusters, University at Buffalo, Buffalo, NY, USA
Lindsay G. Cowell, Duke University, Durham, NC, USA
Louis J. Goldberg, University at Buffalo, Buffalo, NY, USA
Lawrence Hunter, University of Colorado, Denver, CO, USA
Yves A. Lussier, University of Chicago, Chicago, IL, USA
Maryann E. Martone, University of California at San Diego, San Diego, CA, USA
Jose L. V. Mejino, Jr., University of Washington, Seattle, WA, USA
Mark A. Musen, Stanford University, Stanford, CA, USA
Christopher J. Mungall, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
Darren Natale, Georgetown University, Washington, DC, USA
Bjoern Peters, University of California at San Diego, San Diego, CA, USA
Helen Parkinson, European Bioinformatics Institute, Hinxton, UK
Alan Rector, University of Manchester, Manchester, UK
Peter N. Robinson, Charité Hospital, Berlin, Germany
Alan Ruttenberg, Science Commons, Cambridge, MA, USA
Susanna-Assunta Sansone, European Bioinformatics Institute, Hinxton, UK
Richard H. Scheuermann, University of Texas Southwestern Medical Center, Dallas, TX, USA
Nigam Shah, Stanford University, Stanford, CA, USA
David J. States, University of Texas, Houston, TX, USA
Christian J. Stoeckert, Jr., University of Pennsylvania, Philadelphia, PA, USA
Cathy H. Wu, Georgetown University, Washington, DC, USA; University of Delaware, Newark, DE, USA

TABLE OF CONTENTS

PAPERS



ICBO

International Conference on Biomedical Ontology

July 24-26, 2009
Buffalo, New York, USA

An OWL-DL Ontology for Classification of Lipids

Hong-Sang Low[1], Christopher J.O. Baker[2], Alexander Garcia[3], Markus R. Wenk[1]
[1]National University of Singapore, Singapore
[2]University of New Brunswick, Saint John, Canada
[3]University of Bremen, Bremen, Germany

Abstract

Lipids can be systematically classified according to functional properties, structural features, biochemical origin or biological system. However Lipid nomenclature has yet to become a robust research tool since no rigorous definitions exist for membership of specific lipid classes. Lipids need to be defined in a manner that is systematic yet at the same time semantically explicit. We report on the reuse of existing lipid nomenclature, ontology describing chemical structure and the extension of the OWL-DL Lipid Ontology to support the classification of lipid molecules. We applied definitions, DL-axioms, to describe lipids classes and illustrate suitability of the ontology for the classification of Fatty Acyl lipids and Mycolic acids.

Introduction

IUPAC-IUBMB proposed a systematic nomenclature for lipids which received limited adoption by the lipid community. The proposed classification was complicated and prone to erroneous application by scientists. Moreover the naming scheme was not extended and does not adequately represent many novel lipid classes discovered in the recent decades. As a result lipids still lack systematic classification and a nomenclature that is universally adopted by the biomedical research community. The LIPIDMAPS consortium[1] aims to resolve this by introducing a scientifically robust, comprehensive and extensible classification system evolved from the IUPAC nomenclature. This classification scheme organizes lipids from different phyla and synthetic domains yet uptake by the lipid community has been slow and the literature is steeped with instances of lipid synonyms that fail to reflect the new nomenclature.

Hierarchical Classification of Lipid Nomenclature

Lipids are organic compounds and can be systematically classified according to various features e.g. atomic connectivity, physicochemical properties, presence of functional groups, or types of bioactivities. Albeit an important contribution, the LIPIDMAPS central repository of lipids has primarily used is-a relationships[2] to categorize lipids and many definitions describing LIPIDMAPS lipid classes remain implicit. Moreover they are often dependent on a chemical diagram in the form a molecular graphic file that can only be accurately classified by a trained lipid expert. No rigorous definition, independent of a graphical diagram, exists and the graphical definitions are not flexible, nor are they extensible. Changes in such definitions require the redrawing of the chemical diagram/definition. Subsequently, communicating, storing and transferring of such structural definitions in the current format is inefficient and there is much reliance on trained experts. There is therefore a need for lipids to be defined in a manner that is systematic and explicit. A rigorous definition would involve a minimal necessary and sufficient declaration for class membership that could adequately describe a lipid without requiring a molecular structure diagram.

Description logics (DL) describe a domain using class descriptions according to a logic based semantic. In previous work DL has been used to represent chemical knowledge[3,4]. Using DL, it is possible to define a lipid with necessary conditions such that an alpha mycolic acid is defined as a lipid that minimally has alpha-hydroxyl acid and cyclopropane groups. Moreover, we can define necessary and sufficient conditions limiting the definition of an alpha mycolic acid to a lipid that has only alpha-hydroxyl acid and cyclopropane functional groups. Consequently molecules that have functional groups other than alpha-hydroxyl acid group and cyclopropane groups cannot be considered as an alpha mycolic acid.

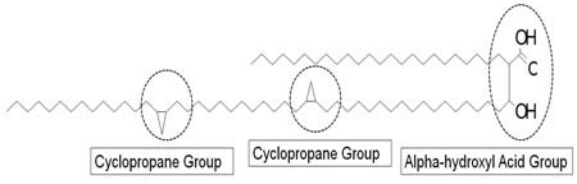


Figure 1. An example of alpha mycolic acid

The Lipid Ontology

The Lipid Ontology was exclusively developed to conceptualize and capture knowledge in the domain of lipids through the use of concepts, relations, instances and constraints on concepts[5]. It was designed to provide a common terminology for the

lipid domain, a basis for interoperability between information systems and to support navigation of text mining results from lipid literature. The ontology has been extended to describe the LIPIDMAPS nomenclature classification explicitly using description logics (OWL-DL) and to support reasoning and inference tasks. Prior to extending the ontology for classification tasks we reviewed existing chemo-ontologies[6,7] for reusable components. We reviewed the Chemical Ontology[6] for reuse of functional group specifications in the *Organic_Group* hierarchy. We enriched the Lipid Ontology with 32 functional groups from Chemical Ontology and 63 new concepts were added under the *Organic_Group* super-concept. The *Organic_Group* hierarchy was reorganized and asserted with new is-a relationships. From Chemical Ontology, we also used hasPart to relate concepts of lipids to concepts under *Organic_Group*. In reviewing the ChEBI[7] Ontology we identified that it is currently undergoing major revisions to correct inconsistent use of 'IsA' and 'IsPartOf' properties. We opted not to re-use its organization and relationship definitions, moreover to represent a systematic lipid nomenclature using formal logical definition of classes, we do not yet need all the relationship definitions found in ChEBI To further facilitate the reuse of the formal definitions in the lipid ontology we provide a high level alignment to ChEBI using SAMBO[12]. The alignment is available online at: http://www.lipidprofiles.com/LipidOntology/Others/SAMBO_0.rdf

Functional Groups Used in Lipid Classification

Lipids can have a wide range of distinct functional organic groups that should be accommodated in their conceptualization and classification. Distinct combinations of these organic groups underpin the definitions of lipid classes and membership of lipid classes can be restricted by formal descriptions which refer to functional groups. While the Chemical Ontology[6] describes basic functional groups, a wider range of functional groups are needed to describe lipids. To equip the Lipid Ontology for use as a classification tool we added 400 DL definitions to all lipid classes, with the exception of polyketides (Table 1). Primarily we re-used, from Chemical Ontology, the axiom "*Organic_Compound* hasPart *Organic_Group*" to relate *Lipid* concepts to *Organic_Group* concepts. We then defined concepts to describe lipid functional groups, namely *Organic_Group* and *Ring_System*. *Organic_Group* has three sub-groups (i) *Simple_Organic_Group*, (ii) *Complex_Organic _Group*, (iii) *Chain_Group*. *Simple_Organic_Group* subsumes concepts that describe basic functional groups whereas *Complex_Organic_Group* encapsulates glycans and amino acids. Glycans, in particular, are used to classify lipids such as sacharrolipids, and other sugar-linked lipids such as sphingolipids. *Chain_Group* consists of the *Carbon_Chain_ Group* and the *Sphingoid_Base_ Chain_Group*. The *Sphingoid_ Base_Chain_Group* is used exclusively for sphingolipids whereas *Carbon_Chain_ Group* is applied to other lipid classes accordingly. The *Ring_System* consists of (i) *Isoprenoid_ring_ derivative*, (ii) *Monocyclic_Ring_ Group* and (iii) *Polycyclic_Ring_System*. These concepts are used to define lipids that have one or more rings, primarily sterol, prenol and other ring lipids. In Lipid Ontology these concepts are extensively used to provide the necessary structural descriptions to define the identity of lipid-based compounds.

Total No. of Classes	715
No. of Lipid Classes	428
Primitive Lipid Classes	162
Defined Lipid Classes	266
Total No. of Restrictions	901
Total No. of Properties	41
DL Expressivity	ALCHIQ(D)

Table 1. Summary of the current Lipid Ontology

Hierarchical Classification of Lipids

Lipid concepts are organized hierarchically with the super-classes restricted by generic necessary conditions. More specific necessary conditions are used to define membership requirements for sub classes of lipid. At the end of a hierarchy, lipid classes are restricted by necessary and sufficient conditions and closure axioms. Super-classes are not closed by closure axiom to avoid inconsistency among disjointed sibling classes. More specific lipid classes are defined in two ways. In the first approach we specify the subclass of the present class to restrict the definition of a lipid. Necessary conditions such as "hasPart some *Carboxylic_Acid_derivative_ Group*" can be further specified by the subclass of *Carboxylic_Acid_Derivative_Group*, e.g. aldehyde. The second approach uses a Cardinality Axiom that restricts the number of a particular concept to be allowed in a restriction. Lipid classes can be defined by the number of certain functional group concept or *Chain_Group* concept. For example, a triacylglycerol is an acylglycerol with 3 acyl chains. Its superclass is restricted with an existential axiom "has some *Acyl_ Chain*". This is further specified with the following cardinality axiom "hasAcyl_Chain exactly 3". We are currently exploring Qualified Cardinality Axioms in

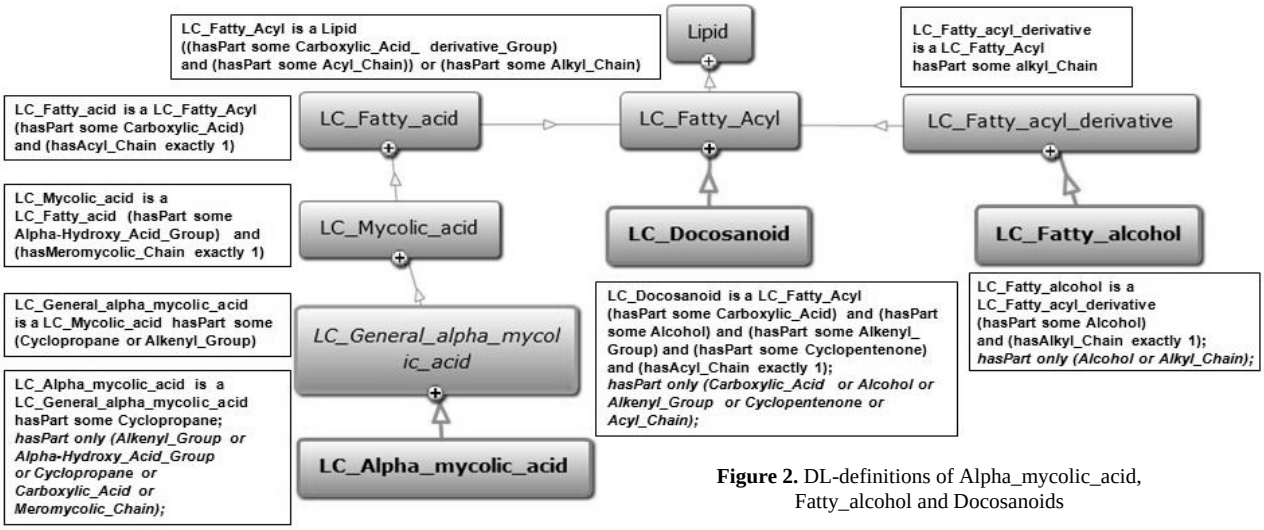


Figure 2. DL-definitions of Alpha_mycolic_acid, Fatty_alcohol and Docosanoids

OWL 2.0 as a means of defining lipid classes. For example we can define the class *Triacylglycerol* with "hasPart exactly 3 *Acyl_Chain*" without additional properties such as hasAcyl_Chain. However, an *Acyl_Chain* would have a carboxylic acid functional group encapsulate within it. When defining lipids with "hasPart exactly 3 *Acyl_Chain*", care must be taken not to add functional groups that entail others.

DL Axioms for the definition of Fatty_Acyl

Fatty acyls are a diverse lipid group synthesized by chain-elongation of an acetyl-CoA primer with malonyl-CoA/methylmalonyl-CoA groups[1]. We define a fatty acyl as a lipid that has at least one *Carboxylic_Acid_derivative_Group* and at least one *Acyl_Chain*. *Docosanoid* is a subclass of fatty acyls that inherits from *Fatty_Acyl* class the *Carboxylic_Acid_derivative_Group* as well as *Acyl_Chain*. The *Carboxylic_Acid_derivative_ Group* in *Docosanoid* is further specified to be a *Carboxylic_Acid*, whereas the *Acyl_Chain* is specified with a cardinality axiom and the property hasAcyl_Chain. Consequently, *Docosanoid* is defined to have only 1 *Acyl_Chain*. In addition, *Docosanoid* can have multiple and distinct functional groups such as *Carboxylic_Acid*, *Alkenyl_Group*, *Alcohol* and *Cyclopentenone*. These functional groups are associated with the class *Docosanoid* via the property "hasPart" in conjuction with the existential axiom "some". A closure axiom is needed to restrict the type of relationship constraints allowed for a lipid class. The closure axiom is applied to the class *Docosanoid* so that lipids of this class can only have the following functional groups, namely, *Carboxylic_Acid*, *Alkenyl_Group*, *Alcohol*, *Cyclopentenone* and *Acyl_Chain*. (Figure 2). As LIPIDMAPS nomenclature classifies lipids based on chemical structure or biosynthetic origin, lipids such as fatty alcohols are classified as fatty acyls in spite of having no *Acyl_Group*. When considered structurally, this classification of lipids is not systematic. We address this shortcoming in LIPIDMAPS nomenclature by expanding the definition of *Fatty_Acyl* to include *Alkyl_Chain*, a characteristic structure of un-usual *Fatty_Acyl* classes. In doing so a *Fatty_alcohol* inherits an *Alkyl_Chain* from *Fatty_Acyl* and is further defined to have a single *Alkyl_Chain* in the necessary and sufficient condition. This definition includes a "hasPart" property that connects *Fatty_alcohol* to an *Alcohol* class allowing inclusion of a lipid without an acyl group as a member of *Fatty_Acyl* (Figure 2). In addition we create a new lipid class, namely *Fatty_Acyl_derivative*, a subclass of *Fatty_Acyl*. Using the flexibility of OWL-DL, we can begin to address inconsistencies in lipid classification grounded in lipid definitions that are non systematic.

Extension of the Mycolic Acid Class

Mycolic acids are a key component of the cell wall of *Mycobacterium tubeculosis sps*. and are implicated mycobacterial disease. By 1998 there existed 500+ known chemical structures of related mycolates[8] and yet LIPIDMAPS currently contains only 3 mycolic acid records. Consequently many mycolic acids with known structures have yet to be systematically classified. Classification of these lipids is critical for system-level analysis of mycobacterial pathogenesis. We illustrate extension of Lipid Ontology to include new *Mycolic_Acid* classes and demonstrate classification of a real instance of an alpha mycolate (Figure 1) to the appropriate class. Based on LIPIDMAPS nomenclature, we assign *Mycolic_acid* as a member of *Fatty_Acid* and extend *Mycolic_acid*

classification to 9 defined subclasses (Table 2 http://www.lipidprofiles.com/LipidOntology/Others/Table2.jpg), distributed among three primitive superclasses. Alpha mycolic acid is a mycolic acid that has cyclopropane and alpha-hydroxyl acid (a special class of carboxylic acid) groups. The carboxylic acid group is a member of the acyl group, an ester group. Therefore, according to the classification scheme below, alpha mycolic acid must be a member of *Fatty_Acyl*. Among members of *Fatty_Acyl*, only *Octadecanoid, Docosanoid, Eicosanoid* and *Fatty_Acid* have carboxylic acid. Alpha mycolic acid does not have a cycloketone group and therefore, it cannot be *Docosanoid, Eicosanoid* or *Octadecanoid* and must be a member of *Fatty_Acid*. Among members of *Fatty_Acid*, only *Mycolic_acid* has an *Alpha-Hydroxy_Acid_Group* and a *Meromycolic_ Chain*. Therefore, alpha mycolic acid is classified under this class of *Fatty_Acid*. Since *Alpha_mycolic_acid* is the only class that accepts mycolic acid with Cyclopropane, the lipid in Figure 1 is classified as a member of *Alpha_mycolic_acid*. (Figure 2).

Conclusion

Lipid research is increasingly integrated within systems level biology such as lipidomics[9] where lipid definition and classification are required before annotation of chemical functions can be applied. In this paper we have sought to address the ongoing challenge of classifying lipids through the adoption of W3C standard knowledge representation and the application of DL axioms. In other domains of metabolomics, e.g. glycomics, the adoption of ontologies such as, GlycO – a focused ontology representing complex carbohydrates, have enabled correlation of structural features of glycans to the biosynthesis and metabolism[10]. We initiated the process of defining lipids according to appropriate functional groups with the intent of using the ontology for classification of lipids. Ontology driven classification has been applied to proteins[11] and small molecules[6] through the coordination of protein domain or pharmacophore analysis, OWL-DL ontology, and DL reasoning. By adding precisely defined DL-axioms to the lipid ontology we can apply a similar approach for the automated classification of lipids. Our approach is extensible to accommodate novel lipids and we extended the use of DL-axioms to classify all lipid classes (except for polyketides). In support of mycobacterial disease research, we extended lipid nomenclature and classification of mycolic acids. We have made available systematic and formalized OWL-DL definitions of lipids for testing the appropriateness of existing nomenclature to lipid structures. This will serve as a reusable standard for lipid researchers and the lipid bioinformatics community. The Lipid Ontology is available online at NCBO's Bioportal and at: http://www.lipidprofiles.com/LipidOntology/LiPrO-02042009.owl

Acknowledgements

NUS Office of Life Science (R-183-000-607-712), ARF (R-183-000-160-112), BMRC A*STAR (R-183-000-134-305), Singapore NRF under CRP award No. 2007-04, NBIF, New Brunswick, Canada.

References

1. Fahy E, *et al*. A comprehensive classification system for lipids. *J. Lipid Res.* 2005, 46: 839–62.
2. Sud M, *et al*. LMSD: LIPID MAPS structure database. *Nucl. Acids Res.* 2007, 35: D527–32.
3. Baader FI, *et al*. Description logics as ontology languages for the semantic web. In *Festschrift Jörg Siekmann, LNAI* 2003.
4. Villanueva-Rosales N, Dumontier M. Describing chemical functional groups in OWL-DL for the classification of chemical compounds. OWLED 2007, co-loc. ESWC2007 Innsbruck, Austria.
5. Baker CJO, *et al*. Towards ontology-driven navigation of the lipid *bibliosphere. BMC Bioinformatics*. 2008, 9 (Suppl 1):S5.
6. Feldman HJ, *et al*. CO: A Chemical Ontology for Identification of Functional Groups and Semantic Comparison of Small Molecules. *FEBS Letters.* 2005, 579: 4685–4691.
7. Degtyarenko K, *et al*. ChEBI: A database and ontology for chemical entities of biological interest. *NAR.* 2008, 36: D344–D350.
8. Barry CE, *et al*. Mycolic acids: structure, biosynthesis and physiological functions. *Prog. Lip. Res.* 1998, 37: 143–179.
9. Wenk MR. The emerging field of Lipidomics. *Nat. Rev. Drug Discov.* 2005, 4: 594–610.
10. Thomas CJ, *et al*. Modular ontology design using canonical building blocks in biochemistry domain. FOIS pp. 115–127, 2006.
11. Wolstencroft K, *et al*. Protein classification using ontology classification. *Bioinf. 2006*, 22: e530–8.
12. http://www.ida.liu.se/~iislab/projects/SAMBO/

The RNA Ontology (RNAO): An Ontology for Integrating RNA Sequence and Structure Data

Colin Batchelor[1], Thomas Bittner[2], Karen Eilbeck[3], Chris Mungall[4], Jane Richardson[5], Rob Knight[6], Jesse Stombaugh[6], Craig Zirbel[7], Eric Westhof[8], Neocles Leontis[7]

[1]Royal Society of Chemistry, Cambridge, UK; [2] Department of Philosophy, University at Buffalo, Buffalo, NY, USA; [3]Department of Human Genetics, Eccles Institute of Human Genetics, University of Utah, Salt Lake City, UT, USA; [4]Life Sciences Division, Lawrence Berkeley National Lab, CA, USA; [5]Department of Biochemistry, Duke University Medical School, Durham, NC, USA; [6]Department of Chemistry and Biochemistry, University of Colorado at Boulder, Boulder, CO, USA; [7]Departments of Chemistry and of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH, USA; [8]Architecture et réactivité de l'ARN, Université Louis Pasteur de Strasbourg, Institut de Biologie Moléculaire et Cellulaire du CNRS, Strasbourg, France.

Abstract

Biomedical Ontologies are intended to integrate diverse biomedical data to enable intelligent data-mining and facilitate translation of basic research into useful clinical knowledge. We present the first version of RNAO, an ontology for integrating RNA 3D structural, biochemical and sequence data. While each 3D data file depicts the structure of a specific molecule, such data have broader significance as representatives of classes of homologous molecules, which, while differing in sequence, generally share core structural features of functional importance. Thus, 3D structure data gain value by being linked to homologous sequences in genomic data and databases of sequence alignments. Likewise genomic data can increase in value by annotation of shared structural features, especially when these can be linked to specific functions. The RNAO is being developed in line with the developing standards of the Open Biomedical Ontologies (OBO) Consortium.

Introduction

The aim of the RNA Ontology Consortium (ROC)[1] is "to create an integrated conceptual framework—an RNA ontology—with a common, dynamic, controlled and structured vocabulary to describe and characterize RNA sequences, secondary structures, three-dimensional structures and dynamics pertaining to RNA function." Other kinds of experiment that are useful to RNA biochemists and bioinformaticists include chemical probes and thermodynamic measurements. Previous work in this field includes the RiboWeb ontology,[2] which was part of a knowledge base for studying the bacterial ribosome, the Multiple Alignment Ontology for nucleic acid and protein sequences[3] and RNAML,[4] which is an actively-used XML schema for exchanging information about RNA secondary structures, tertiary structures, sequences and sequence alignments. The immediate context of the RNA Ontology is the Open Biomedical Ontologies (OBO) project,[5] which seeks to coordinate the development of biomedical ontologies. Small molecules are dealt with by ChEBI,[6] macromolecular sequences (DNA, RNA and protein) by the Sequence Ontology[7] and proteins by the Protein Ontology.[8] The RNAO is distinct from its neighbors but will share relationships and refer to terms from the other ontologies where necessary.

We set out the paper as follows: we briefly describe the chemical structure of the RNA molecule and then describe how to represent (1) base pairing and other pairwise interactions, (2) motifs and (3) backbone conformations based on the hierarchical nature of RNA structure. We will also describe the relationship to the Sequence Ontology. The RNAO is developed using Protégé as an OWL[*] ontology and is also available in OBO format. We illustrate what can be done within the limitations of OWL; however a full treatment of RNA structure requires first-order logic. RNAO is freely available.[†]

RNA

Ribonucleic acid (RNA) molecules consist of nucleotide (nt) units, which themselves consist of heterocyclic nucleobases covalently bonded to ribose rings which are connected covalently to the ribose rings of other nucleotides through phosphate groups. The combination of base and ribose is called a nucleoside. Each nucleoside has three interacting edges, the Watson-Crick edge, the Hoogsteen edge and the sugar edge as shown in Fig. 1. These edges are sets of hydrogen-bond donors and hydrogen-bond acceptors located on the same stretch of the boundary of the nucleoside. They are illustrated for adenosine in Fig. 1. The nucleotide units themselves are linked

[*] http://www.w3.org/TR/owl-features/
[†] http://code.google.com/p/rnao

one to the next in a directional manner, usually by connection of the 3' position of a nucleotide to the 5' position of the next nucleotide in the chain via the phosphate group (see Fig. 1).

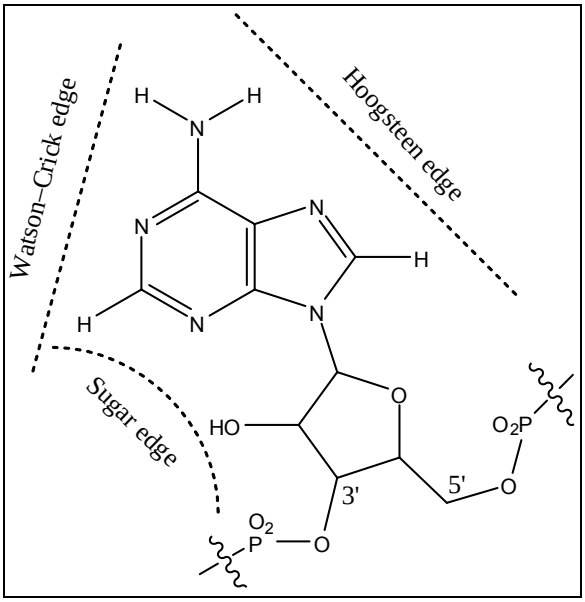


Figure 1. The Watson–Crick, Hoogsteen and Sugar edges on an adenosine nucleotide.

We follow Villanueva-Rosales and Dumontier,[9] who base their ontology on atoms and bonds, but we modify their approach by treating the nucleotides as the objects and the interactions between them as the relations. Thus we have two fundamental relations, the covalently_bonded_to relation, and the pairs_with relation.

The folding of the RNA chain brings together pairs of short sequence segments that are Watson-Crick complementary to form anti-parallel double helices consisting of stacked Watson-Crick basepairs. Helices are the simplest and most regular RNA 3D motifs. The set of Watson-Crick paired helices comprise the secondary structure of the RNA. Some RNA molecules can form more than one secondary structure and can be induced by appropriate perturbations to switch between them. The looping of the chain forms other motifs called hairpin loops, many of which are structured by specific sets of interactions, including base-pairing and base-stacking and often, base-phosphate interactions. Segments of sequence joining two helices can also form structured motifs called internal loops. Finally, multi-helix junction loops result when three or more helical segments are joined together. Junction loops provide branch points in RNA molecules. RNA 3D motifs recur in numerous RNA molecules encoded by genes from different families in very different organisms. Recurrent 3D motifs often play similar roles in different RNA molecules. For example, junction loops provide branch points, kink-turn internal loops provide flexible hinges and GNRA hairpin loops mediate tertiary interactions. Motifs combine to define characteristic RNA folds or domains.

Base Pairing

We start with the basepair classification proposed by Leontis and Westhof,[10] which places RNA basepairs in distinct, geometrically defined classes that are mutually exhaustive and disjoint. The pairwise interactions are hydrogen bonds between atoms in adjacent nucleosides, and as such we define the interactions in terms of *edges* (see Fig. 1). To a first approximation:

(1) *each edge of a nucleoside may interact only with a single edge of a different nucleoside*

Because OWL can only handle binary relations, we have to specialize the pairs_with relation for each combination of interacting edges. With six different combinations of edge interaction (WC-WC, H-H, S-S, WC-H, WC-S and H-S), and two relative orientations (*cis* and *trans*) for the interaction of the nucleosides, there result twelve basepairing classes in the Leontis-Westhof scheme and eighteen base pairing relations as shown in Table 1 (appendix). We can express statement (1) formally by declaring each of these relations to be *disjoint* from other relations, which means for example that if X pairs_with_CWH Y then there is no Z such that X pairs_with_CWW Z. The logical definition for a family 1 base pair is written:

family_1_base_pair = hasPart some (Nucleobase and pairs_with_CWW some Nucleobase)

in OWL Manchester syntax,[11] and this is sufficient for a reasoner to classify a base pair with the correct pairing relation into the correct LW family.

Motifs

By specializing the covalently_bonded_to relation and pairs_with relation it is possible to create rudimentary definitions of most motifs, and it is straightforward to generate RNAO-specific first-order logic representations of a given RNA structure from a plain text file. However, because all but the very simplest motifs contain cyclically-connected nucleotides, and OWL cannot handle cycles, it is impossible for this part of the ontology to be represented in OWL in such a way that reasoners can deal with it.

Further, it is possible that some motifs will be best described by formal definitions, whereas other more

complex motifs may be best described by statistical or machine learning approaches.

Backbone Conformers

The backbone in RNA molecules is a chain of covalently-bonded atoms which are parts either of the phosphate group (O5', P, O3') or of the ribose rings (C3'-C4'-C5'). We are interested in RNA backbone conformations for two reasons: (1) particular RNA motifs can also be described as a sequence of backbone conformers, and (2) they provide sites for catalysis or interaction with ions, proteins, small molecules, proteins, and other nucleic acids or segments of the same RNA.

We are using the ROC backbone committee's 2-character notation[12] for the conformations of suites, which are the stretches of backbone between two ribose rings. Each of their 54 suite conformers is a cluster of datapoints in the 7-dimensional space of the backbone dihedral angles. Suites and nucleotides provide alternative ways to partition the RNA molecule, but we are exploring whether the ontology can simply treat suite conformers as qualities of the covalent connection between nucleotides.

RNAO and the Sequence Ontology

The Sequence Ontology (SO) is a structured controlled vocabulary for the description of biological sequence. SO is used by model organism genome communities for the annotation of genomic sequence and will provide the basic terms to describe sequence features for RNAO. SO will be extended to provide terms to describe discontiguous regions. This will be necessary to describe many secondary and tertiary structural motifs. SO also includes a number of RNA motif terms that will be transitioned to RNAO.

Conclusions

We have presented a rudimentary version of RNAO which contains logical definitions that can be used by a reasoner to classify base pairings into the twelve categories of Leontis and Westhof and outlined how to incorporate 3D motifs and backbone configurations into the ontology. We have also shown what can be done in OWL for interoperability with other OBO ontologies and what needs to be represented in first-order logic.

Acknowledgements

The ROC is supported by a Research Coordination Network (RCN) grant from the National Science Foundation (grant no. 0443508). SO is supported by the NHGRI, via the Gene Ontology Consortium, HG004341.

References

1. Leontis NB, *et al.* The RNA Ontology Consortium: an open invitation to the RNA community. RNA, 2006;12;533
2. Altman R, *et al.* RiboWeb: An ontology-based system for collaborative molecular biology. IEEE Intelligent Systems 1999;14(5);68-76.
3. Thomson JD, *et al.* MAO: A Multiple Alignment Ontology for nucleic acid and protein sequences. Nucleic Acids Research 2005;33(13);4164-4171.
4. Waugh A, *et al.* RNAML: A standard syntax for exchanging RNA information. RNA, 2002;8;707.
5. Smith B, *et al.* The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology, 2007;25; 1251.
6. Degtyarenko K, *et al.* ChEBI: A database and ontology for chemical entities of biological interest. Nucleic Acids Res. 2008;36;D344.
7. Eilbeck K, *et al.* The Sequence Ontology: A tool for the unification of genome annotations. Genome Biol. 2005;6(5):R44.
8. Natale DA, *et al.* BMC Bioinformatics 2007;8(Suppl 9);S1.
9. Villanueva-Rosales N and Dumontier M. 2007. Describing chemical functional groups in OWL-DL for the classification of chemical compounds, in OWL: Experiences and Directions (OWLED 2007), co-located with European Semantic Web Conference (ESWC2007), Innsbruck, Austria.
10. Leontis NB and Westhof E. 2001. Geometric nomenclature and classification of RNA base pairs, RNA, 2001;7;499-512.
11. Horridge M, *et al.* 2006. The Manchester OWL Syntax, in OWL Experiences and Directions Workshop, 2006.
12. Richardson JS, *et al.* RNA backbone: Consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution), RNA, 2008;14;465-481.

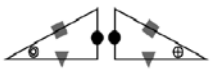
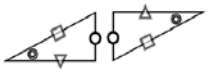
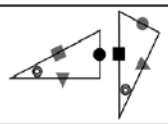
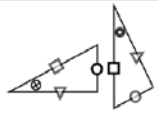
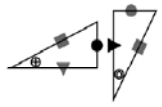
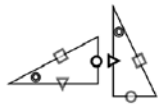
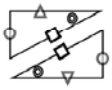
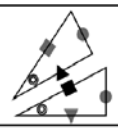
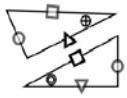
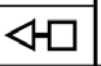
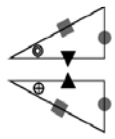
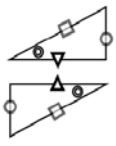
	Relations					
Classes	**Base pairing relations**		**Inverse of**	**Disjoint with other**	**Symbol**	**Triangle Abstraction**
family_1_base_pair	pairs_with_WW	pairs_with_CWW	symmetric	W-first pairings W-second pairings		
family_2_base_pair		pairs_with_TWW	symmetric	W-first pairings W-second pairings		
family_3_base_pair	pairs_with_WH	pairs_with_CWH	pairs_with_CHW	W-first pairings H-second pairings		
	pairs_with_HW	pairs_with_CHW	pairs_with_CWH	H-first pairings W-second pairings		
family_4_base_pair	pairs_with_WH	pairs_with_TWH	pairs_with_THW	W-first pairings H-second pairings		
	pairs_with_HW	pairs_with_THW	pairs_with_TWH	H-first pairings W-second pairings		
family_5_base_pair	pairs_with_WS	pairs_with_CWS	pairs_with_CSW	W-first pairings S-second pairings		
	pairs_with_SW	pairs_with_CSW	pairs_with_CWS	S-first pairings W-second pairings		
family_6_base_pair	pairs_with_WS	pairs_with_TWS	pairs_with_TSW	W-first pairings S-second pairings		
	pairs_with_SW	pairs_with_TSW	pairs_with_TWS	S-first pairings W-second pairings		
family_7_base_pair	pairs_with_HH	pairs_with_CHH	symmetric	H-first pairings H-second pairings		
family_8_base_pair		pairs_with_THH	symmetric	H-first pairings H-second pairings		
family_9_base_pair	pairs_with_HS	pairs_with_CHS	pairs_with_CSH	H-first pairings S-second pairings		
	pairs_with_SH	pairs_with_CSH	pairs_with_CHS	S-first pairings H-second pairings		
family_10_base_pair	pairs_with_HS	pairs_with_THS	pairs_with_TSH	H-first pairings S-second pairings		
	pairs_with_SH	pairs_with_TSH	pairs_with_THS	S-first pairings H-second pairings		
family_11_base_pair	pairs_with_SS	pairs_with_CSS	symmetric	S-first pairings S-second pairings (this will change in v2)		
family_12_base_pair		pairs_with_TSS	symmetric	S-first pairings S-second pairings (this will change in v2)		

Table 1: Base pairing relations in RNAO.

Evolution of the Sequence Ontology Terms and Relationships

Karen Eilbeck[1], Christopher J. Mungall[2]
[1]Department of Human Genetics, University of Utah, Salt Lake City, UT, USA
[2]Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Abstract

The Sequence Ontology is undergoing reform to meet the standards of the OBO Foundry. Here we report some of the incremental changes and improvements made to SO. We also propose new relationships to better define the mereological, spatial and temporal aspects of biological sequence.

Introduction

The Sequence Ontology[1] was begun in 2003 as a means to provide the terms and relations that obtain between terms, to describe biological sequence. The main purpose being the unification of the vocabulary used in genomic annotations, specifically genomic databases and flat file data exchange formats. Genomic data has been notoriously unspecified with a multitude of file formats expressing the same kind of data in different ways. Each gene prediction algorithm for example, exported the gene models in either a different format from other groups, or when they used the same format, the terms often had slightly different meanings. Data integration between groups was therefore not straightforward. Likewise, validation of annotations relied on the programmers understanding the nuances of each kind of annotation and hard-coding their programs to match. The Sequence Ontology provides a forum for the genomic annotation community to discuss and agree on terminology to describe their biological sequence.

The SO was initially divided into aspects to describe the features of biological sequence and the attributes of these features. A sequence feature is a region or a boundary of sequence that can be located in coordinates on biological sequence. SO uses a subsumption hierarchy to describe the kinds of features and meronomy to describe their containment. Features were related by their genomic position. For example polypeptides and transcripts are described by genomic context. This excluded their post-genomic topology.

The SO has a large user community of established model organism databases and newer 'emerging model organism' systems who rely on the GMOD[2] suite of tools to annotate and disseminate their genetic information. GMOD schemas and exchange formats rely on the SO to type their features such as the Chado database[3], with its related XML formats and the tab delimited flat file exchange format GFF3[4]. Several GMOD tools use GFF3, for example GBrowse[5]. SO is also used by genome integration projects such as Flymine[6], modENCODE[7] and the BRC pathogen data repository[8]. There are other uses for SO such as natural language processing initiatives that use the SO terminology[9,10].

The SO is one of the original members of the OBO Foundry[11]. The OBO ontology developers agreed to a set of shared principles for formal ontology design, with the aim of achieving orthogonal, interoperable ontologies. There are 10 principles for OBO Foundry membership which include a common syntax, a data-versioning system, collaborative development, and adherence to the same set of defined relationships[12]. The OBO ontology developers attempt to accurately represent reality. Membership in the OBO Foundry represents a commitment to adhere to the ontology design principles and agree to reform where necessary. The OBO Foundry spans the biomedical domain in steps of granularity from the molecule to the population. It also encompasses the relations to time. Continuants endure through time, where as occurrents, which include processes, unfold through time in stages. The sequence features of SO are instantiated as molecules or parts of molecules.

The SO has orthogonal neighbor ontologies within the OBO Foundry, also describing molecular continuants. Chemical entities of Biological Interest (ChEBI) is a dictionary of small chemical compounds[13]. It does not describe molecules encoded by the genome such as transcripts and peptides. The RNA Ontology[14] describes the secondary and tertiary motifs of RNA as well as providing relationships between bases for base pairing and stacking. The Protein Ontology (PRO) defines the forms of proteins and the evolutionary relationships between protein families[15]. It is natural for these ontologies to interact and create inter-ontology terms in the form of cross products. To do so, the ontologies must all adhere to the same principles.

Coordinated Reform of SO to OBO Standards

The SO, like other pre-exisiting ontologies has begun to undergo reform to meet the OBO Foundry standards.

Textual Definitions

New terms are now defined using OBO Foundry guidelines for definitions. The existing terms in SO were initially either defined by a member of the developer community, or via a cross reference to a reputable source. This has lead to inconsistency between the definitions, and sometimes inconsistency between the definition and placement of the term. The OBO Foundry recommends that terms be defined with respect to the *is_a* parent, and the attributes that differentiate the term from its parent and sibling terms, called the differentiae. This practice forces a self check on the whether the position of the term in the ontology agrees with the defined meaning of the term. New definitions in SO must adhere to the "A is_a B that C's" principle. For example, the new term, **vector_replicon**, a subtype of **replicon**, has the following definition: *A replicon that has been modified to act as a vector for foreign sequence*. Existing terms are undergoing a refinement process.

Logical Definitions

In addition to providing text definitions, the SO includes over 100 'cross-product' definitions in genus/differentiae form[16]. A reasoner can then be used to place the terms in the correct place in the ontology. This is especially useful as it untangles the graph for editing purposes. The SO is released in two forms, either with the logical definitions, or fully classified for use without a reasoner.

Parthood Relations

The SO must adhere to the principles of OBO Relations Ontology (RO). The RO provides a set of defined formal type level and instance level relations. The list of relations may be extended by individual ontologies as required. The class level relations follow the "ALL_SOME" rule[17]. This rule is necessary to improve the ability to reason over data that uses the ontology. In practice, making these changes to SO has required the addition of the '*has_part*' relation to the ontology. Prior to this change the SO stated that:

TATA_box *part_of* **RNApol_II_promoter** and
TATA_box *part_of* **RNApol_III_promoter**.

This was incorrect as all **TATA_boxes** are not part of both kinds of promoter. The ontology now states that:

RNApol_III_promoter *has_part* **TATA_box**.
The *integral_part_of* relation and its inverse have been added to clarify the occasions when the part and the whole must both exist.

Temporal Relations and Spatial Interval Relations

There are several kinds of relation that are needed to describe the complex nature of biological sequence. Mereological relations are needed to describe containment. Spatial relations are needed to relate the positional information about features. Each transformation of sequence requires a temporal relation. Finally as SO is part of a larger suite of ontologies, it will need relations with which to make cross products and refer to other ontologies. We propose to extend SO with the relations outlined in Table 1.

Biological sequence is predominantly instantiated in three kinds of polymeric molecule: DNA, RNA and polypeptide, although man-made polymers such as PNA do exist. The SO will represent the transformation of sequence from one kind of molecule to another using the temporal relations shown in Table 1. A **gene**, manifest in DNA *transcribed_into* the **primary_transcript**, which is expressed as RNA. A **polypeptide** sequence is a *translation_of* the **CDS** sequence. **Transcript** molecules also undergo processing such as splicing and editing which remove or add additional sequences. The relations *processed_from* and *processed_into* relate the primary transcript to its mature processed form.

It is important to understand how the proposed changes will affect the annotation community who already use the terms and relations of SO in their pipelines and processes. This will effect how the changes are released. The terminology used to type the features already in use will not change. The GFF3 format will be unaffected as it lists the feature types and the parent term of a given relation. It does not name the relation – this is maintained in the ontology.

Developers will need to be given notice of new relationships and structures however, as this may have adverse effects of pipelines and programs.

The proposed changes to the SO relationships and structure can be found on the SO website at the following address:
http://www.sequenceontology.org/resources/proposed_relationships.html

	Name	Definition	example
Mereological	part_of	X part_of Y if X is a subregion of Y.	**amino_acid** *part_of* **polypeptide**
	has_part	Inverse of part_of	**operon** *has_part* **gene**
	integral_part_of	X integral_part_of Y if and only if: X part_of Y and Y has_part X	**exon** integral_part_of **transcript**
	has_integral part	X has_integral_part Y if and only if: X has_part Y and Y part_of X	**mRNA** *has_integral_part* **CDS**
Temporal	transcribed_from	X is transcribed_from Y if X is synthesized from template Y.	**primary_transcript** *transcribed_from* **gene**
	transcribed_to	Inverse of transcribed_from	**gene** *transcribed_to* **primary_transcript**
	translation_of	X is translation of Y if X is translated by ribosome to create Y.	**Polypeptide** *translation_of* **CDS**
	translates_to	Inverse of translation _of	**codon** *translates_to* **amino_acid**
	processed_from	Inverse of processed_into	**miRNA** *processed_from* **miRNA_primary_transcript**
	processed_into	X is processed_into Y if a region X is modified to create Y.	**miRNA_primary_transcript** *processed into* **miRNA**
Spatial Interval	contained_by	X contained_by Y iff X starts after start of Y and X ends before end of Y	**intein** *contained_by* **immature_peptide_region**
	contains	Inverse of contained_by	**Pre-miRNA** *contains* **miRNA_loop**
	overlaps	X overlaps Y iff there exists some Z such that Z contained_by X and Z contained_by Y	**coding_exon** *overlaps* **CDS**
	maximally_overlaps	A maximally_overlaps X and Y iff all parts of A (including A itself) overlap both X and Y	**non_coding_region_of_exon** *maximally overlaps* the intersection of **exon** and **UTR**
	connects_on	X connects_on Y,Z,R iff whenever X is on a R, X is adjacent_to a Y and adjacent_to a Z	**splice_junction** *connects_on* **exon**, **exon** **mature_transcript**
	disconnected_from	X is disconnected_from Y iff it is not the case that X overlaps Y	**intron** *disconnected_from* **exon {**on **transcript}**
	adjacent_to	X adjacent to Y if and only if: X and Y share a boundary but do not overlap	**UTR** *adjacent_to* **CDS**
	started_by	X is started by Y, if Y is part_of X and X and Y share a 5 prime boundary.	**CDS** *started_by* **start_codon**
	finished_by	X is finished by Y if Y is part_of X and X and Y share a 3 prime boundary	**CDS** *finished_by* **stop_codon**
	starts	X starts Y is X is part of Y and X and Y share a 5 prime boundary.	**start_codon** *starts* **CDS**
	finishes	X finishes Y if X is part_of Y and X and Y share a 3' boundary.	**stop_codon** *finishes* **CDS**
	is_consecutive_sequence _of	R is_consecutive_sequence_of U if and only if every instance of R is equivalent to a collection of instances of U u1,u2,...,un such that no pair ux uy is overlapping, and for all ux, ux is adjacent_to ux-1 and ux+1, with the exception of the initial and terminal u1 and un (which may be identical).	**region** *is_consecutive_sequence_of* **base** **processed_transcript** *is_consective_sequence_of* **exon**
Cross ontology	site_of	A is a site of B if A is the sequence_feature of a molecule where a GO:biological process B occurs.	**CDS** *site_of* **RNA polymerase activity**
	output_of	A is an output_of B if A is a sequence_feature of a molecule that is produced by GO:biological process B.	**primary_transcript** *output_of* **transcription**
	regulates_expression_of	A regulates expression of B if A is a regulatory region that controls the expression of B, where B is a gene.	**regulatory_region** *regulates_expression_of* **gene**

Table 1. New relations proposed for SO. Definitions are for instance level relations, examples are for class-level relations, which follow from the instance-level definition in the standard all-some pattern.

Conclusions

The updates to the SO, based on OBO Foundry recommendations have strengthened the ontology as a tool for reasoning. The treatment of definitions enforces a tight regulation on the position of a new term in the ontology and synchronizes the textual definition within the subsumption hierarchy. The process of updating all of the definitions is ongoing. Stricter adherence to the OBO Relations Ontology is making SO interoperable with the other OBO ontologies. The SO uses a reasoner to maintain the is_a parents of cross product terms. This aids ontology maintenance and can be used as a model for other OBO ontologies.

The application of sequence features that span the range of the molecular biology central dogma, rather than simply the position of the genomic region that encodes the molecule, is a subtle but important step

forward. It allows the topological relations at each stage from genome to transcript or peptide to be catalogued. It roots the SO within OBO making cross products between the sibling ontologies possible.

The addition of a suite of mereological, topological and temporal relations will dramatically enhance the ability to use the SO as a tool for computational reasoning. Each of the new defined relationships adds another avenue for analysis. This is especially important for the validation of sequence annotations using SO.

Acknowledgments

This work is supported by the NHGRI, via the Gene Ontology Consortium, HG004341.

References

1. Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R and Ashburner M. The Sequence Ontology: A tool for the unification of genome annotations. Genome biology 2005, 6(5):R44.
2. GMOD [www.gmod.org]
3. Mungall CJ and Emmert DB. A Chado case study: An ontology-based modular schema for representing genome-associated biological information. Bioinformatics (Oxford, England) 2007, 23(13):i337-346.
4. GFF3 [www.sequenceontology.org/gff3.shtml]
5. Donlin MJ. Using the Generic Genome Browser (GBrowse). Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al 2007, Chapter 9:Unit 9 9.
6. Lyne R, Smith R, Rutherford K, Wakeling M, Varley A, Guillier F, Janssens H, Ji W, McLaren P, North P, *et al*. FlyMine: An integrated database for Drosophila and Anopheles genomics. Genome biology 2007, 8(7):R129.
7. modENCODE [www.modencode.org]
8. BRC [http://www.brc-central.org/cgi-bin/brc-central/brc_central.cgi]
9. Vlachos A, Gasperin C, Lewin I and Briscoe T. Bootstrapping the Recognition and Anaphoric Linking of Named Entities in Drosophila Articles. In: Pacific Symposium on Biocomputing. vol. 11; 2006: 100-111.
10. RSC [http://www.rsc.org/Publishing/Journals/ProjectProspect/index.asp]
11. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, *et al*. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology 2007, 25(11):1251-1255.
12. Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL and Rosse C. Relations in biomedical ontologies. Genome biology 2005, 6(5):R46.
13. Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcantara R, Darsow M, Guedj M and Ashburner M. ChEBI: A database and ontology for chemical entities of biological interest. Nucleic acids research 2008, 36(Database issue):D344-350.
14. Leontis NB, Altman RB, Berman HM, Brenner SE, Brown JW, Engelke DR, Harvey SC, Holbrook SR, Jossinet F, Lewis SE, *et al*. The RNA Ontology Consortium: An open invitation to the RNA community. RNA (New York, NY 2006, 12(4):533-541.
15. Natale DA, Arighi CN, Barker WC, Blake J, Chang TC, Hu Z, Liu H, Smith B and Wu CH. Framework for a protein ontology. BMC bioinformatics 2007, 8 Suppl 9:S1.
16. [http://wiki.geneontology.org/index.php/SO:Composite_Terms]
17. Smith B, Kumar A, Ceusters W and Rosse C. On carcinomas and other pathological entities. Comparative and functional genomics 2005, 6(7-8):379-387.
18. Allen JF. Maintaining Knowledge about Temporal Intervals. Communications of the ACM 1983, 26(11):832-843.
19. OBO format 1.3 [http://www.geneontology.org/GO.format.obo-1_3.shtml

ChemAxiom – An Ontological Framework for Chemistry in Science

Nico Adams, Edward O. Cannon, Peter Murray-Rust
Unilever Centre for Molecular Science Informatics, University Chemical Laboratory, University of Cambridge, Cambridge, UK

Abstract

We present ChemAxiom as the first ontological framework for chemistry in science. ChemAxiom enables discourse about chemical objects in a computable language and is useful for the management of chemical concepts and data, the retrospective typing of resources, the identification of ambiguity and supports chemical text mining.

Ontology in Chemistry – Current State of the Art

Chemistry is a central scientific discipline and at the heart of a number of other important sciences such as biomedical research. While the latter has derived tremendous benefit from the development of controlled vocabularies, taxomomies and ontologies for the annotation of biological knowledge and text, chemistry has been slow to adopt these technologies and remains, on the whole, an ontological wasteland, although Batchelor and others have made excellent cases for the use of (formal) ontological methods in chemistry.[1, 2]

There have been several attempts to apply ontological techniques to the field of chemistry in the past. Currently, the most widely used ontology in chemistry is the "Chemical Entities of Biological Interest" (ChEBI) ontology.[3] ChEBI contains ontological associations, which specify chemical relationships as well as the biological roles and applications of a molecule. A recent study by Batchelor showed, that ChEBI contains a substantial amount of implicit and disguised semantics, which significantly complicates its use in modern semantic information systems.[1] Other notable ontologies in the chemistry domain are the Chemical Ontology,[4] REX[5] and FIX,[6] which model physicochemical processes and methods respectively, as well as ChemTop, which is a subset of the BioTop ontology.[7] Though valuable for annotation, none of these efforts can be considered to constitute an ontological framework for chemistry.

Case for Formal Ontological Methods in Chemistry

Chemical information systems and resource discovery in chemistry are often predicated on the use of chemical structure (connection table) as an identifier and as annotation for chemical data. This springs from the "central dogma" of chemistry, namely, that molecular structure is correlated to the physico-chemical and biological properties of chemical entities. While this practice has served a subsection of the chemical community relatively well, there are major problems: first and foremost, the use of a connection table as a chemical identifier leads to a fundamental ontological confusion between the universal "molecule" and a "real world" bulk substance. Yet, in many information systems, a physicochemical property of a *substance* is associated with the structure of a molecule. It does not make sense to speak of a melting point in terms of a molecule. Many physicochemical quantities are properties of the mereological sums of the molecules, which make up the substance and not properties of the molecules themselves. In practice, this almost always leads to "lossy" encoding of information and information compartmentalisation. Formal ontology can help by providing a clear distinction between the abstract notion of a molecule and a bulk substance as might be in use in the laboratory. A similar argument can be made for many identifiers: in many information systems, it is not clear whether the identifier applies to a molecule or the substance.

Many chemical entities have dynamic structures (*e.g.* rapidly interconverting isomers - glucose dissolved in water) and cannot be described by one structural representation alone, *i.e.* there exists a parthood relationship between a given chemical entity and the corresponding several structures that can be written. Furthermore, there is a dependence on the notion of time: the fluxional structure of a chemical entity is a function of time. Ontology can assist in defining and specifying both these parthood relationships as well as the time dependence.

Materials and formulations, too, can be composites of several molecular entities or other chemical entities, which, in turn can be composites. Moreover, the "history" (*e.g*. synthesis conditions, post-processing etc.) of a material often has a significant impact on its physical properties, which are not captured by simple structural annotation.

By adopting formal ontological methods, we can clarify ambiguous meanings: if, for example, text mining has identified the term "acid" in a piece of text, then it is not clear whether this refers to a molecule acting as an acid or a chemical substance (a bottle of acid). If, however, the term "pH" has also

been identified in this context, a formal ontology could indicate that the concept "acid" refers to a substance rather than a molecule. When applied in this way, an ontology can be used to "retrospectively type" chemical objects – in this example into chemical substance or molecule.

The ChemAxiom Set of Ontologies

To address some of the points discussed above and to fill the ontological void that currently exists in the chemical domain, we have developed ChemAxiom – a set of separately maintainable, but interoperable and integrated ontologies (Figure 1).

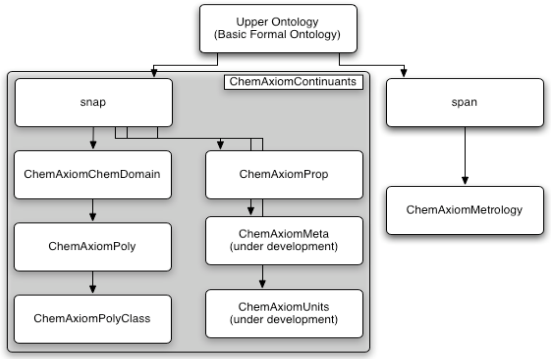


Figure 1. The ChemAxiom set of ontologies.

Each ontology describes a particular aspect of the chemical domain and collectively the ontologies form a framework for the description of chemistry. In designing ChemAxiom, we have borrowed from many bioscience ontologies such as the OBO family, ChEBI and MeSH and have derived some advantage from the fact that chemical concepts have clearer boundaries than biological ones. Consequently, ChemAxiom has been designed to (a) contain no implicit semantics, (b) be useful for the management of both chemical concepts and chemical data, (c) allow retrospective typing of chemical objects and the identification of ambiguity, (d) allow for undecideability either because of lack of knowledge or lack of axiomatisation and (e) allow for community extensibility and interoperability. Currently, the main use case for the ChemAxiom ontologies is the description of chemical data contained in documents of different types as well as machine output and the ability to support machine-generated RDF. We will present a formal evaluation of the ontologies w.r.t. this use case in further work. ChemAxiom complements other ontologies in the chemical field, which focus on, for example, compound taxonomy and biological function (ChEBI) or chemical structure (CO).[4] ChemAxiom has been prepared in the OWL language and is currently under active development, funded in part by both Unilever plc as well as the Dutch Polymer Institute. We are currently exploring the possibility of forming a broad platform around the ontologies with a number of partners and explicitly invite and value community participation in the development process. All ontologies are available at http://www.bitbucket.org/na303.

There are currently several ontology modules, which are integrated via the Basic Formal Ontology[8] as an upper ontology. ***ChemAxiomChemDomain*** is a small ontology which clarifies some fundamental concepts in chemistry, such as the parthood relationships between molecule and substance as well as generic roles which molecules and substances can assume. ***ChemAxiomProp*** currently contains a list of over 150 chemical and materials properties, together with definitions of symbols (where appropriate or available) and axioms for typing (see below). ***ChemAxiomMetrology*** is an ontology of over 200 measurement techniques and also contains a framework for instruments (though currently required metadata such as measurement conditions or specification of minimum information requirements are not included – this will be added at a later stage). It follows the same modeling pattern as *ChemAxiomProp* and thus also allows for typing of objects. ***ChemAxiomPoly*** and ***ChemAxiomPolyClass*** contain terms, which are in common use across polymer chemistry and materials science as well as a taxonomy of polymers in terms of generic chemical structure. ***ChemAxiomMeta*** will allow the specification of the provenance of chemical data and information. ***ChemAxiomContinuants***, finally, represents the integration of all of these sub-ontologies into an ontological framework for chemical continuants. Classes in all ontologies have natural language definitions (which have been omitted in the examples shown in this paper). Further ontologies will include ontologies of chemical reactions and experiments as well as specifying minimum information requirements for properties and measurement methods. We now illustrate some of the capabilities of the framework using a number of select examples.

Clarifying Parthood Relationships and Roles

Key concepts in the ChemAxiomChemDomain ontology are *ChemicalIdentifier*, *ChemicalElement*, *MolecularEntity* and *ChemicalSpecies*. We employ the IUPAC definitions of *MolecularEntity* and *ChemicalSpecies* and understand the former to be a "constitutionally or isotopically distinct atom, molecule, ion, […] etc., identifiable as a separately distinguishable entity", whereas a *ChemicalSpecies* is understood to be "an ensemble of chemically

identical molecular entities". Following Batchelor's suggestion, we map *ChemicalElement*, *MolecularEntity* and *ChemicalSpecies* into the BFO as subclasses of `snap:Object`.[1]

```
ChemDomain:ChemicalSpecies
      a       owl:Class ;
      rdfs:subClassOf snap:Object ;
      rdfs:subClassOf
               [ a     owl:Restriction ;
                 owl:onProperty
ChemAxiomProp:hasProperty ;
                 owl:someValuesFrom
ChemAxiomProp:Property
               ] ;
      rdfs:subClassOf
               [ a     owl:Restriction ;
                 owl:onProperty
ChemDomain:presentInAmount ;
                 owl:someValuesFrom xsd:string
               ] ;
      rdfs:subClassOf
               [ a     owl:Class ;
                 owl:unionOf ([ a
owl:Restriction ;
owl:onProperty ChemDomain:hasPart ;
owl:someValuesFrom ChemDomain:MolecularEntity
                            ] [ a
owl:Restriction ;
owl:onProperty ChemDomain:hasPart ;
                              owl:someValuesFrom
ChemDomain:ChemicalSpecies
                            ])
               ] ;
      owl:disjointWith ChemDomain:MolecularEntity ,
ChemDomain:ChemicalIdentifier ,
ChemDomain:ChemicalElement .
```

ChemicalSpecies, in turn, is composed of (hasPart) *MolecularEntity(s)* or other *ChemicalSpecies*. This crucial distinction now allows "real world" bulk substances (*e.g*. polymers, formulations, an amount of benzene in a bottle) to be modeled and kept ontologically distinct from the notion of the universal "molecule". Concepts such as *Solvent*, *Catalyst* or *Acid* are subclasses of either *ChemicalSpecies* or *MolecularEntity* as appropriate and are modeled in terms of roles: a *Solvent* is a *ChemicalSpecies* which has a role of *SolventRole*. While ChemAxiom makes parthood relationships specific, it is not easy to see how this can be reconciled with the current *de facto* use of many chemical identifiers, which are interchangeably applied to both molecules and substances (e.g. CAS numbers). If there is a unique molecular indentifier, such as InChI may be, then the identifier for the substance (URI) may be viewed as an aggregation of all the identifiers of the discrete molecular entities which are part of the substance. For materials, such as polymers, the situation is even more complex as it is difficult to discern a single uniqueness criterion: uniqueness in materials is often dependent on a material's history and context and it may be the case that several URIs may be required for the same material. This is an important question and subject to ongoing research.

Typing of Chemical Objects and Resources

ChemAxiomProp contains the central class *Property* (subclass of *snap:SpecificallyDependentContinuant*). *Property* has two types of subclass, *NamedProperty*, which is a primitive class and contains a list of concrete properties, which, too, are primitive. The other subclasses are mostly defined classes and represent categorizations in the domain. One *NamedProperty*, for example, is the *MeltingPoint*, which carries the following axiomatisation:

```
:MeltingPoint
      a       owl:Class ;
rdfs:subClassOf :NamedProperty ;
      rdfs:subClassOf
               [ a     owl:Restriction ;
                 owl:onProperty :hasType ;
                 owl:someValuesFrom
:ThermophysicalProperty
               ] ;
      rdfs:subClassOf
               [ a     owl:Restriction ;
                 owl:hasValue "m.p."^^xsd:string ;
                 owl:onProperty :hasSymbol
               ] .
```

In addition to being a direct *rdfs:subClassOf :NamedProperty*, *MeltingPoint* is a also a subclass of the anonymous class "hasType some ThermophysicalProperty" (l. 4-7). The defined class "*ThermophysicalProperty*", in turn, is modeled as the intersection of the two classes "*Property*" and "everything that is of type *ThermophysicalProperty*" (l. 5-12 below):

```
:ThermophysicalProperty
      a       owl:Class ;
      rdfs:label "Thermophysical properties"@en ;
      rdfs:subClassOf :Property ;
      owl:equivalentClass
               [ a     owl:Class ;
                 owl:intersectionOf (:Property [ a
owl:Restriction ;
                              owl:onProperty :hasType
;
                              owl:someValuesFrom
:ThermophysicalProperties
                            ])
               ] .
```

Therefore, a reasoner will be able to infer that a *MeltingPoint* is also a subclass of *ThermophysicalProperty*. This is an example of both ontology normalization and retrospective typing; while all classes have asserted single inheritance, multiple inheritance can be inferred and maintained *via* a reasoner (ontology normalisation). Reasoning of this type can easily be accomplished using reasoners such as Pellet. Furthermore, we do not assert deep hierarchies – rather we allow a user to construct their own taxonomies using a combination of axioms and defined classes. If, for example, text-mining were to discover the term "melting point", it could retrospectively be typed and therefore annotated as a *ThermophysicalProperty* or a *NamedProperty*.

Typing could be part of a larger system, in which a new "perspective" (*i.e.* a view onto a contextual reality, which need not be universally shared and may vary substantially and even conflict with other defined perspectives) can be constructed. This definition can

be implemented *a posteriori* without needing to re-code the data. Typing of this sort is informing our object-oriented code generation in physical science applications. ChemAxiomProp does not yet contain notions of dimensionality, nor a subdivision of properties into qualities and dispositions. This is the subject of future development work.

Management of Chemical Data – Data in RDF

ChemAxiomContinuants is the result of the integration of all the sub ontologies discussed so far, and facilitates the modeling of chemical objects and data in RDF. We show how this can be done by creating an instance of the *NamedChemicalSpecies* benzene in ChemAxiomContinuants:

```
:benzene
      a        ChemDomain:NamedChemicalSpecies ;
      ChemAxiomProp:hasProperty
               :Density_1 , :BoilingPoint_1 ;
      ChemDomain:hasPart :benzeneMolecule .
:BoilingPoint_1
      a        ChemAxiomProp:BoilingPoint ;
      ChemAxiomProp:hasValue
               "80.1"^^xsd:string ;
      :measuredBy Metrology:Ebulliometry .
:Density_1
      a       ChemAxiomProp:Density ;
      ChemAxiomProp:hasValue
              "0.8786"^^xsd:string .
:benzeneMolecule
      a       ChemDomain:MolecularEntity ;
      ChemDomain:hasIdentifier
              :MolecularFormula_1 , :CASNumber_1 .
:CASNumber_1
      a       ChemDomain:CASNumber ;
      ChemDomain:hasValue "71-43-2"^^xsd:string .
:MolecularFormula_1
      a        ChemDomain:MolecularFormula ;
  ChemDomain:hasValue "C6H6"^^xsd:string .
```

In future work we will use the ontologies to describe data and chemical entities extracted from papers, theses and other sources of chemical information using our OSCAR3 entity extraction system. When coupled with the ability of retrospective typing of the extracted information, this opens the door to document classification and faceted search.[9]

Conclusions

The adoption of ontological methods in the chemistry domain is lagging far behind that of other disciplines. However, the integration of biomedical and chemical data is important for the future progress of science. We have developed a set of ontologies that enables the description and typing of chemical objects and data in a semantically rich way. This work should go some way towards facilitating the integration of data from other scientific disciplines with chemical data.

References

1. Batchelor C. (2008) An Upper Level Ontology for Chemistry. 5th International Conference on Formal Ontology in Information Systems: Saarbruecken, Germany
2. Frey JG, Hughes GV, Mills HR, *et al*. (2003) Less is more: Lightweight ontologies and user interfaces for smart labs. UK e-Science All Hands Meeting: 500-507 Nottingham, UK
3. de Matos P, Ennis M, Zbinden M, *et al*. (2006) ChEBI – Chemical entities of biological interest. http://www3.oup.co.uk/nar/database/summary/646, Accessed December 12, 2008
4. Feldman HJ, Dumontier M, Lng S, *et al*. (2005) CO: A chemical ontology for identification of functional groups and semantic comparison of small molecules. FEBS Letters 579:4685-4691
5. Degtyarenko K. (2007) The Rex Ontology. http://obofoundry.org/cgi-bin/detail.cgi?id=rex, Accessed December 30, 2008
6. Degtyarenko K. (2007) The FIX ontology. http://obofoundry.org/cgi-bin/detail.cgi?id=fix, Accessed December 30, 2008
7. ChemTop ChemTop. http://purl.org/chemtop/dev, Accessed February 28, 2009
8. Grenon P. (2003) Spatio-temporality in Basic Formal Ontology. http://www.ifomis.org/Research/IFOMISReports/IFOMIS%20Report%2005_2003.pdf, Accessed Feb. 19, 2009
9. Corbett P and Murray-Rust P. (2006) High-throughput identification of chemistry in life science texts. Computational Life Sciences II, Lecture Notes in Computer Science 4216:107-118

A Formal Ontology of Sequences

Robert Hoehndorf[1,2], Janet Kelso[2], Heinrich Herre[1]
[1]IMISE, University of Leipzig, Germany
[2]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

Abstract

The Sequence Ontology is an OBO Foundry ontology that provides categories of sequences and sequence features that are applied to the annotation of genomes. To facilitate interoperability with other domain ontologies and to provide a foundation for automated inference, we provide here an axiom system for the Sequence and Junction categories in first- and second-order predicate logics.

Introduction

Biological sequences play a major role in genetics and bioinformatics research. They are important in the description of DNA, RNA and proteins. To describe sequences and their features semantically, the Sequence Ontology (SO)[2] was developed.

The SO distinguishes between sequence features, qualities of sequences, operations on sequences and sequence variants. A sequence feature is an extended or non-extended biological sequence. Extended sequence features are regions such as genes, intergenic regions or sequences of polypeptides. Non-extended sequences are called junctions – the boundaries between two extended sequences. Operations on sequences include insertions and deletions. Qualities of sequences include whether or not a sequence encodes a protein, whether a sequence acts enzymatically when transcribed, or whether the sequence is conserved. Although some formal definitions are available for the SO categories, most categories are defined using English.

Formal ontologies are intended to specify a conceptualization of a domain[5], and therefore provide the foundation for data and information integration and exchange. Definitions alone are insufficient to achieve this goal. Axioms are required to provide meaning for primitive, undefined categories. To formalize the basic categories used in the SO, several ontological questions about sequences must be answered, among them: What kind of entity is a biological sequence and how does it relate to categories in a top-level ontology? What are the properties of biological sequences? What relations are applicable to sequences? How do sequences relate to other kinds of entities, in particular to molecules, organisms or processes (of selection, mutation)?

Here we provide an axiom system for the SO's top-level categories. We use first- and second-order logics for this purpose. The axiom system is intended to serve as a foundation for the SO, and as a means to achieve interoperability between the SO and other domain ontologies through the provision of an explicit formalization of the basic categories and relations used in the context of sequences. For the construction of the axiom system, we employed the axiomatic method.[8]

Method

We consider a formal ontology to be a specification of a conceptualization, i.e., a particular view on the world[5]. A formal ontology uses a vocabulary whose terms denote concepts and relations which refer to things in reality.

One method that is used to specify the meaning of a term is an explicit definition. An explicit definition for a relation or category P provides a sentence ϕ in which P does not occur, such that every occurrence of P can be replaced with ϕ.

When explaining the meanings of a set of terms through explicit definitions, other terms must be used to define the terms in the set, and in turn the meaning of these terms must be specified (without creating a circular definition). Therefore, specifying the meanings of terms solely through explicit definitions will either lead to an infinite regress or leave several terms unspecified. In the latter case, the meaning of all terms for which a definition is provided depends on the meaning of the terms without definition, therefore leaving the meaning of all terms in the ontology unspecified.

We call the terms that are not explicitly defined *primitive terms*. The meaning of all terms in the ontology depends on the meaning of these primitive terms: because non-primitive terms are introduced through explicit definitions, every sentence involving a non-primitive term can be replaced with a sentence containing only primitive terms.

The problem remains how the meaning of the primitive terms can be described formally. We may construct complex sentences containing only primitve terms. These sentences can be understood as descriptions of formal interrelations between the

primitive terms. Some of these sentences are chosen as axioms: they are accepted as being true within the domain under consideration. Such axioms provide restrictions on the interpretation of the primitive terms, and therefore on the terms defined using these primitive terms. For a formal theory, and therefore for a formal ontology, the axioms are the central component, because only they can give meaning to terms used in the theory.

Results

The theory of biological symbols and sequences that we propose here is intended to be compatible with the Sequence Ontology (SO).[2] The SO uses two basic categories in the characterization of sequences, *Sequence* and *Junction*. Both can have attributes, i.e., properties. For example, a sequence may be a gene or a base, a junction an insertion site, and a sequence attribute enzymatic.

Sequences are linear entities and can come in two facets. Sequences can either have a start and an end point (such as an mRNA sequence), or form circles (such as the sequence of mitochondrial DNA). There are sequence atoms, which we call *Primitive biological symbols*. Primitive biological symbols have no proper sequence parts.

We introduce an important distinction that is currently neglected in the SO. The SO contains as its only basic category a sequence region, and employs an extensional mereological system on it. However, we will show that it is important to distinguish between a *sequence* and the *tokens* of a sequence. To illustrate the difference between a sequence and its token, consider all constituents (parts) of the sequences $ACAC$ and $CAAC$. The first sequence has as parts the sequences $ACAC$, ACA, CAC, AC, CA, A and C. The sequence $CAAC$ has as parts sequences $CAAC$, CAA, AAC, CA, AA, AC, A and C. It is remarkable that, although both sequences apparently have the same *length*, use the same primitive symbols (only A and C), and every primitive symbol occurs exactly twice in each sequence, $ACAC$ has seven sequences as part, while $CAAC$ has eight. This is due to the fact that there is *only one* sequence AC, which occurs in $ACAC$ *twice*. On the other hand, each *token* of $ACAC$ and of $CAAC$ will have ten parts.

The theory we propose here assumes that *Sequence*, *Molecule*, *Junction* and *Abstract sequence* are primitive categories. In particular, they are not defined, but characterized axiomatically. *Sequence* and *Junction* refer to representations of sequences such as those found in biological databases. Sequences have tokens which belong to the *Molecule* category. Molecules are material entities which are located in space and time. Instances of *Sequence* represent abstract, information bearing entities which are instances of *Abstract sequence*.

We make no commitment to a particular top-level ontology. The ontology of sequences presented here can stand on its own, and axioms are presented for all relations used in the theory. However, the foundation in a top-level ontology can benefit the interoperability between the presented ontology and other domain-specific ontologies, because the top-level ontology can provide a common interface for multiple domain ontologies.

The theory is based on these primitives: the categories *Seq* of biological sequences, *Jun* of junctions, *Mol* of molecules, *ASeq* of abstract sequences, and the relations **sPO** (sequence-part-of), **PO** (part-of), **aPO** (abstract-part-of), **binds**, **::** (token-of), **Rep** (representation), **between**, **end** and **conn**.

The first part consists of axioms that restrict the arguments of some of the relations.[§] Additionally, an axiom requiring all sequences to have only molecules as tokens is introduced.

$$sPO(x,y) \rightarrow Seq(x) \wedge Seq(y) \quad (1)$$

$$PO(x,y) \rightarrow Mol(x) \wedge Mol(y) \quad (2)$$

$$Seq(x) \rightarrow \forall y(y :: x \rightarrow Mol(y)) \quad (3)$$

Based on the relation **sPO**, we first define **sPPO** (proper sequence part) and the category of primitive biological symbols (*PBS*) as well as the **soverlap** and **sdisjoint** relations:

$$sPPO(x,y) \leftrightarrow sPO(x,y) \wedge x \neq y \quad (4)$$

$$PBS(x) \leftrightarrow Seq(x) \wedge \neg\exists y(sPPO(y,x)) \quad (5)$$

$$soverlap(x,y) \leftrightarrow \exists z(sPO(z,x) \wedge sPO(z,y)) \quad (6)$$

$$sdisjoint(x,y) \leftrightarrow \neg soverlap(x,y) \quad (7)$$

The relation sPO is a parthood relation that holds for sequences when one sequence contains the other as a sequence part. It satisfies reflexivity, transitivity and antisymmetry, and therefore forms a partial order.

$$sPO(x,y) \wedge sPO(y,z) \rightarrow sPO(x,z) \quad (8)$$

$$Seq(x) \rightarrow sPO(x,x) \quad (9)$$

$$sPO(x,y) \wedge sPO(y,x) \rightarrow x = y \quad (10)$$

[§]The remaining relations take defined categories as arguments and are introduced later.

The relation sPO also satisfies the strong supplementation principle, leading to an extensional mereology for sequences[6]:

$$\neg sPO(x,y) \rightarrow \exists z(sPO(z,x) \wedge sdisjoint(z,y)) \quad (11)$$

Sequences consist entirely of atoms with respect to the relation **sPO**. The following two axioms require that all sequences have atoms as part, and that they are constituted of only atoms:

$$Seq(x) \rightarrow \exists y(PBS(y) \wedge sPO(y,x)) \quad (12)$$

$$Seq(x) \rightarrow \neg\exists y(sPPO(y,x) \wedge \forall u(sPPO(u,x) \wedge PBS(u) \rightarrow sPO(u,y))) \quad (13)$$

Next, we restrict the arguments for the **between** and **end** relation, and introduce the relation **in** through an explicit definition.

$$between(j,p_1,p_2,s) \rightarrow Jun(j) \wedge PBS(p_1) \wedge PBS(p_2) \wedge Seq(s) \quad (14)$$

$$end(j,p,s) \rightarrow Jun(j) \wedge PBS(p) \wedge Seq(s) \quad (15)$$

$$conn(j_1,j_2) \rightarrow Jun(j_1) \wedge Jun(j_2) \quad (16)$$

$$in(j,s) \leftrightarrow \exists p_1,p_2(between(j,p_1,p_2,s)) \vee \exists p(end(j,p,s)) \quad (17)$$

$$Seq(x) \rightarrow \neg Jun(x) \quad (18)$$

$$Jun(x) \rightarrow \neg Seq(x) \quad (19)$$

The following set of axioms pertains to the **conn** relation of connectedness between junctions. It is used to represent the order of the sequence through an order of junctions.

$$conn(j_1,j_2) \rightarrow conn(j_2,j_1) \quad (20)$$

$$conn(j_1,j_2) \rightarrow j_1 \neq j_2 \quad (21)$$

$$in(j_1,s_1) \wedge in(j_2,s_2) \wedge \neg soverlap(s_1,s_2) \rightarrow \neg conn(j_1,j_2) \quad (22)$$

$$conn(j_1,j_2) \wedge in(j_1,s) \rightarrow in(j_2,s) \quad (23)$$

The axioms presented here are mostly first-order axioms and do not suffice to require connectedness of sequences. Instead, a second-order axiom is required to express the fact that sequences must be connected:

$$\forall s \forall P(\forall x(P(x) \leftrightarrow in(x,s)) \wedge \forall Q(\exists a Q(a) \wedge \forall x(Q(x) \rightarrow P(x)) \wedge \forall u,v(Q(u) \wedge conn(u,v) \rightarrow Q(v)) \rightarrow \forall x(P(x) \rightarrow Q(x)))) \quad (24)$$

The remaining axioms pertain to molecules, relate sequences to their tokens or the abstract sequences they represent. They can be found in Hoehndorf [9] and in the machine implementation we provide with this paper.

A question that is not answered with these axioms is how sequences and junctions relate to categories commonly found in the top-level ontology. We believe these axioms to be compatible with most major top-level ontologies, in particular BFO[4], DOLCE[11] and GFO[7]. However, the foundation in these ontologies varies substantially.

In BFO, sequences and their junctions should be considered subcategories of *Generically dependent continuant*. A category A is generically dependent on the category B if for every instance of A, some instance of B must exist. In the framework of the BFO, sequences are generically dependent on their tokens. The difficulty that arises with such a view is that not every sequence is the sequence of a molecule. Therefore, the tokens must not be restricted to molecules which have the structure specified by the sequence, but must include textual and other digital representations as tokens of sequences. Junctions, on the other hand, always belong to a sequence and cannot exist without a sequence. Therefore, junctions should be considered as specifically dependent continuants which are dependent on sequences.

In DOLCE, the category *Abstract* is a sub-category of *Particular*. The main characteristic of abstract entities is that they do not have spatial nor temporal qualities, and they are not qualities themselves.[3] Sequences as well as junctions have this property, and the axioms we provide can be founded in the DOLCE ontology through an addition axioms:

$$Seq(x) \vee Jun(x) \rightarrow dolce:abstract(x) \quad (25)$$

Integration of our theory in GFO is similar to the scenario described in the DOLCE. Alternatively, GFO provides the category *Symbol structure*, to which both sequences and junctions can be assigned. Symbol structures are higher-order categories in the GFO, and the token-of relation $(::)$ falls together with the instantiation relation.

Implementation

We implemented the axiom system using the SPASS first-order theorem prover.[12] The implementation can be found on our project webpage.[10] Due to the restriction of SPASS to first-order logic, we could not implement the axiom requiring connectedness of sequences. This axioms necessitates the use of monadic second-order logics. Furthermore, a condition that sequences must be finite could not be implemented due to the restrictions of first order logic.

We employed the SPASS theorem prover on our axioms and attempted to prove the proposition $\phi \wedge \neg\phi$. If this logical contradiction can be derived from the

axioms we provide, our axioms would be inconsistent. On the other hand, if our axioms are consistent, we expect SPASS to never terminate, because, in the general case, an automated consistency proof for first-order theories is impossible.[1]

The SPASS theorem prover could not find a proof for the contradictory statement $\phi \wedge \neg\phi$ in three weeks time. However, this is merely an indication for consistency. A formal proof of the consistency, e.g., through the construction of a model, is subject to future work.

Conclusion

We provide an axioms system for sequences, junctions and molecules in predicate logics. Most of the axioms are available in first-order logic, although some require the use of second-order logic. The axiom system is intended to serve as a foundation of the Sequence Ontology's top-level categories *Sequence* and *Junction*. As a corollary from the axiom system, we introduced a class of sequence tokens, which we called *Molecule*. We find that in order to understand the category *Sequence*, it is necessary to consider the tokens of a sequence.

The axiom system we provide is not based on a particular top-level ontology, but is compatible with multiple top-level ontologies. We discuss how to include the theory of sequences in the BFO, DOLCE and GFO top-level ontologies. Depending on the top-level ontology used, sequences and junctions are considered different kinds of entities: generically dependent continuants in BFO, abstract individuals in DOLCE and higher-order categories in GFO.

This axiom system for sequences is – to the best of our knowledge – the first extensive axiom system for basic categories of an OBO Foundry ontology. With increasing demands for semantic interoperability and information flow between OBO Foundry ontologies, the importance of developing axiom systems likely will increase, because only axioms can provide a formal specification of a category's meaning, and therefore provide the foundation for automated inferences, information flow and integration. The new axioms are implemented for the SPASS theorem prover and can be downloaded from our website.[10]

References

1. Church A. A note on the Entscheidungsproblem. *Journal of Symbolic Logic*, 1:40–41, 1936.
2. Eilbeck K, Lewis S, Mungall C, Yandell M, Stein L, Durbin R and Ashburner M. The Sequence Ontology: A tool for the unification of genome annotations. *Genome Biology*, 6(5), 2005.
3. Gangemi A. DOLCE Lite 397 OWL File. http://www.loa-cnr.it/Files/DLPOnts/DOLCE-Lite397.owl, 2003. OWL Annotation of *abstract* class.
4. Grenon P. BFO in a Nutshell: A bi-categorial axiomatization of BFO and comparison with DOLCE. Technical report, University of Leipzig, Leipzig, 2003.
5. Guarino N. Formal ontology and information systems. In *Formal Ontology in Information Systems: Proceedings of the First International Conference (FOIS'98), Trento, Italy, 6-8 June 1998*, volume 46 of *Frontiers in Artifical Intelligence and Applications*, pages 3–15, Amsterdam, June 1998. IOS Press.
6. Guizzardi H. *Ontological foundations for structural conceptual models*. PhD thesis, University of Twente, Enschede, The Netherlands, Enschede, October 2005.
7. Herre H, Heller B, Burek P, Hoehndorf R, Loebe F and Michalek H. General Formal Ontology (GFO) – A foundational ontology integrating objects and processes [Version 1,0]. Onto-Med Report, Research Group Ontologies in Medicine, Institute of Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, 2006.
8. Hilbert D. Axiomatisches Denken. *Mathematische Annalen*, 78:405–415, 1918.
9. Hoehndorf R. *Basic considerations for improving interoperability between ontology-based biological information systems*. PhD thesis, University of Leipzig, 2009. Forthcoming: http://1c2.eu/thesis.pdf.
10. Hoehndorf R. Formal ontology of sequences. http://bioonto.de/pmwiki.php/Main/FormalOntologyOfSequences, 2009.
11. Masolo C, Borgo S, Gangemi A, Guarino N and Oltramari A. WonderWeb Deliverable D18: Ontology library (final). Technical report, Laboratory for Applied Ontology, ISTC-CNR, Trento, Italy, 2003.
12. Weidenbach C, Brahm U, Hillenbr T, Keen E and Theobald C. SPASS version 2.0. In *In Proc. CADE-18*, pages 275–279, 2002. Springer.

Hematopoietic Cell Types: Prototype for a Revised Cell Ontology

Alexander D. Diehl[1], Alison Deckhut Augustine[2], Judith A. Blake[1], Lindsay G. Cowell[3], Elizabeth S. Gold[4], Timothy A. Gondré-Lewis[2], Anna Maria Masci[3], Terrence F. Meehan[1], Penelope A. Morel[5], NIAID Cell Ontology Working Group[§], Anastasia Nijnik[6], Bjoern Peters[7], Bali Pulendran[8], Richard H. Scheuermann[9], Q. Alison Yao[2], Martin S. Zand[10], Christopher J. Mungall[11]

[1]The Jackson Laboratory, Bar Harbor, ME, USA; [2]National Institute of Allergy and Infectious Disease, Bethesda, MD, USA; [3]Duke University Medical Center, Durham, NC, USA; [4]Institute for Systems Biology, Seattle, WA, USA; [5]University of Pittsburgh, Pittsburgh, PA, USA; [6]University of British Columbia, Vancouver, BC, Canada; [7]La Jolla Institute for Allergy and Immunology, La Jolla, CA, USA: [8]Emory University, Atlanta, GA, USA: [9]U.T. Southwestern Medical Center, Dallas, TX, USA; [10]University of Rochester Medical Center, Rochester, NY, USA; [11]Lawrence Berkeley Laboratory, Berkeley, CA, USA

Abstract

The Cell Ontology (CL) aims for the representation of in vivo and in vitro cell types from all of biology. Although the CL is a reference ontology of the OBO Foundry, it requires extensive revision to bring it up to current standards for biomedical ontologies, both in its structure and its coverage of various subfields of biology. A recent workshop sponsored by NIAID on hematopoietic cell types in the CL addressed both issues. The section of the ontology dealing with hematopoietic cells was extensively revised, and plans were set for structuring these cell type terms as cross-products with logical definitions built from relationships to external ontologies, such as the Protein Ontology and the Gene Ontology. The methods and improvement to the CL in this area represent a paradigm for improvement of the whole of the ontology over time.

Overview

The Cell Ontology (CL) is an OBO Foundry candidate ontology originally built to represent *in vivo* and *in vitro* cell types, including developmental stages, of all the major model organisms.[1] The CL now aims to become a reference ontology within the OBO Foundry.[2] The CL both serves the terminology needs of data annotation, and provides a base ontology from which compound terms in other ontologies can be derived by means of cross-product term formation.[3] At Mouse Genome Informatics, the CL is used in conjunction with Gene Ontology (GO) annotation of mouse gene products to indicate the cell type in which a gene product is active. The GO itself uses CL terms in the formation of new GO terms: for instance, the GO term "leukocyte differentiation" is a cross-product of the CL term "leukocyte" with the GO term "cell differentiation."

The Cell Ontology is constructed using two relationships, *is_a* and *develops_from*. The first relationship is used to build an ontology of cellular subtypes; the latter relationship is used to indicate cell lineage relationships. The ontology as it was initially developed relied upon a number of artificial high level terms to capture types of cellular qualities, such as "cell in vivo," "cell by organism," and "cell by class," a term which itself has the *is_a* child terms "cell by function," "cell by histology," "cell by lineage," "cell by ploidy," etc. These subclasses of cells have further *is_a* children denoting more specific qualities of cells. Depending on the qualities of a particular cell type it may have one or more of these terms as *is_a* ancestors. For instance, the well-defined cell type "erythrocyte" is a type of "erythroid lineage cell," "oxygen accumulating cell," "transporting cell," and "blood cell." It also has a *develops_from* relationship with "reticulocyte."

With its multiple inheritance structure, the original CL could be described as having separate ontologies of cell types delineated by particular cellular qualities overlaid upon each other, i.e. an ontology with multiple axes of differentia that are variously and sometimes arbitrarily applied to individual cell types. Furthermore the high level terms themselves are not actual cell types, so the ontology is not a true *is_a* hierarchy. This unwieldy ontological construct is not ideal for developing proper inference about cell types, nor does it always provide obvious placement of new cell type terms.

Informal discussions among interested parties in the past few years have focused on how best to

[§] Other members of the NIAID Cell Ontology Working Group: Christopher C. Cavnor, Patrick Dunn, Thomas B. Kepler, Jingming Ma, Yuri N. Naumow, Elena N. Naumova, Jeremy Seto, and Alessandro Sette.

restructure the CL to eliminate the complexity of its multiple inheritance structure with the aim of finding a single axis of differentia upon which to base the ontology. Participants in these general discussions about the CL gradually recognized that no consistent differentia such as cellular structure or lineage can adequately describe all cell types, and that the best solution for biologists is to represent the differences and relations between cell types as scientists working in various subfields of biology do, depending on their specific criteria for differentiating cell types.

Other criticisms about the CL include the fact that many terms do not have definitions or a complete set of synonyms. Also, cell types in many subfields of biology are poorly represented within the CL. A compounding issue has been the lack of a full-time curator for the ontology as a whole. Efforts at improvement have been made in certain areas of the ontology, and hematopoietic cell types in particular have been the focus of two rounds of intensive curation in recent years. Here we report on these revisions and examine the process as an example for the future development of the Cell Ontology.

Hematopoietic Cell Type Revisions

The first set of improvements for hematopoietic cells was done in 2006 in conjunction with the revision of the terms for immunological processes in the GO.[4,5] At that time 80 new hematopoietic cell type terms were introduced, many other terms were revised, and many improvements in ontology structure were made for these cell types.

A second, larger round of revisions to the hematopoietic cell type terms in CL is described herein. These revisions are the product of a National Institute of Allergy and Infectious Disease (NIAID) sponsored "Workshop on Immune Cell Representation in the Cell Ontology," held in May 2008, where domain experts and biomedical ontologists worked together on two goals: 1) revising and developing additional specific terms for T cells, B cells, natural killer cells, monocytes and macrophages, and dendritic cells, and 2) establishing a new paradigm for development of the CL. These changes in the representation of hematopoietic cells were needed to represent these cell types in a more complete manner so that all major cell types identified in the literature are found in the ontology and so that these cell types are defined in an in-depth manner that greatly increases the descriptiveness of the ontology for data annotation and logical inference.

Methods

The NIAID workshop attendees discussed both specific groups of cell types of interest to immunologists as well as how to improve the overall ontological structure of these groups and the CL ontology in general. The consensus view was that the current multiple inheritance structure of the CL is unsustainable and that existing and new terms for hematopoietic cells should be logically defined via their qualities as represented in other ontologies. Much discussion centered on what might be the optimal axis of differentia for these hematopoietic terms. It was recognized in many cases that these cell types are defined largely, but not solely, by the expression of particular marker proteins either at the cell surface (e.g. receptor proteins) or internally (e.g. transcription factors). The presence of these proteins as part of a cell is considered a structural feature of the cell, and participants agreed that the relationship *has_part* from the OBO relationship ontology would be used to relate particular cell types to protein terms from the Protein Ontology.[6,7]

However, for certain cell types, such as macrophages, it was seen that the full molecular characterization of different types of macrophages is still not complete in the literature, and that anatomical location serves as a major differentia for these cells. For other cell types, functional or lineage criteria serve as differentia for the complete definition of the cells. Functional criteria include the ability to execute or participate in particular GO processes that relate to individual cells, such as "cytotoxicity" or "cytokine production," or GO processes that involve coordination of multiple cell types, such as "T-helper 1 type immune response." Thus, the participants at the workshop agreed to focus on structural criteria where possible as the primary differentia, but to accept other types of differentia when necessary. This flexibility should make it possible to stick to the commonly accepted biological definitions of individual cell types and to organize the ontology according to sound ontological principles while still reflecting organization of hematopoietic cell types seen in the literature.

The primary goal in revising the hematopoietic cell terms is to define all the terms according to logical definitions based on relationships to external ontologies. The workshop participants recognized that reaching the full development of these terms as cross-products would be difficult at this time due to the lack of a full-time curator for the CL. Also, external ontologies, such as the Protein Ontology, are not yet complete in all the required terms. Yet at the same time, the new hematopoietic cell terms are

needed for data annotation and development of cross-products in the GO and other ontologies.

Results: A Two-Stage Process

Reflecting the above considerations, the participants at the NIAID workshop agreed upon a two-stage approach to further development of the hematopoietic cells in the Cell Ontology. In the first stage, which is now complete, current terms were revised and new terms added by the experts at the workshop. The textual definitions for these terms contain all the necessary details to define the cells logically. These terms have been directly incorporated into the existing ontology. It was also decided to separate the hematopoietic terms from the multiple inheritance hierarchy of the original CL as much as possible, so that the section of the ontology containing these terms represents a true ontology hierarchy. This first-stage ontology has been given the working name "CL1.5." Figure 1A shows a typical OBO term stanza for one of these new terms, "induced T-regulatory cell."

The second stage will then be the development of the hematopoietic terms into full cross-products as discussed above. The extended definitions provided in the first step will hopefully enable this to be done in a fairly efficient manner depending upon the availability of the necessary terms in external ontologies. Ideally, this approach will be extended to the whole of the CL to create version "CL2.0." For the moment we plan to develop the hematopoietic terms of the CL into an external mini-ontology based on these cross products, "hemo-CL." Figure 1B shows the OBO term stanza for term "induced T-regulatory cell" as it will be represented in hemo-CL and CL2.0, illustrated graphically in Figure 1C. We have already been working with the curators of the Protein Ontology to ensure that protein terms needed for hemo-CL are found in the Protein Ontology.

The initial step towards hemo-CL and CL2.0 has been taken by Masci and colleagues, who have developed a dendritic cell ontology, DC-CL, which is based on cross-product principles and is the foundation of the revised dendritic cell terms in CL1.5.[8] DC-CL terms for types of dendritic cells are primarily based on structural criteria (surface protein expression) with a few cell types also defined by relationships to functions or dispositions. DC-CL utilizes an expanded range of relationship types based on those in the relationship ontology in order

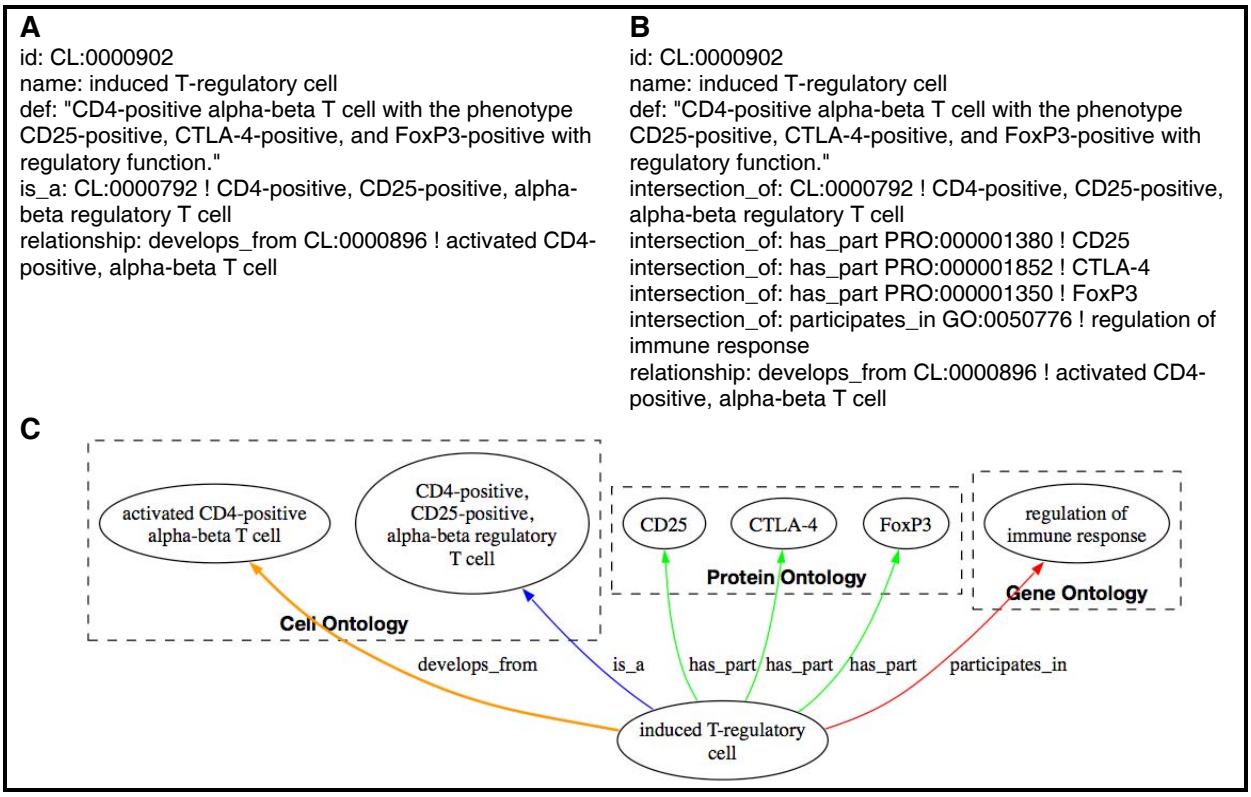


Figure 1. Examples of improvement in the representation of hematopoietic cells.
A. OBO term stanza representative of CL1.5 term definitions for the term "induced T-regulatory cell."
B. OBO term stanza representative of CL2.0 showing logical definition of the same term as in A.
C. Graphical view of the term relationships in B.

to be more expressive about the cellular location and degree of protein expression (*has_membrane_part*, *has_high_membrane_amount*). It is likely that similar relationships will be employed in the construction of hemo-CL and CL2.0.

Specific Improvements in the Representation of Hematopoietic Cell Types

With the work initiated at the NIAID workshop and carried on afterwards, many concrete improvements to CL content in the area of hematopoietic cells have been achieved. Many new terms for individual cell types have been created, including over 40 terms for T-lineage cells, over 40 terms for B-lineage cells, several natural killer cell terms, over 30 terms for monocytes and macrophages, and over 30 terms for dendritic cells. Other new terms were introduced for various hematopoietic progenitor cell types. As discussed above, most of these new terms have been defined by structural criteria (protein expression) sometimes in conjunction with functional or anatomical relationships. The exception to this general rule is that most of the new macrophage terms are defined based on their anatomical location with protein expression criteria added where supported by the literature.

The ontology structure has been improved as well in important areas such as T cell and B cell development. Lineage relationships via the *develops_from* relationship have been provided for many additional cell types. In general the hematopoietic terms are intended to be species neutral, but species-specific information has been incorporated in some definitions where necessary and comments added to provide clarity to data annotators.

Discussion

The Cell Ontology is an essential core component of the OBO Foundry and has great potential for aiding data annotation and analysis. With the improvements described herein, implemented for CL1.5, and planned for hemo-CL/CL2.0, we expect the CL to fulfill much more of its promise in the area of hematopoietic cell representation. The ontology now has fairly complete coverage of these cell types in an improved hierarchy and using up-to-date molecular definitions. These changes will provide for more robust inference across the ontology and greater utility for annotation of hematopoietic cell type data, and will strengthen the use of the CL as a reference ontology for cross-product development.

The workshop approach, aided by an acting editor for this section of the ontology, has worked reasonably well in carrying out the needed additions and revisions in the ontology content in this area, and in outlining a clear plan for the future of the ontology. The section-by-section approach for improvement of defined parts of the Cell Ontology represents a paradigm for continued development of the CL and should prove even more useful once dedicated funding is achieved.

Acknowledgements

We thank NIAID for the support of the workshop and follow-up teleconferences. ADD and JAB are supported by NHGRI grant HG002273, RHS by NIAID grant N01AI40076, PAM by NIAID grant N01AI50018, and BP by NIAID grant N01AI50019.

References

1. Bard J, Rhee SY and Ashburner M. An ontology for cell types. Genome Biol. 2005;6:R21.
2. Smith B, Ashburner M, Rosse C, *et al*. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007;25:1251–1255.
3. Hill DP, Blake JA, Richardson JE and Ringwald M. Extension and integration of the gene ontology (GO): Combining GO vocabularies with external vocabularies. Genome Res. 2002;12:1982–91.
4. Diehl AD, Lee JA, Scheuermann RH and Blake JA. Ontology development for biological systems: immunology. Bioinformatics. 2007;23:913–5.
5. Diehl AD, unpublished.
6. Smith B, Ceusters W, Klagges B, *et al*. Relations in biomedical ontologies. Genome Biol. 2005;6:R46.
7. Natale DA, Arighi CN, Barker WC, *et al*. Framework for a protein ontology. BMC Bioinformatics. 2007;8 Suppl 9:S1.
8. Masci AM, Arighi CN, Diehl AD, *et al*. An improved ontological representation of dendritic cells as a paradigm for all cell types. BMC Bioinformatics. 2009; 10:70.

Using OWL Metamodeling to Create an Ontology of Neurons

Matthew E. Holford[1], Luis N. Marenco[1], Pradeep Mutalik[1], Gordon M. Shepherd[2], Perry L. Miller[1], Kei-Hoi Cheung[1]
[1]Center for Medical Informatics, [2]Department of Neurobiology, School of Medicine, Yale University, New Haven, CT, USA

Abstract

This paper presents the Senselab Neuron Ontology (SLNO), a new ontology to describe neuron types and their properties. The ontology is generated dynamically from the Senselab databases. The project makes extensive use of the new metamodeling capabilities offered in OWL 2. This paper also discuss the consequences of modeling relationships between classes as collections of statements. The SLNO will serve as the basis of the new SWISL web application, a context-aware browser for neurological information on the semantic web.

Introduction

In recent years, the semantic web has begun to see exploratory use in biomedical research. Scientists hope that the terminological precision and logical rigor offered by semantic modeling will resolve some of the challenges posed by the recent influx of diverse biological information. Ontologies may be particularly helpful in the integration of heterogeneous data models and the resolution of diverse nomenclatures. These issues plague neuroscience much as they do any area of biomedicine[5]. Two notable recent endeavors are NIF (Neuroscience Information Framework)[6] and SWAN (Semantic Web Application in Neuromedicine)[7]. The NIF project is focused upon producing a controlled vocabulary of neuroscience terms and a taxonomy of cells and their properties. SWAN, on the other hand, hopes to provide an ontology of biomedical discourse, a framework for integrating statements made in neurological research.

It is often useful in the process of modeling data to make statements about classes of objects in much the same way as we would make statements about individual objects. This technique is known as metamodeling. As a generic example, take the statement: “This box contains grapes.” It is clear that “this box” should be modeled as an individual, an instance of a hypothetical Box class. Less certain is how to treat the grapes contained within the box. In a certain sense we are treating Grapes as an individual object, an instance, suppose, of a Cargo class of which other individuals could be Textiles or Treasure. This is less than ideal, however, because, in reality, Grapes are a class of object. We are not referring to the word “Grapes” or to an idealized concept of a “Grape.” More importantly, by treating Grapes as an individual object we are painting ourselves into a corner. If we wish to model Grapes more specifically elsewhere in our ontology, to describe, for example, individual varieties of grape, or to share data with an ontology that does so, we will be unable to associate these data with the grapes that are in our box. Thus, it is important to preserve the class identity of Grapes.

In the past, this raised an additional problem. While making a statement about a class in the same way that one would about an individual is perfectly legal in RDF and OWL Full, it was forbidden in the initial version of OWL DL, which requires that the pool of object names be distinct from the pool of individual names. If one wanted to take advantage of the features offered by description logic reasoners, a workaround was required. The most common solution is sometimes referred to as the Relationship Transfer Pattern[8]. In our example, we would define a subclass of Box called GrapeBox which would contain the restriction that GrapeBoxes contain Grapes. Then we define the box we are referring to as an instance of the GrapeBox class. Though this technique works, it is much less straightforward than our original statement. It requires three declarations rather than one and if used repeatedly leads to a class model that is complicated and difficult to read. Although we have used a simple example, the same issues arise when one tries to model the fact that a specific type of neuron may be found in a specific region of the brain.

Fortunately, versions 1.1 and 2 of OWL DL allow a technique called 'punning', whereby class names can be treated as individual names[9]. Although OWL 1.1 and OWL 2 are not yet official standards, punning is already supported by several OWL tools, including Pellet[10], a widely used reasoner. The Senselab Neuron Ontology makes extensive use of punning to model statements about neurons and their properties.

Results

The Senselab Neuron Ontology (SLNO) is made up of two components, a taxonomic part and a data descriptive part. The first portion provides a

canonical hierarchy of neurological terms for morphology and cellular function. The second portion provides a format for statements which make use of the foundational terms. A collection of such statements culled from research literature is also included. This second portion provides us with a framework from which we can determine relationships between neurons and their properties. It makes use of the punning technique we describe above. The foundational portion defines classes and properties which build upon the BFO (Basic Formal Ontology)[11] and RO (Relationship Ontology)[12], base ontologies which define terms common to all knowledge models. For example, the SLNO class *Neuron* is defined as a subclass of the BFO term *snap:Object*, which is turn is defined as a subclass of *snap:IndependentContinuant*. The SLNO property *has_neuron*, which is used to indicate that a *BrainRegion* may hold a particular kind of *Neuron*, is a subproperty of the RO term *contains*. It is hoped that our ontology can be easily integrated with ontologies that also build upon these base units such as the OBO (Open Biomedical Ontologies) collection[13]. Thia ontology can be accessed at: http://bioinformatics.med.yale.edu/owl/slno.owl.

Our ontology is built on top of the well-established Senselab project, a collection of databases which outline terms for neurological processes and categorize recent research under these terms[14]. Individual databases exist for areas such as neurons, odor receptors and neuronal models. At present, we focus upon NeuronDB, which describes individual types of neurons, their structure, their membrane properties and where they may be found. Like all of the Senselab databases, NeuronDB is regularly updated to reflect new domain knowledge. It was of great importance to us in designing the SLNO that we remain synchronized with Senselab. This centralizes update of neurological information and allows our ontology to benefit from active curation.

Data is imported into SLNO using Senselab's EDSP format, an XML serialization of the EAV/CR database format[15]. EDSP files consist of two sections. The first part uses metadata properties to describe class-like objects and their fields. The second fills in instances of these classes. Through the use of our conversion library, called Senselib, the data structures described in EDSP are converted into a graph of Python objects. Each EAV/CR class corresponds roughly one-to-one with a python class, although there are important divergences between the taxonomy of the ontology and that of the original NeuronDB. For example, Senselab defines a handful of classes for denoting regions of the central nervous system: *GeneralRegion*, *SpecificRegion* and *Subdivision*. For the SLNO, we found it simpler and more intuitive to merge these classes into a single hierarchy of subclasses of *CNSRegion*.

A handful of core ontology classes are modeled by hand as python classes. These base classes include *Neuron*, which is subclassed into *Interneuron* and *PrimaryNeuron*; *CNSRegion*; *Compartment*, a part of a neuron; *CanonicalForm*, which describes what compartments may be found in what neurons; and *MembraneProperty*, which is subclassed into *Current*, *Ion*, *Neurotransmitter*, and *Receptor*. We recognize that the Neurotransmitter class is problematic in that it defines a role rather than a category of object. It is very convenient, however, to be able to group these entities. Each of these classes inherits from a single base class called *SLObject* which defines three common properties, *has_description*, *has_name* and *has_senselab_id*. These properties are populated dynamically from the EDSP files. The base python classes are serialized into OWL using the rdflib package. All further subclasses are added dynamically by parsing EDSP files. Individual subclasses are marshaled into instances of the core python classes and then serialized into OWL classes. The basic properties describing Senselab metadata are attached as OWL AnnotationProperties. Further properties include a *neuron_in* property which is used to indicate in what regions a neuron may be found and a *has_neuron* property to declare what neurons are contained within a region. Similar properties are used to describe the relationships between neuron compartments and canonical forms. Most of the broader classes are annotated with definitions to clarify their meaning within the ontology. Chemical entities are further annotated with their ChEBI identifiers[16].

The second or, data interchange, component of our ontology consists of a single class called *NeuronPropertyStatement*. The OWL ObjectProperties *describes_neuron*, *describes_compartment* and *describes_property* are used to hold pointers to the particular class to which the statement refers. This is accomplished by using the punned name of the class as described earlier. An example insertion could be expressed using the OWL2 functional syntax as:

```
ClassAssertion(NeuronPropertyStatement NA-1)
SubClassOf(Interneuron Thalamic_reticular_neuron)
SubClassOf(Compartment Soma)
SubClassOf(I_Sodium I_Na_p)
ObjectPropertyAssertion(describes_neuron NA-1
        Thalamic_reticular_neuron)
```

```
ObjectPropertyAssertion(describes_compartment NA-1
    Soma)
ObjectPropertyAssertion(describes_property NA-1
    I_Na_p).
```

Other properties are used to provide specifics on the the source of the observation and more detailed notes about it. An advantage of this approach is that individual statements can be associated with particular publications. We hope, eventually, to represent formally the citations mentioned in the notes, using properties from the SWAN ontology. Statements that disagree can be compared; observations can be validated or refuted. Querying of the statement collection is nearly as straightforward as insertion of new data. For example, if we wish to know for which neurons, the glutamate receptor mGluR2 has been found to be present, we would issue the simple SPARQL query:

```
select distinct ?n
where {
  ?s :describes_neuron ?n .
  ?s :describes_property :mGluR2_Receptor }
```

To issue a more general query such as to find what neurons have been found to contain any of the glutamate receptors, we would alter our query to take advantage of OWL DL subclass inferencing:

```
select distinct ?n
where {
  ?s :describes_neuron ?n .  {
    ?s :describes_property :Glutamate_Receptor
  } union {
    ?s :describes_property ?p .
    ?p rdfs:subClassOf :GlutamateReceptor }}
```

Our decision to model neurological observations as individual statements reflects a certain strategic tradeoff in that we use ontological terms to describe data rather than to attempt a canonical representation of reality. The advantages to this approach are not limited to practicality, although the format is similar to the EAV structure of the Senselab database. New statements can be added as information becomes available and spurious or conflicting information can be easily filtered out. Of course, at some point it may become useful for information about what properties might be observed in particular neurons to be included within a foundational ontology of neurology. Such an ontology could be constructed dynamically in a relatively straightforward manner by querying the pool of *NeuronPropertyStatements* for results that meet our criteria.

An earlier version of the NeuronDB ontology attempted to provide this type of foundational model[17]. In this model, individual neuron types are further subclassed by a restriction on the membrane property found in that neuron and the compartment in which that property was found. This is accomplished using a compound restriction on the has_part property which is imported from the OBO Relation Ontology; for example: "ro:has_part some (Soma and not (has_Current some I_Na_t))". These subclasses are given names, although this is not strictly required, such as *CA1_oriens_alveus_interneuron_with_ GabaA_ receptor_in_Soma*. Although this approach has the advantage of merging information about membrane properties with the neuron taxonomy, it does so at the expense of modeling clarity and ease of querying. A more intuitive approach might be to use separate properties to indicate presence of a compartment and presence of a membrane property, rather than expressing both of them with the *has_part* property. We could also punning to state directly the properties and compartments that are present instead of enforcing them through a restriction object.

We expect that end users will interact with our ontology in two ways. They may wish to use the terms from the foundational portion of our ontology to further describe neurons and neuronal properties, either by subclassing or through statement objects such as those of our *NeuronPropertyStatement* class. Secondly, the data representative portion can be used to facilitate a browser for information about neurological research.

Future Directions

In line with this second purpose, we are using the Senselab Neuron Ontology as the basis for SWISL, the Semantic Web Interface to Senselab, a re-designed version of the Entrez Neuron neuroinformatics browser[18]. This web application allows users to perform searches using the hierarchy of terms from our ontology and obtain results collected from research literature. Because the search terms are encoded with their meaning, the results are potentially more informative than those obtained using a standard search engine. SWISL is implemented using the Turbogears python web development framework[19]. Data is stored in RDF triples using the Virtuoso Open Source database[20]. To facilitate the full range of OWL DL reasoning while maintaining optimal performance, all logical entailments are generated beforehand by running the data through the Pellet reasoner prior to insertion into Virtuoso. Interaction with the data store is handled through the rdfalchemy library[21], which provides an object-relational mapping layer between python

objects and SPARQL in much the fashion as tools like Hibernate, SQLAlchemy and LINQ do for SQL. Several extensions to rdfalchemy were created to allow it to handle more complex joins and filters. SWISL is available for use at: http://bioinformatics.med.yale.edu/swizzle.

In the near future, we hope to strengthen our ontology through integration with the NIF project. As developers of Senselab, we are part of the multi-institutional consortium which is collaborating to define the NIF standard ontology. As we benefit from the use of standardized neurological vocabulary defined by NIF, we hope that the observations about neuron property relationships that our ontology makes possible can be used to extend the NIF ontology. We also would like to incorporate information from other neurological databases into our neuron ontology. To start, this would involve incorporating some of the other Senselab databases to allow us to express relationships such as those between neurons and genes or neurons and research models. Additionally, the modeling strategies we employed in the creation of SLNO can be used in a similar fashion to allow the SWISL application to query neurological data from a wide variety of sources.

Acknowledgements

This work was supported in part by NIH grants P01 DC04732, R01 DA021253 and U24 NS051869.

References

1. Ashburner M, Blake J and Stein L. Extreme curation. Comput Biomed Res. 1990;23:514–28.
2. Rubin, DL, Shah, NH and Noy, NF. 2007. Biomedical Ontologies: a functional perspective. Briefings in Bioinformatics, 9(1): 75–90.
3. RDF Primer [online]. Accessed February 23, 2009. URL: http://www.w3.org/TR/REC-rdf-syntax.
4. OWL Web Ontology Language Overview [online]. Accessed February 23, 2009. URL: http://www.w3.org/TR/owl-features.
5. Lam YK, Marenco L and Shepherd GM. 2006. Using Web Ontology Language to Integrate Heterogeneous Databases in the Neurosciences. AMIA 2006 Symposium Proceedings: 464–468.
6. Bug WJ, Ascoli GA and Grethe, JS. 2008. The NIFSTD and BIRNLex Vocabularies Building Comprehensive Ontologies for Neuroscience. Neuroinformatics, 6: 175–194.
7. Ciccarese P, Wu E and Wong, G. 2008. The SWAN biomedical discourse ontology. Journal of Biomedical Informatics, 41: 739–751.
8. Allemang D and Hendler J. 2008. Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL. San Francisco: Morgan Kaufmann Publishers.
9. OWL 2 Web Ontology Language: Profiles [online]. Accessed February 23, 2009. URL: http://www.w3.org/TR/2008/WD-owl2-profiles-20081202.
10. Sirin E, Parsia B and Grau BC. 2007. Pellet: A practical OWL-DL reasoner. Web Semantics Science, Services and Agents on the World Wide Web, 5: 51–53.
11. The Basic Formal Ontology (BFO) [online]. Accessed February 23, 2009. URL: http://www.ifomis.org/bfo/1.1#.
12. Smith B, Ceusters W, Klagges B, *et al*. 2005. Relations in biomedical ontologies. Genome Biology, 6: R46.
13. Smith B, Ashburner M, Rosse C, *et al*. 2007. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology, 25(11): 1251–1255.
14. Craso CJ, Marenco LN, Liu N, *et al*. 2007. SenseLab: new developments in disseminating neuroscience information. Briefings in Bioinformatics, 8: 150–162.
15. Marenco L, Tosches N, Crasto C, *et al*. 2003. Achieving evolvable web-database bioscience applications using the EAV/CR framework: recent advances. Journal of the American Medical Informatics Association, 10: 444–453.
16. ChEBI [online]. Accessed June 9, 2009. URL: http://www.ebi.ac.uk/chebi.
17. Experiences with the conversion of SenseLab databases to RDF/OWL [online]. Accessed February 28, 2009. URL: http://www.w3.org/TR/hcls-senselab.
18. A portal to the Semantic Web of Neuroscience [online]. Accessed February 23, 2009. URL: http://neuroweb3.med.yale.edu:8090.
19. Turbogears [online]. Accessed February 23, 2009. URL: http://turbogears.org.
20. VOS: Virtuoso Open-Source Edition [online]. Accessed February 23, 2009. URL: http://virtuoso.openlinksw.com/wiki/main/Main.
21. RDFAlchemy – Openvest – Trac [online]. Accessed February 23, 2009. URL: http://openvest.com/trac/wiki/RDFAlchemy.

Development of Neural Electromagnetic Ontologies (NEMO): Ontology-Based Tools for Representation and Integration of Event-Related Brain Potentials

Gwen Frishkoff[1,2], Paea LePendu[2], Robert Frank[2], Haishan Liu[2], Dejing Dou[2]
[1]Medical College of Wisconsin, Milwaukee, WI, USA; [2]University of Oregon, Eugene, OR, USA

Abstract

We describe a first-generation ontology for representation and integration of event-related brain potentials (ERPs). The ontology is designed following OBO "best practices" and is augmented with tools to perform ontology-based labeling and annotation of ERP data, and a database that enables semantically based reasoning over these data. Because certain high-level concepts in the ERP domain are ill-defined, we have developed methods to support coordinated updates to each of these three components. This approach consists of "top-down" (knowledge-driven) design and implementation, followed by "bottom-up" (data-driven) validation and refinement. Our goal is to build an ERP ontology that is logically valid, empirically sound, robust in application, and transparent to users. This ontology will be used to support sharing and meta-analysis of EEG and MEG data collected within our Neural Electromagnetic Ontologies (NEMO) project.

Introduction

In the last two decades, neuroscience has witnessed the development of some exciting new methods for research on human brain function—including high-density electroencephalography (EEG), whole-head magnetoencephalography (MEG), and functional Magnetic Resonance Imaging (fMRI). Each of these methods has contributed important insights on human brain function. At the same time, the proliferation of data has made clear the need for large-scale summary and integration of research results. To meet this need, several groups have been working to develop formal ontologies that can be used for consistent annotation, sharing, and meta-analysis of neuroscience data[1,2].

In the present paper, we describe initial steps in the development of an ERP (event-related potentials) ontology. ERPs are measures of brain electrical activity (EEG or "brainwaves") that are time-locked to experimental events (e.g., the appearance of a word). These measures provide a powerful technique for studying brain function, because they are acquired noninvasively and can therefore be used in a variety of populations —e.g., children and patients, as well as healthy adults. In addition, they provide detailed information about the time dynamics, as well as the scalp spatial distribution, of neural activity during various cognitive and behavioral tasks.

ERP research is likely to enjoy several benefits from the development of ERP ontologies. Historically, progress in this area has been hampered by debates over how to define high-level concepts[3]. As a result, it is has been hard to achieve even informal consensus, let alone quantitative syntheses of results across experiments (i.e., statistical "meta-analysis"). In this context, the process of building an ontology may prove to be a healthy exercise. Where there are debates over concepts, the need to make these concepts explicit will bring controversies into the open. Where there is mere inconsistency in labeling, the existence of a common reference may lead to standards for reporting that will better support cross-lab integration of research results.

To address these aims, we have assembled an international team of ERP researchers and computer scientists to found the Neural Elecromagnetic Ontologies (NEMO) consortium[3,4]. The major goal of our project is to address basic scientific questions in ERP research using ontology-based classification and labeling of ERP data, particularly in studies of language comprehension. The present paper gives an overview of the NEMO project and describes how it builds on and extends other efforts in bio- and neuro-ontology development.

NEMO Framework

Our framework includes the following components:

1. Top-down (knowledge-driven) specification and coding of domain concepts (*NEMO ontologies*);
2. Bottom-up (data-driven) validation and refinement of complex concepts, including tools for *ontology-based labeling of ERP data*;
3. An international *consortium of experts* in ERP methods, with a shared interest in language;
4. An *ontology database* and *inference engine* to enable semantic queries over labeled data.

Each of these components is described in the following sections.

Top-down Ontology Development

Traditional methods for ontology development can be described as top-down or *knowledge-driven*, and are largely manual. The process typically begins with knowledge capture, that is, expert identification of a relatively small set of domain concepts. In NEMO, we have focused on defining concepts that represent spatial and temporal attributes of ERP patterns, as well as some functional (i.e., cognitive) concepts that are of immediate interest for analysis of ERP experiment data (building on previous efforts in the development of cognitive ontologies[1,2]). In addition, because our goal is to use ontologies to develop tools for labeling of ERP data, we have represented data-level concepts in a separate but linked namespace. These first steps in ontology development are documented in NEMOlex, a text document that was modeled after Neurolex (formerly BirnLex[2]). NEMOlex contains natural language descriptions of concepts (classes and relations), organized by categories (e.g., spatial, temporal, and functional).

In the next step, domain experts work with ontology engineers to develop a formal conceptualization of domain-specific concepts. These concepts are subsequently coded in the Web Ontology Language (OWL), and Protégé is used to generate a set of web-accessible documents that can be viewed online (see nemo.nic.uoregon.edu for links to owl ontologies).

Throughout this process we have worked to implement recommendations of the Open Biomedical Ontologies (OBO) community[5]. Domain-specific concepts in NEMO are linked to more basic or foundational concepts, as implemented in the Basic Formal Ontology (Figure 1). Similarly, to facilitate reuse and integration of NEMO with other neuroscience ontologies, we have aligned our efforts with members of the OBO, including fBIRN and NIFSTD. For example, the NEMO concept *scalp* is defined as a proper_part_of NeuroLex class *head.* In addition, we have designed NEMO ontologies to be modular wherever possible. Concepts representing spatial, temporal, and functional objects and

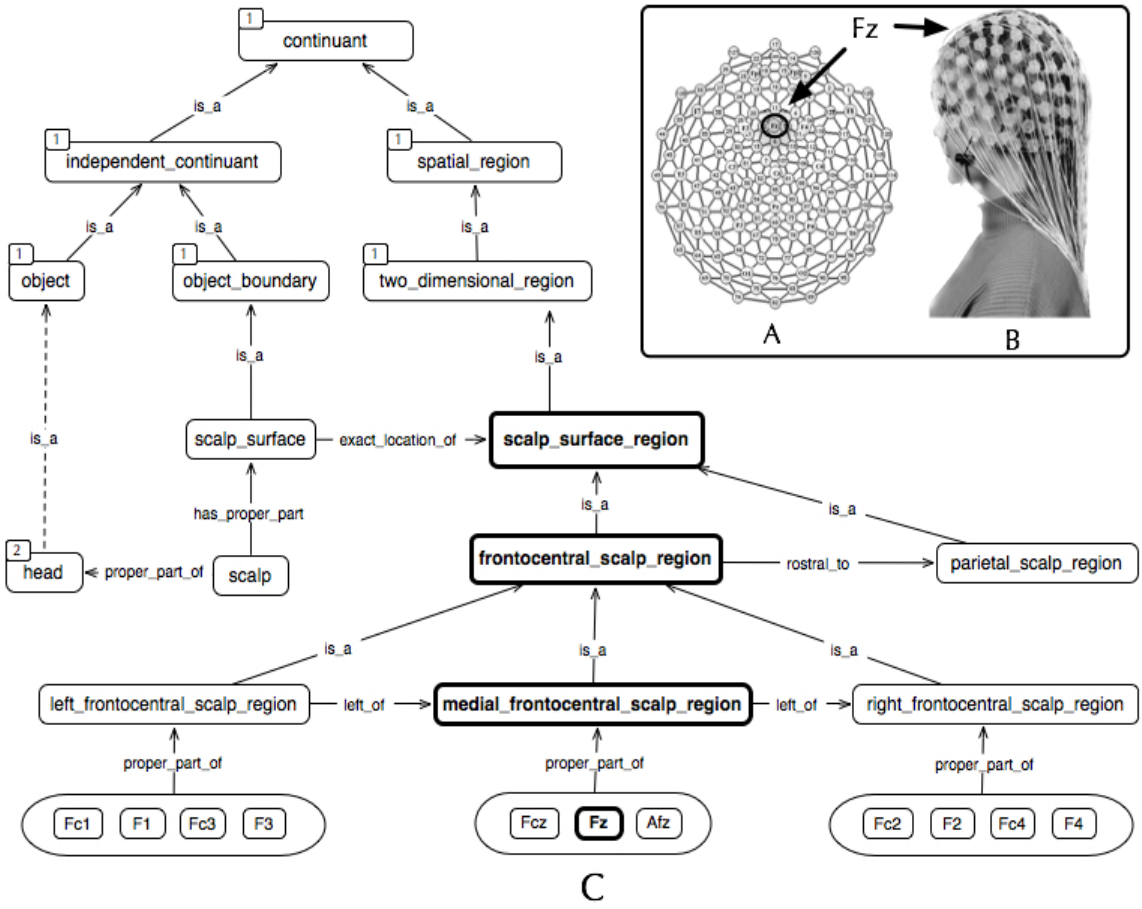


Figure 1. A) International 10-10 electrode layout (i.e., *scalp_surface_region*), with *electrode_location Fz* highlighted; B) EEG net applied to the scalp surface. C) A subset of concepts from NEMO_spatial. Concepts marked by superscript '1' are from BFO; superscript '2' denotes concepts from NeuroLex (formerly BirnLex).

properties are therefore stored in different name spaces (NEMO_spatial, NEMO_temporal, and NEMO_functional). Of key importance are ERP spatiotemporal patterns that are seen in particular experiment contexts. These patterns have distinctive spatial, temporal and functional attributes as described in the following section. Pattern definitions are represented as first-order rules in our merged NEMO_erp ontology.

Bottom-up Validation and Refinement

While ontologies are intended to capture expert (i.e., domain) knowledge, knowledge in certain areas may be uncertain or changing. For example, spatiotemporal ERP patterns, which are the main concepts of interest in the ERP domain, are often ill-defined. The same label (e.g., "N400") can be used to pick out manifestly different entities[3]. Conversely, the same pattern may be called by different names in different experiment paradigms or research groups.

The existence of a standard ERP ontology can help to address this lack of consistency, but there is no guarantee that concepts defined using "top-down" methods will be optimal for classification, labeling, and annotation of actual ERP data. To address this concern, NEMO has adopted a *data-driven* strategy for validating and refining high-level patterns before encoding this knowledge in our ontologies. This strategy is used to augment first-pass ontology engineering steps described in the previous section.

Our approach is outlined in Figure 2. It begins with expert specification of spatial, temporal, and functional concepts, including definitions of patterns that are commonly found in ERP data. These definitions represent expert hypotheses about (a) the ERP patterns that exist and (b) the spatial and temporal attributes that define these patterns (see Ref. [3], Appendix B for concrete examples). To test these hypotheses, we encode these pattern rules in an automatic data classification and labeling tool. ERP datasets are summarized by extracting attribute vectors that constitute a compact summary of the measured data. The values of these spatiotemporal metrics are then compared to rule-specific thresholds for each ERP pattern of interest. Results are recorded in a true/false table, and observations meeting pattern criteria are flagged as instances of that pattern.

Next, we perform clustering on the spatial and temporal values of these summary metrics using the Expectation-Maximization (EM) algorithm[3,4]. The resulting clusters represent candidate pattern classes, which are characterized by the central tendencies of their cluster attributes (e.g., mean latency and amplitude over scalp regions of interest). Based on these results, we refine our initial hypotheses about the number of pattern classes in the ontology and the definitions of these patterns. If similar results are obtained across multiple datasets, this leads in turn to a revision of NEMO ontologies and ontology-based tools for pattern classification and labeling.

We have applied these methods to several datasets[3,4], and results have led to refinements of our methods for ontology-based labeling. In our current ERP labeling tools, for example, we have omitted reference to high-level ERP pattern concepts, such as the "N400." Concepts are still coded in the NEMO_erp ontology, but with provisional notes that indicate they are based on working hypotheses that

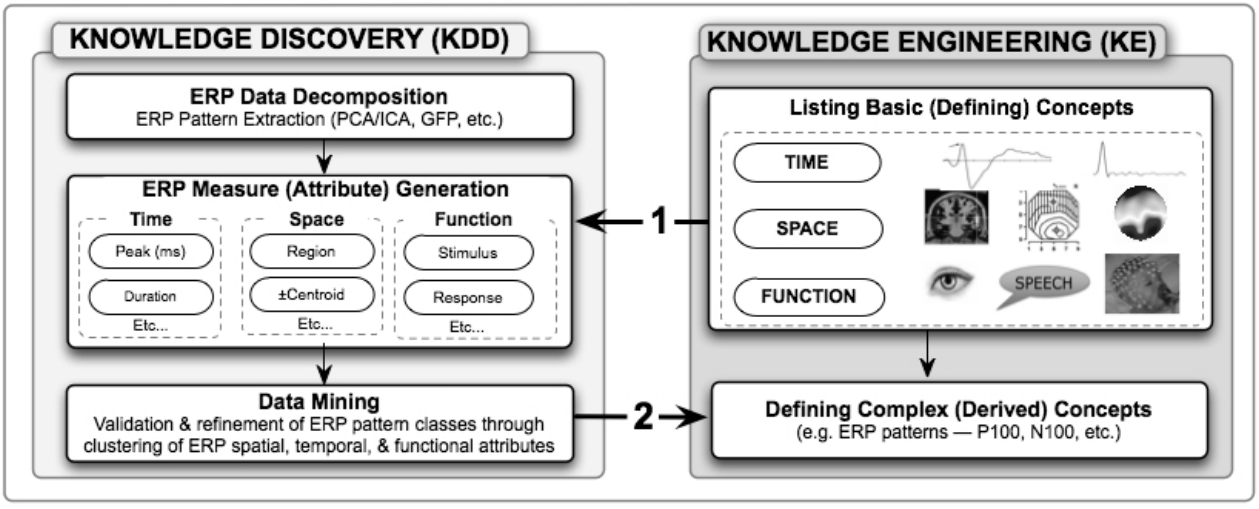


Figure 2. NEMO framework for deriving complex concepts for ERP ontology.

are awaiting robust empirical testing and validation.

The NEMO Consortium

The application domain for our project is language processing. We have established an international consortium of experts in this area who contribute ERP data from experiments and collaborate on the design and testing of ontology-based tools developed for NEMO. Consortium members include John Connolly (McMaster U.), Timothy Curran (U. Colorado), Dennis Molfese (U. Louisville), Charles Perfetti (U. Pittsburgh), Joseph Dien (U. Maryland), and Kerry Kilborn (Glasgow U.).

The NEMO Ontology Database

The NEMO database will store large numbers of ERP datasets collected from multiple research sites (e.g., from members of our research consortium). As described above, we have developed MATLAB scripts that automatically decompose, classify, label, and annotate ERP data using ontological terms. On the backend, we will support ontology-based querying and reasoning by using specialized databases designed to model the class (subsumption) hierarchy as well as most integrity and cardinality constraints. These databases will be coupled with a reasoning engine (OntoEngine[6]) to support efficient and scalable knowledge-based query answering. For example, consider the following database query:

Return all data instances that belong to ERP pattern classes which have a surface positivity over frontal regions of interest and are earlier than the N400.

In this query, "frontal region" is a clear generalization that can be unfolded into constituent parts (e.g., right frontal, left frontal; see Figure 1). At an even more abstract level, the "N400" is a pattern class that is associated with spatial, temporal, and functional properties (Figure 2). As described above, these three types of concepts are encoded in separate namespaces, and linking concepts are used to combine them for definition of high-level pattern concepts in NEMO. This design allows for a rich and flexible range of queries, which we refer to as *ontology-based queries*[7].

NEMO has investigated several methods of using databases to support ontology-based queries. A view-based approach is commonly used to simplify instance-checking and subsumption-based reasoning by unfolding views at query time. By contrast, we have developed a new method that uses asynchronus, event-driven triggers to forward-propagate the knowledge model so that queries are answered more quickly and efficiently[7].

Summary and Conclusions

In conclusion, we have described a first-generation ontology for representation and integration of event-related brain potentials. The ontology is designed following OBO "best practices" and is augmented with tools to perform ontology-based labeling and querying of ERP data.

We have further described how data mining (i.e., clustering) is used to help validate and refine top-down ERP ontologies. These ontologies will be used to support sharing and meta-analysis of cognitive neuroscience data collected within the NEMO project.

Acknowledgements

This work is supported by a grant from the National Institutes of Health (NIH), award #1R01EB007684. We thank Maryann Martone, Jessica Turner and Angela Laird for helpful discussions.

References

1. Laird A, Lancaster J and Fox P. BrainMap: The social evolution of a human brain mapping database. Neuroinformatics. 2005; 3:65–78.
2. Bug W, Ascoli G, Grethe J, Gupta A, Fennema-Notestine C, Laird A, Larson S, Rubin D, Shepherd G, Turner J and Martone M. The NIFSTD and BIRNLex vocabularies: Building comprehensive ontologies for neuroscience. Neuroinformatics. 2008; 6:175–194.
3. Frishkoff G, Frank R, Rong J, Dou D, Dien J and Halderman L. A framework to support automated classification and labeling of brain electromagnetic patterns. Comput Intell Neurosci. 2007; 14567.
4. Dou D, Frishkoff G, Rong J, Frank R, Malony A and Tucker D. Development of NeuroElectro-Magnetic Ontologies (NEMO): A Framework for Mining Brainwave Ontologies. Proc KDD. 2007; 270–279.
5. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg L, Eilbeck K, Ireland A, Mungall CJ, *et al*. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007; 25:1251–1255.
6. Dou D, McDermotte V and Qi P. Ontology translation on the semantic web. Journal of Data Semantics. 2005; 2: 35–57.
7. LePendu P, Dou D, Rong J and Frishkoff G. Ontology Database: A New Method for Semantic Modeling and an Application to Brainwave Data. Proc SSDBM. 2008; 313–330.

Generating Homology Relationships by Alignment of Anatomical Ontologies

Frederic B. Bastian[1,2], Gilles Parmentier[1,2], Marc Robinson-Rechavi[1,2]
[1]**Department of Ecology and Evolution, University of Lausanne, Switzerland;**
[2]**Swiss Institute of Bioinformatics, Lausanne, Switzerland**

Abstract

The anatomy of model species is described in ontologies, which are used to standardize the annotations of experimental data, such as gene expression patterns. To compare such data between species, we aim to establish homology relations between ontologies describing different species. We present a new algorithm, and its implementation in the software Homolonto, to create new relationships between anatomical ontologies, based on the homology concept. These relationships and the Homolonto software are available at http://bgee.unil.ch/

Introduction

To be able to compare biological data, we need to use ontologies, to ensure that a biological concept is unambiguously associated to a unique identifier. To achieve this, ontologies such as the Gene Ontology[1] are increasingly used. Websites dedicated to model species also rely on the use of ontologies, for example the zebrafish anatomy for ZFIN[2], or the Mouse gross anatomy and development[3].

We are interested in integrating and comparing gene expression patterns between several species[4]. This raises the question of encoding corresponding information between ontologies which describe different anatomies (e.g. zebrafish and human). The most widely accepted criterion to make such comparisons in biology is homology[5]. Homology is classically defined as the relation between structures which derive from a same ancestral structure, although other definitions are discussed. It should be noted that the exact definition is up to the user, whose input will define which pairs of terms are defined as homologous.

To apply this concept in practice, hundreds of terms must be compared between ontologies. Although a purely manual annotation of homologies is possible, it would be too time consuming to be done for all terms between several divergent species. Kruger et al.[6] used a manual approach to find similarities between simplified anatomy ontologies for human and mouse. There is also an on-going effort to integrate anatomical ontologies, the Common Anatomy Reference Ontology project CARO[7].

Since the problem of aligning anatomical ontologies is to find correspondences between the concepts of two ontologies, we draw on methods from "schema matching", or "ontology alignment"[8,9]. Ontology alignment is the process of determining correspondences between ontology concepts. Usually, this technique is used to find the common concepts present in two ontologies. In the case of anatomical ontologies, the concepts to align are not strictly common, but rather, related: a homology relationship is not an equivalence relationship. For this reason, ontology alignment approaches developed for other applications (e.g. medicine oriented descriptions of human[9, 10]) cannot be applied as such: these methods would be misled by the existence of elements of same names and related to the same concept, but not homologous (e.g. eye of insects and of vertebrates), or reciprocally, homologous elements with different names (e.g. pectoral fin and upper limb).

We present here a new algorithm, and its implementation in the Java software Homolonto, to create new relationships between anatomical ontologies, based on the homology concept. Thus the basic aim of Homolonto is to propose in priority to the user the best candidate pairs of homologs, and avoid the need to consider many irrelevant pairs.

Homolonto Algorithm

1) *Computing word specific scores:* Score modifiers are computed for all words of the ontologies being aligned. Each word present at least once in both ontologies being aligned (O1 and O2) is given a score modifier based on its number of occurrences f(word, O):

$$Mod(word, Oi) = 1/(1+\log10(f(word, Oi))) \quad \textit{eq. 1}$$

$$Mod(word) = Mod(word, O1) * Mod(word, O2) \quad \textit{eq. 2}$$

2) *Starting list of propositions:* To initialize the algorithm we define first obvious similarities between the terms of the ontologies to align. Based on the assumption that two structures that have the same name are likely homologous, the initial propositions are formed of terms with identical names. In this process, we also consider the synonym field of the terms. Each pair of names n1, n2, is given a base score, dependent on the words shared:

$$Base\_score(n1, n2) = base\_homonymy\_score * \max(Mod(word)) * |n1 \cap n2| / \max(|n1|, |n2|) \quad \textit{eq. 3}$$

Where |n| is the number of words in n, |n1∩n2| is the number of words shared by n1 and n2, and max(Mod(word)) is computed over all shared words. In the starting list, |n1∩n2| = |n1| = |n2| by definition, but this is not the case at further iterations of the algorithm.

3) *Initial propagation step:* The score of these propositions is propagated between neighbors. This initial propagation is bidirectional, and limited to already defined propositions. For example, the score of the "optic cup" pair is added to the score of the "eye" pair, as "optic cup" is part of "eye", and both pairs are initial propositions. Symmetrically the score of the "eye" pair is added to the "optic cup" pair. But the score of "eye" is not propagated to e.g. the pairing of "visual system" (ZFA[2] parent of "eye") with "sensory organ" (EHDAA[11, 12] parent of "eye"), because this pair is not an initial proposition. The aim of this step is to increase the score of the most likely homologs.

4) *Cleaning the initial proposition list:* The design of some ontologies may generate many false positives, typically through repetition of the same name as a child of diverse structures (e.g. 76 occurrences of "mesenchyme" in EHDAA). To avoid this, if a term is a member of several propositions with different scores, we initially keep only the best scoring proposition. If there are more than 5 highest scoring propositions for a given term, we remove all propositions for this term.

5) *Evaluation step:* Each proposition is presented to the user, in descending order of scores. The user has to validate, invalidate, or delay decision regarding the proposed homology.

6) *Computation step:* If one of the terms of a validated pair is already a member of an homology group, then the other term is added to the homology group. Otherwise, a new homology group is created, containing both terms of the validated pair. The information of homology is propagated through the hierarchy by the use of a validated homology score (eq. 4). The underlying idea is that if two terms A and B are homologous, then one of the children of A is probably homologous to one of the children of B. During the propagation the validated homology score is added to the base score (eq. 3) of pairs of terms:

Propagated_score(a, b) = validated_homology_score * (max_depth + 1 - present_depth) / (max_depth + 1) *eq. 4*

Total_score(a, b) = Propagated_score(a, b) + Base_score(n_a, n_b) *eq. 5a*

Where n_a is the name of term a. In the present implementation, the max_depth is 1, and the validated homology score is 1.5 times the base homonymy score. For pairs of terms which are not yet a proposition, a new proposition is created, and the base score is computed. This will include cases of partial homonymy, for which eq. 3 down weights names which share a lower proportion of words. Pairs which have been previously invalidated by the user will not receive a propagated score, and will remain invalidated.

To down weight potential false positives due to validation of terms with many children, the propagated score is reduced proportionally to the number of new propositions for each term of the ontology to align (eq. 5b).

Total_score(a, b_i) = Propagated_score(a, b_i)/ ((|b| + 1) * 2) + Base_score(n_a, n_{bi}) *eq. 5b*

Where a is a term of the ontology to align, b_i is a term of the reference ontology, and |b| is the number of new propositions for term a. When a proposition (a, b_i) is invalidated, |b| is updated, and the Total score(a, b_i) increases for the remaining propositions.

When the terms of an invalidated proposition share common words, then the score modifiers of all shared words is diminished (eq. 6). As this is repeated, words which tend to generate false positives will be increasingly down weighted.

Mod'(word) = Mod(word) * 0.9 *eq. 6*

7) *Iteration:* Evaluation of propositions (step 5), ordered by total score (base score + propagated score), and computation (step 6), is repeated until the user decides to terminate, or no more propositions are generated.

Homolonto Results

Homolonto has been used to align six anatomical ontologies to date, representing four vertebrate species (human and mouse have different ontologies for adult and embryonic stages). We will present more in detail two alignments: zebrafish (ZFA ontology[2]) / Xenopus (XAO ontology[13]), which illustrates a best case scenario of two recently updated ontologies, conforming to the CARO standards[7], with annotations of synonyms and definitions, and low redundancy. And human (EHDAA ontology[11, 12]) / mouse (EMAPA ontology[11, 3]) which, despite the similarity in anatomy, illustrates a more difficult scenario of large ontologies, with issues such as repetition of names (76 occurrences of "mesenchyme" in human, 93 in mouse), due to splitting of concepts among morphological structures or among developmental stages.

The main observation is that our algorithm is successful at ordering propositions. In the "easy" case of zebrafish / Xenopus, there are only seven invalidated propositions in the first 150 (95% validation). This is followed by a relatively short interval of iterations where validated and invalidated propositions are mixed: 46% of validations between iterations 151 and 200, and 20% between 201 and 250. Further iterations generate mostly invalidated propositions (3% validation from 251 to 735). Thus 93% of all validations occurred in the first 250 iterations.

The pattern is similar for the human / mouse alignment. In the first 1400 iterations, 99% of propositions are validated. In the next 600 iterations, the figure reduces to 63%, and in the last 962 iterations it falls to 21%. This slower decrease illustrates the complexity of this alignment. The validation rate of 66% shows that the propositions were mostly worth considering, and that the high number of propositions was due indeed to the size of the ontologies, not to a default in the algorithm. Results also show that manual expertise is necessary, since even in the high scoring propositions some are invalid. Overall, 27% of invalidations are pairs of terms with identical names. Interestingly, Homolonto manages to give these misleading homonyms low priority: homonyms within the first 1000 iterations have a 99% chance of being homologs, whereas homonyms within the last 1000 iterations only have a 19% chance of being homologs. Thus 93% of invalidated homonyms appear after iteration 1400.

Generating Relationships between Groups of Homologs

Homolonto is used to generate pairwise homology relationships between anatomical ontologies. As homology relationships are transitive, these pairwise alignments can be merged into homologous organs groups (HOGs). Homolonto thus generates HOGs, and mapping of species-specific anatomical structures to these HOGs. HOGs then need to be structured as an ontology to allow reasoning on them. This means that, at a minimum, relationships amongst them have to be designed. Another algorithm has thus been developed to infer relationships between HOGs.

1) *Initial Step:* all possible paths between HOGs are retrieved. For instance, if an anatomical structure "a", mapped to the HOG "A", has a *part_of* relationship to the anatomical structure "b", mapped to the HOG "B", then a putative *part_of* relationship is defined between HOGs "A" and "B".

Relationships between HOGs are often indirect (e.g. structure "a", mapped to HOG "A", *part_of* structure "c", *part_of* structure "b", mapped to HOG "B"). If the first relation (the relation "outgoing" from the child HOG, "A" in the previous example) and the last relation (the relation "incoming" to the parent HOG, "B" in the previous example) are of the same type (e.g. *part_of*, *is_a*), then the putative relationship is defined as this type. Otherwise, the relationship is defined as the SKOS[14] type *broader_than*.

2) *Skipping relations from not-trusted ontologies:* some ontologies do not follow the OBO principles, and implement for instance only one type of relation amongst all concepts (e.g. EV[15] only uses *is_a* relationships). The user may choose to not use these ontologies to define relation types. All the putative relations inferred by these ontologies at step 1 are then set as *broader_than*. But the final relation type between these HOGs can still be inferred thanks to other ontologies.

3) *Skipping relations defined by too few species:* if the proportion of species defining a relation, compared to the total number of species involved in the creation of the HOGs, is below a threshold defined by the user ("species coverage"), then the relation is defined to the type *broader_than*, and the algorithm stops examining relations between these HOGs. Indeed, in such case, inferred relation types may not be trusted.

4) *Defining within-ontology agreement*: several anatomical structures from the same ontology can belong to the same HOG. This can generate a within-ontology conflict for defining a relation type. For instance, structures "a" and "b" allow to define a putative *part_of* relationship between HOGs "A" and "B", while structures "a'" and "b'", belonging to the same ontology, define a putative *is_a* relationship between these HOGs. The algorithm then calculates, for each relation type, the proportion that the number of paths defining this relation type represents, compared to the total number of paths between these two HOGs for this ontology. If, for a type, this proportion exceeds a threshold ("within-ontology agreement"), defined by the user and at least greater than 0.5, then this relation type is attributed for this species between these HOGs. Otherwise, the relation is defined to the type *broader_than* for this ontology.

5) *Defining inter-ontology agreement:* different ontologies can define different relation types between two related HOGs. This conflict is resolved in the same way as at step 4, by using a threshold ("inter-ontology agreement"), defined by the user and at least greater than 0.5.

6) *Removing cyclic relationships:* by inferring automatically the relationships between HOGs, cycles

may be generated (e.g. HOG "A" *part_of* HOG "B" *part_of* HOG "A"), whereas the ontology has to be acyclic. If such cycles are detected, the algorithm stops with an error message prompting the user to make a decision: the user has then to manually remove one of the involved relationships.

7) *Removing redundancies:* if several relationships are redundant, only the deepest relationship is conserved; for instance, if a HOG "A" has two substructures by a *part_of* relationship, "B" and "C", and if "C" is also a substructure of "B", then the direct relationship between the HOGs "A" and "C" is removed.

8) *Curation step:* a curator has then to manually review all the *broader_than* relations, to attribute them to a type defined by the OBO Relation Ontology[16]. Some custom relationships, not inferred by the algorithm, can also be added at this step.

Conclusion

To date, the use of Homolonto, followed by a curation process, allowed to define 1004 HOGs, involving 4088 structures from 6 anatomical ontologies (ZFA[2], EHDAA[11, 12], EV[15], EMAPA[11, 12], MA[17], and XAO[13]).

The algorithm to design relationships amongst the HOGs inferred 1188 relations. With the more stringent parameters (species coverage = 1, within-ontology agreement = 1, inter-ontology agreement = 1), 341 of them are defined as *part_of*, all the others as *broader_than*. The curation step to review these *broader_than* relations is currently under process.

The HOG ontology has been successfully implemented into Bgee[4], a database for studying gene expression evolution, and already allows to perform automated, cross-species, gene expression pattern comparisons.

The Homolonto software and source code, and the HOG ontology, are available from the download section of the Bgee website (http://bgee.unil.ch). The algorithm to generate relationships between groups of homologs will be available soon.

References

1. Gene Ontology Consortium. The Gene Ontology (GO) project in 2006. Nucleic Acids Res 2006; 34: D322–6.
2. Sprague J, *et al*. The Zebrafish Information Network: The zebrafish model organism database. Nucleic acids research 2006; 34: D581–5.
3. Baldock RA, *et al*. EMAP and EMAGE: A framework for understanding spatially organized data. Neuroinformatics 2003; 1: 309–25.
4. Bastian F, *et al*. Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species. LNBI Springer 2008; 124–31.
5. Hall B. Homology: The Hierarchical Basis of Comparative Biology. Academic Press 1994.
6. Kruger A, *et al*. Simplified ontologies allowing comparison of developmental mammalian gene expression. Genome Biol 2007; 8: R229.
7. Haendel MA, *et al*. CARO – The Common Anatomy Reference Ontology. Springer 2008; 327–49.
8. Euzenat J and Shvaiko P. Ontology Matching. Springer Verlag 2007.
9. Lambrix P and He T. Ontology alignment and merging. Springer 2008; 133–49.
10. Mork P and Bernstein PA. Adapting a Generic Match Algorithm to Align Ontologies of Human Anatomy. IEEE 2004.
11. Aitken S. Formalizing concepts of species, sex and developmental stage in anatomical ontologies. Bioinformatics 2005; 21: 2773–79.
12. Hunter A, *et al*. An ontology of human developmental anatomy. Journal of anatomy 2003; 203: 347–55.
13. Bowes JB, *et al*. Xenbase: A Xenopus biology and genomics resource. Nucleic Acids Res 2008; 36: D761–7.
14. Miles A and Brickley D. Simple Knowledge Organisation System (SKOS). Aug 2008; http://www.w3.org/TR/2008/WD-skos-reference-20080829/.
15. Kelso J, *et al*. eVOC: A controlled vocabulary for unifying gene expression data. Genome Res 2003; 13: 1222–30.
16. Smith B, *et al*. Relations in biomedical ontologies. Genome Biol 2005; 6: R46.
17. Smith CM, *et al*. The Mouse Gene Expression Database (GXD): 2007 update. Nucleic Acids Res 2007; 35: D618–23.

Towards Desiderata for an Ontology of Diseases for the Annotation of Biological Datasets

Olivier Bodenreider[1], Anita Burgun[2]
[1] LHNCBC, National Library of Medicine, Bethesda, MD, USA
[2] U936, IFR 140, Faculté de Médecine, University of Rennes 1, France

Abstract

There is a plethora of disease ontologies available, all potentially useful for the annotation of biological datasets. We define seven desirable features for such ontologies and examine whether or not these features are supported by eleven disease ontologies. The four ontologies most closely aligned with our desiderata are Disease Ontology, SNOMED CT, NCI thesaurus and UMLS.

Introduction

Ontologies have been developed for the annotation of biological datasets from multiple perspectives including functional annotation of gene products (Gene Ontology), molecular sequences (Sequence ontology) and phenotypes (Mammalian Phenotype Ontology, Phenotypic Quality Ontology). Entries in biological datasets also need to be linked to diseases, either human diseases or experimental models of diseases in model organisms. Ontologies of diseases include the Disease Ontology (DO), from the Open Biomedical Ontology (OBO) family. The NCI Thesaurus was developed for the annotation of cancer research and includes many diseases, but its focus on cancer can be a limitation for use in other domains.

On the other hand, terminologies have been long been developed for the purpose of annotating clinical records, including the International Classification of Diseases (ICD) and SNOMED CT. However, these terminologies have not been widely adopted by biomedical researchers for annotating disease entities in biological datasets. Moreover, neither terminology is free of intellectual property restrictions and a license or fee may be required for their use, which represents a limiting factor.

Finally, terminology integration resources such the Unified Medical Language System (UMLS) Metathesaurus and NCBO's BioPortal both integrate more than one hundred biomedical terminologies, including all those mentioned above. Moreover, both resources provide mappings across terminologies, facilitating the integration of biological and clinical data required for translational medicine. However, their use may necessitate significant training.

The objective of this study is to propose a list of desirable features for an ontology of diseases suitable for the annotation of biological datasets, and to analyze a list of candidate terminologies through the framework provided by these features.

Desiderata for selecting ontologies have been established by the OBO Foundry[1]. While interesting and potentially relevant to the domain of diseases, we find some of these criteria unnecessarily restrictive for the purpose of annotating biological datasets, while key criteria (from our specific perspective) are missing. A brief analysis of the OBO Foundry criteria in the context of our study is proposed in the discussion.

This work also differs from Cimino's desiderata for controlled medical vocabularies[2] in that we focus on content and usability for a particular purpose in addition to representation issues and development process.

Methods

We first select a list of biomedical terminologies and ontologies (hereafter referred to simply as ontologies) potentially suitable for the annotation of diseases in biological datasets. We establish a list of characteristics from these ontologies, focusing on those characteristics which represent potential barriers to adoption of these terminologies by biomedical researchers. We apply the list of features to each candidate ontology and summarize our findings in a feature x ontology matrix.

Candidate Ontologies

In order to identify candidate ontologies for diseases, we explored the two major repositories of biomedical ontologies: The Unified Medical Language System (UMLS) and NCBO's BioPortal. We investigated ontologies whose focus is on human diseases and phenotypes, as well as ontologies which contain a significant number of disease entities. In practice, we exploited the metadata provided with OBO ontologies and selected those ontologies for which the domain is contains "phenotype" or "health". No similar mechanism is available for the UMLS and we simply used our knowledge of the source vocabularies to make our selection. References to the ontologies

discussed below are listed in Table 1 (appendix). This selection process led to the identification of eleven ontologies potentially suitable for the annotation of diseases in biological datasets.

- **Disease Ontology** (DO): Controlled terminology from the OBO family created for annotation purposes as part of the NuGene project at Northwestern University. Coverage restricted to diseases.
- **Online Mandelian Inheritance in Man** (OMIM): Knowledge base on human genetic diseases developed at John Hopkins University and available through the NCBI Entrez system. Its terminological component – including clinical synopses – is available through the UMLS. Coverage restricted to genetic diseases.
- **International Classification of Diseases** (ICD): Classification from the World Health Organization (WHO) family of health classifications, with many local adaptations. ICD9-CM, developed by the Center for Medicare & Medicaid Services (CMS) for use in the US, includes clinical modifications. Coverage restricted to diseases and health problems.
- **SNOMED CT**: The largest clinical terminology developed by the International Health Terminology Standard Development Organization (IHTSDO) for use in electronic health records and adopted by eleven countries to date. Broad coverage including diseases.
- **Medical Subject Headings** (MeSH): Controlled vocabulary developed by the U.S. National Library of Medicine for the indexing and retrieval of the biomedical literature, especially in the MEDLINE bibliographic database. Broad coverage including diseases.
- **NCI Thesaurus** (NCIt): Controlled vocabulary developed by the National Cancer Institute to support the integration of information related to cancer research. Broad coverage including diseases.
- **Unified Medical Language System** (UMLS): Terminology integration system developed by the U.S. National Library of Medicine, establishing a correspondence among terms from different terminologies for a given biomedical entity. Broad coverage including diseases.
- **Human Phenotype Ontology** (HPO): Controlled vocabulary for the phenotypic features encountered in human hereditary and other diseases. Developed by a consortium including Charite Hospital (Berlin) and the University of Cambridge (UK). Coverage restricted to monogenic diseases listed in OMIM.
- **Phenotypic Quality Ontology** (PATO): Ontology of phenotypic qualities, intended for use in a number of applications, primarily defining composite phenotypes and phenotype annotation. Coverage restricted to phenotypes.
- **Mammalian Phenotype Ontology**: Controlled vocabulary for the "robust annotation of mammalian phenotypes" currently used for the annotation of phenotypic data in mouse and rat databases. Developed at the Jackson Laboratory. Coverage restricted to phenotypes.
- **Logical Observation Identifiers Names and Codes** (LOINC): Set of names and codes for laboratory and other clinical observations (elements of clinical phenotypes). Developed at the Regenstrief Institute. Coverage restricted to clinical observations.

Phenotype ontologies for organisms other than *Homo sapiens* were ignored. (e.g., **Yeast phenotypes**). Ontologies of diseases included as part of a broader ontology were ignored when they were unlikely to provide additional coverage or characteristics useful for the discussion in this paper (e.g., **National Drug File Reference Terminology** and **International Classification of Primary Care**). Specialized resources (e.g., **Online Congenital Multiple Anomaly/Mental Retardation Syndromes**, **Infectious Disease Ontology** and **Diagnostic and Statistical Manual of Mental Disorders**), while providing deep coverage of a narrow subdomain of medicine, are unlikely to provide the broad coverage expected from an ontology of diseases and were ignored.

Desirable Features

Starting from the ten OBO Foundry principles, we have identified seven desirable features for an ontology of diseases. In each case, the absence of a feature represents a potential barrier to the adoption of a biomedical ontology for the annotation of diseases in biological datasets. Differences with the set of OBO Foundry principles are discussed later in this paper.

- **No Intellectual Property Restrictions**. The use of some vocabularies is limited to certain contexts (e.g., restriction for research purposes vs. production systems for some vocabularies in the UMLS) or to certain countries (e.g., member countries of the IHTSDO for SNOMED CT), or

subject to the payment of a fee (e.g., ICD 10). This feature is aligned with Foundry principle #1.

- **Standard, Friendly Format**. Availability of terminologies in formats that are standard (e.g., RDF, OWL) or friendly to biologists (e.g., OBO) is likely to foster adoption. In contrast, proprietary formats (e.g., RRF for the UMLS Metathesaurus) may represent a barrier to adoption. This feature corresponds roughly to Foundry principle #2.
- **Existence of a Mapping to Clinical Terminologies**. In the era of translational medicine, biological datasets must be linkable to clinical datasets. The existence of mappings between an ontology of diseases used for the annotation of biological datasets and clinical terminologies used in patient records is strong requirement.
- **Harmonization with Other Biological Ontologies**. Similarly to the requirement for integration with clinical terminologies, there is a need for a disease ontology to be integrated – if possible natively – with other biological ontologies. This feature corresponds roughly to Foundry principle #5.
- **Regular Maintenance**. The domain of diseases is in constant evolution and an ontology of disease shall reflect emerging diseases and changes in our understanding of the domain of diseases. This feature corresponds roughly to Foundry principle #4.
- **Exhaustive Coverage of Diseases**. At a given level of granularity, the ontology shall provide an exhaustive coverage of the domain. Terminologies focusing on a specific subdomain may have limited applicability outside this subdomain (e.g., focus on cancer in NCIt).
- **Support for Automatic Reasoning**. Annotations made to ontologies often form the basis for gaining new knowledge about biomedical entities. In order to process annotations efficiently and automatically, ontologies need to have a robust, formal structure and provide support for automated reasoning (e.g., through subsumption).

A Framework for Comparing Disease Ontologies

The desirable features listed above do not all have the same importance from the perspective of an ontology of diseases for annotation purposes. For example, coverage of diseases it of the outmost importance for an ontology of diseases and was given the highest weight (5). Interoperability with other ontologies (clinical and biological) and support for automatic reasoning correspond to major uses of ontologies and are also weighted more (2) than the remaining features (1).

We examined the eleven candidate ontologies through the prism of the seven desirable features. More precisely, for each feature, we rated the ontology semi-quantitatively: 0 (no or minimal support for the feature), 0.5 (partial support of the feature) or 1 (reasonable support for the feature), assessed by the authors. The weights were applied to the ratings. Finally, the score of each ontology was computed by comparing the sum of the scores for each feature to the sum of all weights (14).

Results

The result of assessing the presence of the desirable features in the candidate ontologies is summarized in Figure 1 (appendix). Support for the desirable features ranges from 32% (OMIM, LOINC) to 68% (NCIt, UMLS). Seven ontologies have a score of 50% or more.

Discussion

Applying the Desiderata. The top four contenders identified in our matrix of desirable features x ontologies (Figure 1) are Disease Ontology, SNOMED CT, NCIt and UMLS. Interestingly, these four ontologies made it to the top for slightly different reasons. Depending on what features are most important in a given use case, the ontologies corresponding to this profile of features should be selected.

Phenotypes vs. Diseases. Precisely defining phenotype and disease is beyond the scope of this paper. However, we observed that phenotype ontologies containing pre-coordinated concepts (e.g., **Mammalian Phenotype Ontology**, **LOINC**) or supporting post-coordination (e.g., the **Phenotypic Quality Ontology – PATO**), cover low-level phenotypes and clinical observations (e.g., individual anatomical and physiological abnormalities) rather than diseases. Examples of phenotypes form MPO include *enlarged liver*, found in ontologies including MeSH, NCTt and SNOMED CT. In contrast, they mostly contain terms indicating deviation from normal anatomical structures or physiologic states (e.g., *decreased liver weight*), typically absent from the clinically-oriented disease ontologies. Phenotype ontologies seem suitable for the annotation of data with low-level phenotypes, whereas disease ontologies have application in the annotation of higher-order information about diseases, i.e., resulting from some elaborate diagnostic process.

Differences with OBO Foundry Criteria. Although some of our desirable features are aligned with

principles of the OBO Foundry[1], we found the Foundry principles to be generally too rigid for the purpose of annotating biological datasets and lacking consideration for legacy ontologies. Applying these principles strictly to the selection of ontologies would potentially result in unnecessarily excluding from consideration the datasets annotated to these legacy ontologies.

Many legacy disease ontologies are not available in OWL or OBO format, but are widely used. Borrowing from "orthogonal" ontologies is a good principle for the coordinated development of ontologies (i.e., applied in a prospective manner). However, this principle can hardly be held against legacy disease ontologies. The absence of textual definition is a common feature to many legacy disease ontologies. It can be offset in part by the presence of formal definitions (in description logic-based systems) and usage information. Finally, most widely used disease ontologies are developed outside the OBO Foundry and not always in a collaborative manner.

Limitations. The framework provided here for analyzing disease ontologies is relatively coarse and somewhat arbitrary. The list of desirable features and the weights would need to be adapted to specific annotation scenarios. For example, the presence of synonyms is required if annotations are to be discovered automatically in text corpora using text mining techniques.

Conclusions

The plethora of disease ontologies available to biomedical researchers for annotation purposes is not necessarily good news. In this domain in particular, reusing existing ontologies should be carefully considered before starting the development of a new one. Annotations made to different ontologies, including legacy ontologies, will likely need to be reconciled in order to enable interoperability among datasets, which is a strong requirement for translational medicine. Terminology integration systems such as the UMLS are thus expected to play a key role in data integration tasks.

Acknowledgements

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM).

References

1. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, *et al*. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 2007;25(11):1251–5.
2. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998;37(4-5):394–403.

DO	*Disease Ontology* - http://diseaseontology.sourceforge.net/
OMIM	*Online Mandelian Inheritance in Man* - http://www.ncbi.nlm.nih.gov/omim/
ICD	*International Classification of Diseases* - http://www.who.int/classifications/icd/en/
SNOMED CT	*SNOMED CT* - http://www.ihtsdo.org/
MeSH	*Medical Subject Headings* - http://www.nlm.nih.gov/mesh/
NCI Thes.	*NCI Thesaurus* - http://cancer.gov/cancerinfo/terminologyresources/
UMLS	*Unified Medical Language System* - http://www.nlm.nih.gov/research/umls/
HPO	*Human Phenotype Ontology* - http://www.human-phenotype-ontology.org/index.php/hpo_home.html
PATO	*Phenotypic Quality Ontology* - http://www.bioontology.org/wiki/index.php/PATO:Main_Page
MPO	*Dictionary of Medicines and Devices* - http://www.informatics.jax.org/searches/MP_form.shtml
LOINC	*Logical Observation Identifiers Names and Codes* - http://loinc.org

Table 1. List of potential disease ontologies discussed in this paper

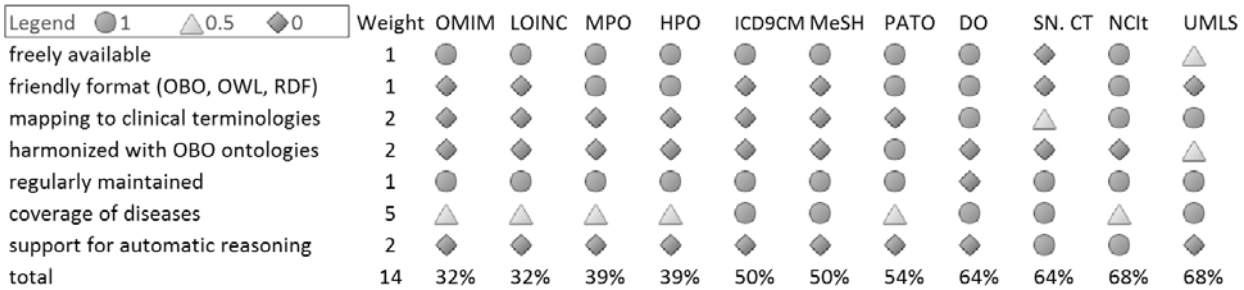


Figure 1. Desiderata applied to candidate disease ontologies (matrix of desirable features x ontologies)

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

A Set of Ontologies to Drive Tools for the Control of Vector-Borne Diseases

Pantelis Topalis[1], Emmanuel Dialynas[1], Elvira Mitraka[2], Elena Deliyanni[1], Inga Siden-Kiamos[1], Christos Louis[1,2]

[1]Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology-Hellas, Heraklion, Crete, Greece; [2] Department of Biology, University of Crete, Heraklion, Crete, Greece

Abstract

We are developing a set of ontologies dealing with vector-borne diseases and the arthropod vectors that transmit them. For practical reasons (application priorities), we initiated this project with an ontology of insecticide resistance followed by a series of ontologies that describe malaria as well as physiological processes of mosquitoes that are relevant to, and involved in, disease transmission. These will be expanded to encompass other vector-borne diseases as well as non-mosquito vectors. The aim of the whole undertaking, which is worked out in the frame of the international IDO (Infectious Disease Ontology) project, is to provide the community with a set of ontological tools that can be used both in the development of specific databases and, most importantly, in the construction of decision support systems (DSS) to control these diseases.

The Problem of Vector-Borne Diseases

Epidemiologists have brought together in one "functional" group a series of diseases of different etiology and pathogenesis that share one key component: their mode of transmission (see Goddard, 1999[1] and several chapters of Marquardt, 2005[2] for specific questions addressing insect-borne diseases and their vectors). These diseases are transmitted by the bite of a specific arthropod vector, which is usually an insect. The pathogenic agent is passed with the saliva transferred during the bite to the potential patient. Two additional characteristics are shared by most vector-borne diseases, namely most people affected live in the tropical regions of the world and, connected to this, the diseases affect mostly populations that are also heavily affected by poverty. The pathogens responsible for these diseases are very diverse, ranging from protozoan parasites (e.g. *Plasmodium* spp. in malaria, *Leishmania* spp. in leishmaniosis) and bacteria (e.g. *Borrelia* spp. in Lyme disease), to worms (e.g. Nematodes in filariasis and river blindness) and to viruses (e.g. Dengue, Yellow fever). Similarly, the vectors are also very diverse and range from mosquitoes (e.g. malaria and Dengue) and flies (e.g. Tsetse in African trypanosomiasis) to kissing bugs (Chagas' disease) and ticks (e.g. Lyme disease). The extreme variation in the biology of both pathogens and the vectors makes it difficult to address vector-borne diseases as a whole. Importantly, these difficulties also affect significant aspects such as prevention, epidemiology, therapy, etc.

A common theme, which in a sense unites these diseases, is the fact that their transmission can be blocked if the agents that transmit them, i.e. the arthropod vectors, are removed from the pertinent chain of events[3]. Vector control has therefore historically become a *conditio sine qua non* for the control of these infections[4,5], and this fact has been exemplified by the elimination of malaria from most non-tropical areas of the globe[6]. While leading to about half a billion cases in the tropics every year, and still being responsible for anything between one and three million deaths (mostly children in sub-Saharan Africa), this killer illness has practically disappeared from Europe and North America through intense insecticidal measures aimed at eliminating the *Anopheline* vectors. It should be stressed that, with the exception of the Yellow fever[7], no vaccine is currently available for any vector-borne disease as an alternative prevention strategy that would act on a different level than that of the actual vector. Prevention focused on the vector includes not only control of insect populations through environmental management or the use of chemicals, but also the protection of individuals through the use of clothing, repellents, nets and screens[8].

Although greatly successful in the previous century, insect-control programmes are now immensely obstructed by a variety of reasons. These range from community opposition to a vast usage of chemicals[9], to the development of resistance against these very chemicals by the insect vectors to be controlled[10]. Moreover, these problems are aggravated by several facts: resistance against drugs is also encountered in the pathogens[11]; vaccine development, if at all possible, is slow[12]; new drug development is not only slow but extremely expensive and the areas affected by the diseases in question are certainly not the ones that can easily spearhead such efforts due to the lack of economic and scientific resources in them[13]. It is therefore of utmost importance to develop innovative strategies for the control of vector-borne diseases. One novel approach is to use IT technologies as a complement

to the application of developments in the biology of disease vectors. While the latter projects make use of scientific research products such as whole genome sequences[14,15], transgenesis[16], and the use of other "intelligent" approaches[17] the former potentially brings new specific tools that can be used for a more efficient, and often close-to-the-field management of pertinent disease data, including entomological ones.

In this context, our group has embarked on a long project that involves the development of ontologies dealing with disease vectors and vector-borne diseases[18,19]. The obvious rationale behind this is the need of these ontologies to unify the "language" spoken by vector biologists and epidemiologists. The ultimate end is to build a comprehensive ontology for insect-borne diseases that may consist of sub-ontologies, each addressing a specific aspect of the whole. In the frame of the Infectious Disease Ontology project (http://www.infectiousdisease ontology.org/Home.html), we initiated this effort focusing on malaria, but we are already expanding this to encompass most other vector-borne diseases as well. These ontologies, some of which are already available and some under development, will be presented below in a summary form.

Ontologies and Vector-Borne Diseases: A Brief Description

The aspects of vector-borne diseases that are in need of an ontological description range from those that deal with the diseases as such (e.g. pathogenesis, clinical aspects, therapy, etc.), to vector biology (physiological processes of the vectors) and to epidemiology and control in the widest sense of the terms (prevention, insect control, etc.). As stated earlier, these aspects are extremely diverse and complex, simply given the multitude of organisms involved (vectors and pathogens in addition to the human host) and the fact that we are often dealing with populations (additional level of granularity!). The construction of a comprehensive ontology, thus, if at all feasible, must be addressed using a piecemeal approach. It is clear that certain fundamental decisions have to be taken at the initial phases, and an open-ended advance is, in our mind, a must. We therefore decided, early on, that the end product would have to follow i) the rules set by the OBO Foundry[20] and ii) be based on the basic formal ontology[21,22]. If long-term interoperability of future databases is to be achieved, these two choices are a prerequisite. This rule, of course, is the end goal and we decided to keep a certain degree of flexibility throughout the project until a "unified" ontology is constructed. One example for such a flexible approach is the fact that the ontology of insecticide resistance in mosquitoes that we developed (MIRO) does not follow the BFO in its initial versions but, rather, it is structured such that it can be adopted without many problems by the community that immediately needs to apply it in the field[23]. The MIRO forms the core of the related database on insecticide resistance (IRbase) that we also developed, and which was adopted for immediate use by the World Health Organization (Dialynas et al., 2009, in preparation). We should state that we are nevertheless in the process of long term restructuring the ontology along BFO standards, such that its contents can be later included in the comprehensive ontology on vector-borne diseases.

Although already submitted to and listed by the OBO Foundry, MIRO is a pure application ontology that is being used to drive a dedicated database, IRbase (http://anobase.vectorbase.org/ir/). It consists of four specially devised sub-ontologies that cover all aspects of insecticide resistance, with an emphasis on field work and monitoring. Thus, although mechanisms of resistance are covered, this is not done in detail. Furthermore, MIRO's fifth component, a geographical one, uses *in toto* the controlled vocabulary Gazeteer to provide IRbase curators with records describing the areas in which data were collected. The MIRO is constantly being updated upon request by members of the international community that is involved in the study of insecticide resistance.

The second ontology, which is still nameless, covers physiological processes of mosquitoes that are involved in disease transmission. The processes covered do not only address the actual transmission, i.e. the interplay between vectors and pathogens but, importantly, also the actual progression of events in the vector. We want to stress that the processes mentioned here are, in their vast majority, processes on the level of the organism and not cellular or sub-cellular ones, such as the ones covered by the GO[24, 25]. Thus, (near) top level classes are, among others, behaviour, sensory perception, processes of the immune system and nutrition. As an example, when looking at the children of "behaviour", one will find a line of terms leading through the adult feeding behaviour to entities such as the four phases of "interrupted feeding" (exploratory phase, imbibing phase, probing phase and withdrawal phase). The ontology also covers processes that are not directly "linked" to disease transmission and this, obviously, for reasons of completion. For reasons of orthogonality, in all cases in which terms are already covered by established ontologies we adhere to these, along with their descendants. This is notably the case

for the Processes sub-ontology of the GO. Our ontology is far from complete, although it already covers more than 600 terms, which are all fully defined.

The next ontology that we are in the process of populating with terms is the one describing malaria. This is the actual ontology that we decided to develop in the frame of IDO, and which we plan to expand in the near future in order to cover other vector-borne diseases as well. It is built based on BFO and the IDO reference ontology and it is meant to cover malaria on all possible levels. These obviously include both the clinical aspects of the disease in the widest sense (i.e. including epidemiology, etc.) and the biology of the disease that describes processes and objects of not immediate clinical relevance. We consider as such items (e.g. proteins) involved in the penetration of both mosquito and human/vertebrate cells as well as their interacting partners in the *Plasmodium* parasites. Again, similarly to the case of the ontology of physiological processes, we have taken care to include, wherever possible, direct imports of pre-existing ontologies. One such example is the *Plasmodium* parasite life cycle stage and its descendants that all have cross-references to the, at the moment, inactive *Plasmodium* life cycle ontology. The malaria ontology has at this time about 600 terms.

Concluding Remarks

The ontologies that we are constructing can be described as pure application ontologies that are meant to form the basis for specific tools such as specific databases or decision support systems for various diseases. The need for such tools became apparent immediately after the first working version of the MIRO and its “cousin” IRbase were made public. Not only did the international community immediately decide to adopt both tools, but also already within a few months after the initiation of data population, there are about 1350 sampled populations that are shown in the database. This is about 1250 more than what the insecticide resistance section in VectorBase carried, the only repository for data of this kind. In addition to databases that are driven by ontologies in an increasing fashion (see for example databases using the ontology-depending schema Chado[26], such as FlyBase[27,28] and VectorBase[29,30], ontologies are ideal tools for the design of intelligent DSS. In cases such as vector-borne diseases, whose control is also hampered by weak infrastructure in endemic countries, these DSS can be used by medical workers and health agencies in remote areas, either for ongoing studies or in cases that need immediate attention[31,32].

One of the intricacies that we are already faced with is the planned expansion of the malaria-oriented ontologies, to cover many other vector-borne diseases. To understand the magnitude of the problem one should think of the fact that vector-borne diseases represent major threats to public health in wide and ecologically diverse areas of the world, that they are caused by completely different pathogens and that they are transmitted by completely different vectors. Thus, the challenge now is how to cover this broad spectrum of facts in a single ontology. There is naturally the possibility of cutting through the Gordian knot, by devising separate ontologies for each disease. The counter-argument in this case would be that, brought to an extreme, each malaria form (i.e. tertian, malignant and benign, and quartan, should have its own ontology) similar to the different forms of filariasis that are caused by different species of nematodes and whose clinical aspect differ only slightly. In addition, similarities between these diseases and the agents that transmit them may be obscured if different ontologies were used, and this would certainly have a negative impact on their value in the long term. Therefore, we are still trying to solve the knot in a non-Alexandrian way. By “merging” the ontologies into one, we can also actively support the rules of the OBO Foundry and provide an example of how the construction of a large and comprehensive ontology can, later on, provide advantages to its users.

Acknowledgements

The work was funded by contract HHSN 266200400039C from the National Institute of Allergy and Infectious Diseases in the frame of the VectorBase project and by the BioMalPar European Network of Excellence supported by a European grant (LSHP-CT-2004-503578) from the Priority 1 "Life Sciences Genomics and Biotechnology for Health" in the 6th Framework Programme. The authors would numerous colleagues who helped at different stages of the work, and in particular Drs. John Vontas for his contribution to the MIRO, Frank Collins for his encouragement and support in the frame of VectorBase and Barry Smith and Lindsay Cowell for accepting us in the IDO community.

References

1. Goddard J. Infectious diseases and arthropods. Totowa, N.J.: Humana Press; 2000.
2. Marquardt WC and Kondratieff BC. Biology of disease vectors. 2nd ed. Burlington, MA: Elsevier Academic Press; 2005.

3. Hemingway J, Beaty BJ, *et al*. The Innovative Vector Control Consortium: improved control of mosquito-borne diseases. Trends Parasitol 2006;22: 308–12.
4. della Torre A, Arca B, *et al*. The role of research in molecular entomology in the fight against malaria vectors. Parassitologia 2008;50: 137–40.
5. Peter RJ, Van den Bossche P, *et al*. Tick, fly, and mosquito control--lessons from the past, solutions for the future. Vet Parasitol 2005;132: 205–15.
6. de Zulueta J. The end of malaria in Europe: An eradication of the disease by control measures. Parassitologia 1998;40: 245–6.
7. Roukens AH and Visser LG. Yellow fever vaccine: Past, present and future. Expert Opin Biol Ther 2008;8: 1787–95.
8. Hill J, Lines J and Rowland M. Insecticide-treated nets. Adv Parasitol 2006;61: 77–128.
9. Schapira A. DDT: a polluted debate in malaria control. Lancet 2006;368: 2111–3.
10. Hemingway J and Ranson H. Insecticide resistance in insect vectors of human disease. Annu Rev Entomol 2000;45: 371–91.
11. Laufer MK. Monitoring antimalarial drug efficacy: Current challenges. Curr Infect Dis Rep 2009;11: 59–65.
12. Langhorne J, Ndungu FM, *et al*. Immunity to malaria: More questions than answers. Nat Immunol 2008;9: 725–32.
13. Craft JC. Challenges facing drug development for malaria. Curr Opin Microbiol 2008;11: 428–33.
14. Holt RA, Subramanian GM, *et al*. The genome sequence of the malaria mosquito Anopheles gambiae. Science 2002;298: 129–49.
15. Nene V, Wortman JR, *et al*. Genome sequence of Aedes aegypti, a major arbovirus vector. Science 2007;316: 1718–23.
16. James AA. Preventing the spread of malaria and dengue fever using genetically modified mosquitoes. J Vis Exp 2007: 231.
17. Rasgon JL. Using predictive models to optimize Wolbachia-based strategies for vector-borne disease control.Adv Exp Med Biol 2008;627: 114–25.
18. Topalis P, Tzavlaki C, *et al*. Anatomical ontologies of mosquitoes and ticks, and their web browsers in VectorBase. Insect Mol Biol 2008;17: 87–9.
19. Topalis P, Lawson D, Collins FH and Louis C. How can ontologies help vector biology? Trends Parasitol 2008;24: 249–52.
20. Smith B, Ashburner M, *et al*. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol l2007;25: 251–5.
21. Simon J, Dos Santos M, Fielding J and Smith B. Formal ontology for natural language processing and the integration of biomedical databases. Int J Med Inform 2006;75: 224–31.
22. Grenon P, Smith B and Goldberg L. Biodynamic ontology: Applying BFO in the biomedical domain. Stud Health Technol Inform 2004;102: 20–38.
23. Dialynas, E, Topalis, *et al*. MIRO and IRbase: IT Tools for the Epidemiological Monitoring of Insecticide Resistance in Mosquito Disease Vectors. PLOS Negl Trop Dis, in press.
24. Ashburner M and Lewis S. On ontologies for biologists: The Gene Ontology--untangling the web. Novartis Found Symp 2002;247: 66–80; discussion 80–3, 84–90, 244–52.
25. Harris MA, Clark J, *et al*. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 2004;32: D258–61.
26. Mungall CJ and Emmert DB. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. Bioinformatics 2007;23: i337–46.
27. Gelbart WM, Crosby M, *et al*. FlyBase: A Drosophila database. The FlyBase consortium. Nucleic Acids Res 1997;25: 63–6.
28. Tweedie S, Ashburner M, *et al*. FlyBase: Enhancing Drosophila Gene Ontology annotations. Nucleic Acids Res 2009;37: D555–9.
29. Megy K, Hammond M, *et al*. Genomic resources for invertebrate vectors of human pathogens, and the role of VectorBase. Infect Genet Evol 2008.
31. Lawson D, Arensburger P, *et al*. VectorBase: A data resource for invertebrate vector genomics. Nucleic Acids Res 2009;37: D583–7.
32. Thomson MC, Connor SJ, *et al*. The ecology of malaria--as seen from Earth-observation satellites. Ann Trop Med Parasitol 1996;90: 243–64.
33. Coleman M, Sharp B, *et al*. Developing an evidence-based decision support system for rational insecticide choice in the control of African malaria vectors. J Med Entomol 2006;43: 663–8.

An Ontology for Designing Models of Epidemics

**Geoffrey A. Frank, William D. Wheaton, Vesselina Bakalov,
Philip C. Cooley, Diane K. Wagener**
RTI International, Research Triangle Park, NC, USA

Abstract

Models of epidemics allow decision makers to explore the consequences of different interventions. The Models of Infectious Disease Agent Study (MIDAS) project has been collecting studies, models, data supporting the models, and publications providing historical evidence about epidemics.

An ontology has been developed for MIDAS to support this collection, documentation, and dissemination. It uses relations to link taxonomies (including a subset of the infectious disease ontology) that define the range of potential models or supporting documentation.

The ontology is used to aid in the navigation process that is part of the user interface for identifying which studies and publications are available in the MIDAS repository (MREP) that are consistent with the many parameters associated with a particular study.

Keywords: Infectious disease models, disease transmission, ontology, taxonomy

Introduction

Models of epidemics that allow decision makers to explore the consequences of different interventions are a valuable tool for public health officials. For these tools to be effective, the models must be validated and the resulting conclusions must be convincing to the decision makers. The process of creating models, designing the series of modeling studies needed to calibrate them, and providing useful feedback on the consequences of interventions requires close collaboration among an interdisciplinary team. The team must be able to integrate data from databases for multiple disciplines and understand the implications of modeling requirements from multiple viewpoints. The MIDAS program is an example of such collaboration.

MIDAS

MIDAS was created to help improve the nation's ability to respond to biological threats promptly and effectively.[1] MIDAS is a research partnership between the National Institute of General Medical Sciences and the scientific community to develop computational models for policymakers, public health workers, and researchers. To date, MIDAS primarily has developed agent-based influenza transmission models to assess influenza prevention and containment strategies that include the broad categories of antiviral medications, vaccines, and nonpharmaceutical interventions, such as case isolation, household quarantine, school or workplace closure, and travel restrictions. The models have been used to generate simulations to study the following:

- Influenza epidemics and containment options in Southeast Asia[2,3]
- The potential spread of a pandemic strain of influenza virus through a U.S. population of approximately 300 million individuals[4,5]
- An epidemic in Chicago, Illinois.[6] These research efforts are supported by the MIDAS repository (MREP), which stores and manages the computerized models, model results, model parameters, and the specifications used to develop the models.[7] Each element in the MREP is cataloged in a relational database. The database links these elements so the inputs to a specific model and the corresponding results are connected. The ontology described herein is an evolving method for making this information more accessible to policymakers, public health professionals, and researchers.

Components of an Epidemic Model

To conduct a study, the modeler must create a simulation by assembling and calibrating several components. This usually involves a series of tradeoffs in terms of how these components are defined and how the validity of the associated parameter values are established. The ontology is designed to show the user which studies are available that meet the user's requirements across the range of models. The ontology also assists the user in finding publications and other data sources that are available to calibrate a model meeting the user's requirements. The following paragraphs describe the major components and indicate the scope of the ontology.

The Disease Model

A primary element of an agent-based epidemic model is a representation of the natural history of the disease in individuals in the model, whom we refer to as agents. One common approach is to represent the history in terms of a set of transitions between states

such as the Susceptible–Infectious–Removed (SIR) model. The initial conditions of an epidemic model place all of the agents into either a *susceptible* state or an immune state (which is a special case of the *removed* state). Upon contact with agents in the infectious state, susceptible agents (S) transition to the infectious state (I) with a specified probability. After a number of days, the agents transition to the *removed* state (R). Because these agents are now either dead or immune to the pathogen, they remain in state R for the rest of the time period of interest. A significant step in a model's calibration process is to demonstrate that the probabilities for these transitions in the simulation are consistent with historical data.

Social Network Models

A key tradeoff when constructing a model involves the level and nature of the population disaggregation. The social network model disaggregates the population into subgroups according to certain characteristics (e.g., age, gender) and provides behaviors for each of these subgroups. The subgroup behavior models include appropriate disease state–transition rates and likelihoods of contacts between infected and susceptible agents. Targeting interventions to specific subgroups is a critical element of an effective strategy for combating an epidemic. The social network model must be disaggregated to a level that distinguishes the effects of targeted interventions. At the same time, data sources are needed to substantiate the different subgroup models. The ontology is being developed to help the user make tradeoffs in the level of disaggregation based on the interventions and output resolution and available data sources.

Social network data are combined with transportation data to further specify the behavior of agents. Agents are initially assigned to households in a manner consistent with census data. The transportation data are used in combination with census data to determine where the agents work or go to school. Where the patterns of movement are more diffuse, a gravity model is used that prioritizes the destinations based on a metric that weighs the basic attractiveness of a destination against the distance to be traveled to reach the destination.

Some workplaces, such as hotels and hospitals, have special significance for epidemic models. Hotels are a primary location for the mixing of travelers. Hospitals are significant because their workers are more likely to come in contact with infected agents.

Intervention Models

The primary goal for the development of these models is to assist decision makers in assessing the expected consequences of their interventions through a series of "what if" studies. Part of the development of these "what if" studies involves modeling a potential set of interventions. Typical interventions include vaccinations; treatment of the symptomatic population via antiviral medications; and social distancing method, such as the closing of schools, workplaces, and public means of transportation. If interventions are directed at particular social networks (e.g., prophylactic treatment of children in a school), then the model must provide separate behaviors for each social group of interest.

Study Design

Constructing an epidemic model involves calibrating the model against historical data before running the "what if" analyses. The study design section of the ontology is being designed to indicate what historical data are available to calibrate a study. This section of the ontology also identifies what data are collected as part of the available studies.

Computational Framework

The choice of the computational framework can have a major impact on the computational requirements of a study and is, therefore, one of the most important decisions in the design of a study. While designing a study, the research community would like to be able to choose from a hierarchy of models—from calibrated differential equations to complex agent-based model simulations.[8] Different computational frameworks have different input data requirements. This section of the ontology describes what computational framework was used for each of the studies.

An Ontology for Models of Epidemics

Ontology Architecture

This ontology is defined in terms of a simple high-level entity-relationship model in which the entities are taxonomies constructed with IS_A and PART_OF relations.[9] Attributes are associated with the entities, including lists of the publications and studies that are consistent with the characteristics of the entity. Each taxonomy has an associated class hierarchy[10] that defines the inheritance of attributes to the lowest level entities and also specifies rollup rules for aggregating information lower in the taxonomy. The combination of entity relationships and attributes allows the ontology to communicate with source databases to obtain initial attribute values. The class hierarchies of rules allow the ontology to process hierarchical structures such as taxonomies.

Taxonomy Structure

The taxonomies are constructed as class hierarchies with three different base relations: parts, choices, and options. Each of these relations has a set of functions associated with its class that are inherited to all instantiations of that relation. An *instance object* may have studies and documents associated with it. An *abstract object* may have associated documents and a base relation. The required parts of an object are specified by parts classes. Choices for an object are specified by choices classes. Optional components for an object are represented by options classes.

Component Taxonomies

The following taxonomies are required parts of the ontology built for these models:

- *A taxonomy of infectious diseases.* This taxonomy uses the terminology from the Infectious Disease Ontology (IDO) website.[11] The primary reason for using the same identifiers as the IDO is to link to additional information about diseases being developed by other IDO users. This taxonomy is represented as a hierarchy of choices; only one disease is modeled in a simulation run.

- *The geographic area of interest for a model.* This is selected from a choices taxonomy with two levels: the scope of the region modeled and the instance of the region The top level choice is the size of the region. The second level choice is the location of the region.

- *A taxonomy of transportation models.* This options taxonomy describes the ways in which the modeled individuals move into contact with one another. Multiple transportation models can be used in a single simulation for different modes of transportation.

- *A taxonomy of interventions.* This taxonomy has five options: vaccination, treatment, quarantine, social distancing, and other interventions addressing vectors of a disease. This is another potential application of IDO identifiers.

- *A taxonomy of social groups.* This taxonomy includes households, workplaces, schools, communities, and group quarters as options.

- *A taxonomy of epidemic dates.* This choices taxonomy is used to identify historical data.

- *A taxonomy of model outputs.* This taxonomy includes options for the time resolution of trend analysis, levels of data aggregation, and output visualization methods.

Taxonomy Dependencies

The ontology includes dependency relations that link together the taxonomies. One set of dependencies links options to choices. For example, options for model outputs for school-age children implies appropriate interaction models for children, typically involving the inclusion of schools as social groups. Conversely, if schools are modeled as social groups, then school-age children become a viable output aggregation category.

Another form of dependency links the model to the required parameters for that model and to any sources of historical data to calibrate those parameters. Each of the component taxonomies described above previously has an associated set of parameters. For example, the different genetic structures of influenza react differently to various vaccines, which implies different disease state–transition rates based on the pharmaceutical intervention selected. Part of the ontology development process involves encoding these dependencies as relations in the ontology so the ontology can ensure that the disease natural history parameters are consistent with the particular strain of influenza and the interventions being studied.

Application of the Ontology to Support Modeling

A User Interface

The ontology described has been used to develop a user interface to identify which studies and publications in the MREP match an evolving specification for a given study. The user works through a set of menus arranged in terms of four spaces of parameters[12] as follows:

- The *background space,* where the user specifies the disease and geographic region of interest.

- The *decision space,* where the user specifies the interventions of interest.

- The *situation space,* where the user specifies the characteristics of the disease and initial conditions of the study.

- The *study design space,* where the user defines the range of parameters to be varied, the metrics of the epidemic to be tracked, and the outputs to be calculated from the metrics and how they are to be displayed. The user also selects the timeframes and resolution of the study.

Benefits of an Ontology-Based User Interface

The ontology structure provides two key benefits relative to a simple Boolean query processor to access the MREP study and publication databases. First, the dependencies between the taxonomies reduce the set of options for later menu structures. The menus are generated dynamically, reflecting earlier decisions. Given the large space of parameters associated with these models, there is value in using the relations of the ontology to reduce the number of options facing the user.

Second, the taxonomies provide multiple levels of abstraction. This provides the user with more control over the space of parameters by controlling the level of abstraction along different dimensions. For example, the user may focus on only the H1N1 influenza virus or expand the set of possible studies or publications of interest to include all influenza-like diseases (such as H5N1 flu strains). Similarly, the user can focus on models that represent specific group quarters (such as prisons or nursing homes) or expand the set of possible studies or publications to include any combination of group quarters. This strategy of increasing abstraction has also been used in defining the time period covered by historical studies. A researcher can focus on a specific epidemic year or look for historical information over the span of a decade or more.

Conclusions

This paper presents an ontology that supports the development of models of epidemics and their respective mitigating interventions. It links studies and publications to the parameters of computer models of epidemics. This ontology is in the early stages of development and is currently being used to describe the contents of the MREP. However, the ultimate success of the ontology will depend upon its ability to interface (through related ontologies) with bibliographies of publications describing disease processes and the efficacy of interventions.

Acknowledgments

We thank the National Institute of General Medical Sciences MIDAS Program for research funding (Grant #5U01GM70698-5).

References

1. Cooley P, Ganapathi L, Ghneim G, *et al*. Using influenza-like illness data to reconstruct an influenza outbreak. *Mathematical and Computer Modelling*. 2008;48(5,6):929–939.
2. Ferguson N, Cummings DAT, Cauchemez S, *et al*. Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature*. 2005;437:209–214.
3. Longini I Jr, Nizam A, Xu S, *et al*. Containing pandemic influenza at the source. *Science*. 2005;309:1083–1087.
4. Germann TC, Kadau K, Longini IM Jr., *et al*. Mitigation strategies for pandemic influenza in the United States. *PNAS*. 2006;103(15):5935–5940.
5. Ferguson N, Cummings DAT, Fraser C, *et al*. Strategies for mitigating an influenza pandemic. *Nature*. 2006;442:448–452.
6. Halloran ME, Ferguson NM, Eubank S, *et al*. Modeling targeted layered containment of an influenza pandemic in the USA. *PNAS*. 2008;105(12):4639–4644.
7. Cooley PC, Roberts DJ, Bakalov VD, *et al*. The MIDAS Model Repository (MREP). *IEEE Transactions on Information Technology in Biomedicine*. 2008;12(4):513–522.
8. Grenfell B. Modeling of infectious disease dynamics and control. White paper presented to the NIGMS Advisory Council, May 18, 2007.
9. Smith B and Neuhaus F. Modeling principles and methodologies—relations in anatomical ontologies. In: Burger A, Davidson D and Baldock R (eds.) *Anatomy Ontologies for Bioinformatics: Principles and Practice*. New York, NY Springer; 2007: 289–306.
10. ISO Standard 21127:2006. Information and documentation—A reference ontology for the interchange of cultural heritage information Available at www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=34424. Accessed February 2009.
11. The Infectious Disease Ontology website. Available at www.infectiousdiseaseontology.org/Home.html. Accessed February 2009.
12. Bobashev G, Goedecke M, Morris R *et al*. Distribution strategies for the limited amounts of vaccines in the context of pandemic influenza. Presented to the World Health Organization, October 21, 2008.

Open Biomedical Ontologies Applied to Prostate Cancer

James A. Overton[1], Cesare Romagnoli[2], Rethy Chhem[3]
[1]Department of Philosophy, The University of Western Ontario, London, Canada
[2]Department of Medical Imaging, London Health Sciences Centre, London, Canada
[3]Director of Human Health, International Atomic Agency, Vienna, Austria

Abstract

This paper surveys preliminary results from the Interdisciplinary Prostate Ontology Project (IPOP), in which ontologies from the Open Biomedical Ontologies (OBO) library have been used to annotate clinical reports about prostate cancer. First we discuss why we rejected several controlled vocabularies, including SNOMED, DICOM, and RadLex, preferring instead to use the OBO library. We then briefly describe the database-backed website we have created around the relevant OBO ontologies, and provide excerpts of reports from radiology, surgery, and pathology which we have hyperlinked to the ontology terms. This method allows us to discover which relevant terms exist in the OBO library, and which do not. The final section of this paper discusses these gaps in the OBO library and considers methods of filling them.

Introduction

The Interdisciplinary Prostate Ontology Project aims to develop expertise with biomedical ontologies at the University of Western Ontario and the London Health Sciences Centre. This paper surveys results from the first stage of IPOP, which assessed existing biomedical ontology tools and applied them to clinical reporting about prostate cancer.

The main goal of IPOP is to improve communication between medical practitioners from radiology, oncology, anatomy, surgery, pathology, and other areas. Communication is often impeded by local variations in the use of terminology. Controlled vocabularies are part of the solution to this problem. Biomedical ontologies improve upon controlled vocabularies by linking together terms and thus allowing for better computerized data collection, search, and analysis. We hope that improving communication will ultimately lead to better patient outcomes.[1]

Our first step was to assess different approaches to ontologies and controlled vocabularies. For reasons discussed below, we rejected controlled vocabularies including SNOMED CT, DICOM, and RadLex. Our preferred alternative is to use ontologies from the Open Biomedical Ontologies consortium, and in particular from their OBO Foundry initiative. OBO includes a network of well designed, interoperable ontologies, which cover a wide range of relevant terminology.

We then created a database-backed website around the relevant OBO ontologies. Each ontology term has a web page describing it and linking it to parent and child terms via relations like “is_a” and “part_of”. We collected examples of radiology, surgery, and pathology reports dealing with prostate cancer, added them to our website, and hyperlinked the relevant terms in those reports to terms in the database of ontologies. Our approach is designed to reveal which relevant terms already exist within the OBO ontologies and which are not yet included in any OBO ontologies. We conclude this paper by suggesting fragments of one ontology, which would be useful for completing the annotation of these reports.

IPOP shares much in common with Marwede and Fielding’s recent work on ontologies in clinical radiology.[2] Their focus is on radiology reporting, in many different forms, with a view to annotating medical images using ontology terms. Our focus is on prostate cancer across medical disciplines, but we are also interested in extending the use of ontologies to medical image annotation.

Controlled Vocabularies

We assessed several controlled vocabularies before settling on OBO ontologies. This section outlines these alternatives and our reasons for setting them aside.

SNOMED CT

Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) is a very large collection of medical terms which is becoming widely used throughout the world. The chief advantage of SNOMED is its broad coverage of terms describing medical practice. But SNOMED has been criticized for not following sound ontological theory[3], and for including many errors of classification[4]. OBO and many Semantic Web systems show that a network of specialized tools is often more flexible and comprehensive in the long run than centralized and monolithic approaches like SNOMED.

DICOM

Digital Imaging and Communications in Medicine (DICOM) is an industry driven standard for medical imaging files, their storage, and their transmission over networks. The DICOM Structured Reporting standard (DICOM-SR) extends DICOM to the kinds of textual reports considered in the first stage of IPOP[5]. DICOM covers many terms important for our future work on medical image annotation. But DICOM-SR is not yet in wide use, and since the terms defined in "DICOM Part 16: Content Mapping Resource" are largely derived from SNOMED, for our current purposes they share the same problems.

RadLex

The Radiology Lexicon (RadLex) is a controlled vocabulary developed by the Radiology Society of North America (RSNA). There are ongoing efforts to transform RadLex from a lexicon into an ontology[6], and to reduce its duplication of other more comprehensive ontologies like the Foundational Model of Anatomy (FMA).[7]

OBO Ontologies

Each of the preceding tools has its strengths. In particular, they share a focus on the practice of medicine and medical interventions. However our group has chosen to pursue an alternative approach.

The Open Biomedical Ontologies library includes a large number ontology projects, and the OBO Foundry initiative aims to unite them under a set of shared best practices.[8] Each OBO Foundry ontology is specialized for a particular domain and designed by a group of experts in the relevant field. When two domains overlap, the goal is keep their ontologies separate but link them together, creating a division of labour.

Advantages of the OBO Foundry ontologies include permissive licenses, open source practices, a human-readable common file format, and a shared Basic Formal Ontology (BFO). Although many of the OBO ontologies are incomplete, and although there are many domains relevant to IPOP which are not yet covered by OBO, in our assessment the OBO Foundry approach is the best bet for future work in biomedical ontologies.

Below we describe in brief some of the OBO ontologies which we have found useful for our goal of annotating clinical reports about prostate cancer in humans.

- Foundational Model of Anatomy (FMA): provides IPOP with terms such as FMA:9600 "prostate", as well as terms for the parts of the prostate and its neighbouring organs.
- Disease Ontology (DO/DOID): parallels the structure of the FMA and describes diseases of various portion of the human anatomy. For IPOP the main terms of interest are DOID:47 "prostate disease" and its children, including DOID:514 "prostatic neoplasms" and DOID:8634 "carcinoma in situ of prostate".
- Protein Ontology (PRO): proteins like prostate specific antigen (PSA) are important for the detection of prostate cancer.
- Gene Ontology (GO): provides terms related to PSA, such as GO:0004252 "serine-type endopeptidase activity" and GO:0016525 "negative regulation of angiogenesis".
- Phenotype Quality (PATO): designed to describe phenotypes of organisms, it contains many terms useful for qualitative assessments in general. Relevant to IPOP are the children of the term PATO:0000014 "color" used in biopsies, as well as terms for patterns (e.g. PATO:0000060 "spatial pattern") and textures (e.g. PATO:0000701 "smooth").
- Units of Measurement (UO): organizes the International System of Units (SI) into an ontology, and includes other terms such as UO:0000190 "ratio".

Annotating Reports

In the first stage of IPOP our focus has been on annotating the text of clinical reports using ontology terms. These three samples come from reports on three separate patients. A selection of OBO terms is marked in bold and explained.

Radiology Report Sample

"**Peripheral Zone**: This zone is relatively homogeneous with a **smooth contour** although it is compressed by a large **transition zone**."

- "peripheral zone" corresponds to FMA:19587 "peripheral zone of prostate"
- "smooth" is PATO:0000701
- "contour" corresponds to PATO:0000052 "shape"
- "transition zone" corresponds to FMA:45721 "transition zone of prostate"

Surgery Report Sample

"Once the **prostate** was mobilized in a cephalad direction, I could see **Denonvilliers fascia**. This was opened in the midline. We then dissected out the **ampulla of Vater**, which were clipped and divided.

The **seminal vesicles** were dissected off in their entirety quite easily using clips for **hemostasis**."

- "prostate" is FMA:9600
- "Denonvilliers fascia" is a synonym for FMA:19933 "rectovesical septum"
- "ampulla of Vater" is a synonym for FMA:15076 "hepatopancreatic ampulla"
- "seminal vesicle" is FMA:19386
- "hemostasis" is GO:0007599

Pathology Report Sample
"The specimen consist of 2 cores of **pale tan tissue**, the larger measures 1.3 **cm** and the smaller measures 1.1 **cm**. All **tissue** is submitted in one cassette."

- "pale tan" is a close synonym of PATO:0001268 "desaturated brown"
- "tissue" corresponds to FMA:9637 "portion of tissue"
- "cm" is UO:0000015

Gaps in OBO Ontologies

Although the OBO ontologies provide a broad and rich source of terms which we have been able to use to annotate our reports, there are important terms which cannot be found in any of the existing ontologies. Some of these terms are included in SNOMED CT, DICOM, and RadLex, but have not yet been integrated into the OBO library. In many cases these gaps reflect the focus of SNOMED CT, DICOM, and RadLex on medical practice and interventions, while OBO ontologies tend to focus on biomedical investigations.

Near Synonyms
As might be expected, there are phrases used in the reports which do not match the name of any term in an OBO ontology, nor any of the synonyms given, but are clearly meant to refer to an existing term. The example above of "pale tan" is one such case. Since our reports are annotated by human beings and not automatically, it is possible to discern the author's intent in most cases and add the hyperlink. Variation in synonyms poses a problem for automatic annotation methods.

Missing Composites
There are UO terms for density, including "milligram per milliliter", but no term for "nanogram per milliliter" used in PSA measurements. This points to an issue of compositionality in UO which is shared by other ontologies: not all possible combinations of terms are included in the ontology. This can be remedied either by adding the composite term to the ontology, or by using two ontology terms linked together by a well-defined relation.

Conflicting Fiat Boundaries
Ontologies like FMA include *bona fide* boundaries like those between bones and between organs, as well as *fiat* boundaries drawn for the sake of convenience. While there tends to be good agreement about *bona fide* boundaries, it is easy to find disagreements about how fiat boundaries should be established.

For instance, the FMA divides the regional parts of FMA:9600 "prostate" into the anterior, posterior, right lateral, and left lateral lobes. Each of these lobes is given its own child terms for fibromuscular stroma, vasculature, neural network, and parenchyma. But RadLex divides the prostate's RID:344 "anterior fibromuscular stroma" into outer and inner glands, and then into peripheral, central, and transition zones. Since prostate tumours are much more common in the peripheral zone of the prostate, it is more natural to use the RadLex nomenclature than the FMA fiat boundaries when dealing with prostate cancer. Including two distinct sets of fiat boundaries is likely to be problematic, as is mapping a set of fiat boundaries in one ontology onto a different set in another ontology.

Medical Procedures
OBO does not currently include ontologies for describing the medical procedures relevant to IPOP. Also missing are terms for the tools used to perform these procedures. For instance: digital rectal exam, transrectal ultrasound; biopsy, specimen, core, fragment, cassette; surgery, mobilize, dissect, clips.

There are ontologies being developed like the Ontology for Biomedical Investigations (OBI) which may fill some of these gaps. Investigations are, of course, distinct from interventions, and the latter is the primary concern of IPOP.

If no existing ontologies can be found to fill these gaps, the alternative is to coordinate with others to create a new ontology for these terms using OBO Foundry best practices, and submit it to the OBO consortium.

Image Types
Important for the radiological aspects of IPOP is an ontology for medical image types. Below is a proposed fragment of an "is_a" hierarchy for types of medical image.

- Medical Image
  - Magnetic Resonance Imaging (MRI) Image
    - T1 Weighted MRI Image

        - MRI Image without Contrast
        - MRI Image with Contrast
            - Static Contrast MRI Image
            - Dynamic Contrast MRI Image
    - T2 Weighted MRI Image
        - Proton Density Weighted MRI Image
        - Diffusion MRI Image
- Ultrasound (US) Image
    - A Mode US Image
    - B Mode US Image
        - Conventional US Image
        - US Image with Contrast
        - Doppler US Image
            - Power Doppler US Image
            - Colour Doppler US Image
            - Pulse Doppler US Image
    - M Mode US Image
- X-Ray Image
    - Computed Tomography (CT) Image
        - CT Image with Contrast
        - CT Image without Contrast
    - …
- Nuclear Medicine Image
    - …

RadLex and DICOM contain many useful terms for medical images which could be adapted into a new ontology. It is important to maintain the distinction between medical imaging procedures and the images which are products of those procedures.

Conclusion

The first stage of IPOP has been successful in applying the OBO library to a small set of clinical reports from radiology, surgery, and pathology. The process is labour intensive, which has limited the number of cases we have annotated. Nevertheless, the small set of annotated cases will be useful as an education tool, raising awareness of the availability of ontology and the utility of controlled vocabularies. Analysis of the annotated reports has also helped us to find examples of conflicting term usage in our practice.

It is worth reiterating the difference between controlled vocabularies focused on medical practice, such as SNOMED CT, DICOM, and RadLex, and the OBO ontologies which focus on biomedical research. IPOP requires a firm foundation in biomedical science, but our goal is to improve medical practice. The evolving shared principles of the OBO Foundry promise to avoid many of the pitfalls encountered in the other controlled vocabularies. And we believe that these shared principles will prove just as fruitful in the development of new ontologies, aimed primarily at medical interventions, as they have already in developing ontologies aimed at biomedical investigations.

Our ongoing work on IPOP builds on the successes of this first stage by annotating sample reports from more fields, attempting to fill in the gaps discussed here, and considering the use of ontological annotations for medical images.

Acknowledgements

This research is supported by funding from the Department of Radiology and Nuclear Medicine at the London Health Sciences Centre, and from the Joseph L. Rotman Institute of Science and Values at the University of Western Ontario.

References

1. Arp R, Romagnoli C, Chhem RK and Overton JA. Radiological and biomedical knowledge integration: The ontological way. In Chhem RK, Hibbert KM and Van Deven T. Radiology Education: The Scholarship of Teaching and Learning, 2008;87–104.
2. Marwede D and Fielding JM. Entities and relations in medical imaging: An analysis of computed tomography reporting. Applied Ontology, 2007;2(1):67–79.
3. Ceusters W, Smith B and Flanagan J. Ontology and medical terminology: Why description logics are not enough. In Proceedings of TEPR, 2003;10–14.
4. Ceusters W, Smith B, Kumar A and Dhaen C. Mistakes in medical ontologies: Where do they come from and how can they be detected? Studies in Health Technology and Informatics, 2004;145–164.
5. Clunie DA. DICOM structured reporting. PixelMed, 2000.
6. Rubin DL. Creating and curating a terminology for radiology: Ontology modeling and analysis. Journal of Digital Imaging, 2007;21(4):355–362.
7. Mejino JL, Rubin DL and Brinkley JF. FMA-RadLex: An application ontology of radiological anatomy derived from the Foundational Model of Anatomy reference ontology. In AMIA 2008 Symposium Proceedings, 2008:465.
8. Smith B, *et al*. The OBO foundry: Coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology 2007:25: 1251–1255.

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

SNOMED CT's Ontological Commitment

Stefan Schulz[1], Ronald Cornet[2]
[1]University Medical Center, Freiburg, Germany
[2]Academic Medical Center, Amsterdam, The Netherlands

Abstract

SNOMED CT is a clinical terminology that describes the meaning of terms by logical axioms. This requires an ontological commitment, i.e. precise agreements about the ontological nature of the entities referred to. We provide evidence that SNOMED implicitly supports at least three different kinds of commitments, viz. (i) independently existing entities, (ii) representational artifacts, and (iii) clinical situations. Our analysis shows how the truth-value of a sentence changes according to one of these perspectives. We argue that a clear understanding of to what kind of entities SNOMED CT concepts extend is crucial for the proper use and maintenance of SNOMED CT.

Introduction

SNOMED CT[1] is the inheritor of a dynasty of medical nomenclatures and coding systems[2] which had been constructed to provide:

1. Semantic descriptors to annotate and encode clinical procedures, diagnoses, observables, etc.;
2. Standardized medical terms in different languages;
3. Guidance for the construction of composed terminological expressions.

SNOMED CT's predecessors made only very basic claims with respect to the domain they represented. The meaning of semantic descriptors was given by the intuitive understanding of the terms they were linked to and it was assumed that they were correctly interpreted by the (human) language users. Therefore, none of these systems made any attempt to formally represent any reality beyond a rough mapping of controlled terms to shared concepts with the aim to reduce the high variability of human language through a set of controlled terms or to support the encoding of medical data by means of a coded thesaurus of procedural and administrative terms.

With the advent of SNOMED RT (and later SNOMED CT), logics entered the scene and added a mathematically rigorous layer to the hitherto informal, close-to-human-language representation of medical terms. However, the use of logic axioms and theorems in a terminology (which imposes the assignment of truth-values) requires an equally precise agreement about the objects and relations being denoted by the terms and concepts. This is commonly called ontological commitment[3].

In this paper we will substantiate the claim that SNOMED CT's ontological commitment is inconsistent. To this end we will scrutinize three frequent SNOMED CT design features, viz. (i) qualifiers and their values, (ii) context-dependent concepts, and (iii) multiple parenthood.

Furthermore, we will discuss the pros and cons of the inferences they enable and discuss them in the light of competing ontological commitments.

Description Logics

SNOMED CT's backbone is given by a taxonomy of nodes, called SNOMED CT concepts. Every concept represents the characteristic properties of all its (concrete) instances. This is done by a parsimonious variant of description logics (DL)[4], which we will briefly introduce. Key notions in DL are classes and instances (their extensions). So is the class *Liver* instantiated by every individual liver, just as *Bodily Organ* extends to all individual bodily organs. Putting those two statements together, we get the hierarchy-building principle of taxonomic subsumption: *Liver* is subsumed by *Bodily Organ.* In DL this is expressed by *Liver* $\sqsubseteq$ *Bodily Organ*, which asserts that every *Liver* instance is also an instance of *Bodily Organ*[i].

More complex statements can be obtained by combining representations of classes with operators and quantifiers. In the following example, we employ the $\sqcap$ ("and") operator and add a quantified role, using the existential quantifier $\exists$ ("exists"). For example, the expression *Inflammatory disease* $\sqcap \exists$ *has location.Liver* extends to all instances in which both instantiate *Inflammatory disease* and are further related through the relation *has location* to some *Liver*. This example actually gives us both the necessary and the sufficient conditions needed in order to fully define a class, e.g.: *Hepatitis* $\equiv$ *Inflammatory disease* $\sqcap \exists$ *has-location.Liver*, with the equivalence operator $\equiv$ telling that (i) every particular instance of *Hepatitis* is also an instance of *Inflammatory disease* that is located at some instance of *Liver*, and (ii) that every instance of *Inflammatory disease* that is located at

[i]Another way of referring to taxonomic subsumption (DL operator $\sqsubseteq$) is the use a relation named "is-a"

some *Liver* is an instance of *Hepatitis*. Hence, in any situation, the term on the left can be replaced by the expression on the right without any loss of meaning.

Running Examples

Having introduced SNOMED CT's formal background we will base our forthcoming deliberations on four examples taken from the January 2009 release of SNOMED CT. All these examples are representative as the phenomena they incorporate occur frequently throughout the terminology.

1. Infusion Pump

(430033006), a primitive concept in the *Physical object* branch:
Infusion pump ⊑ *Pump* ⊑ *Instrument, device* ⊑ … ⊑ *Physical object.* All SNOMED CT concepts are inserted in this kind of subsumption hierarchies.

2. Denied Tonsillectomy

(173422009|: 272125009|= 82975001), a postcoordinated concept, refining tonsillectomy by using the qualifier "priority" with the value "denied", in DL notation: *Tonsillecomy* ⊓ ∃ *Priority.Denied.*
All SNOMED CT procedure concepts (~50,000) allow analogous refinements.

3. Heart Operation Planned

(183983001). This concept is in SNOMED CT's *Situation with explicit context* branch and is fully defined as

∃ *rg.*(∃ *associated_procedure.Operation on heart* ⊓
∃ *procedure_context.Planned* ⊓
∃ *subject_rel_ context. Subject of record* ⊓
∃ *temporal_context.Current or specified time*) [ii]

There are currently 17 concepts with the context *Planned*, but hundreds with similar contexts such as *Suspected* or *Known Absent*.

4. Tetralogy of Fallot

(86299006). This concept is subsumed by the four concepts: *Pulmonic valve stenosis*, *Ventricular septal defect*, *Overriding aorta*, and *Right ventricular hypertrophy.* More precisely, it implies the following expression:

∃ *rg.*(∃ *assoc_morphology.Congenital Anomaly* ⊓
∃ *finding_site.Cardiac Ventricular Structure*) ⊓
∃ *rg*. (∃ *assoc_morphology.Defect* ⊓
∃ *finding_site.Intraventricular Septum Structure*) ⊓
∃ *rg*. (∃ *assoc_morphology.Stenosis* ⊓
∃ *finding_site.Pulmonary Valve Structure*) ⊓
∃ *rg*. (∃ *assoc_morphology.Overriding Structures* ⊓
∃ *finding_site.Thoracic Aorta Structure*)

Nearly 77,000 SNOMED CT concepts contain relationship groups.

[ii] *rg* means „role group", cf.[6]

Using these examples we now want to demonstrate the different ontological commitments occurring in SNOMED CT. In other words, we will ask the question: which entities in the clinical context are instantiated by SNOMED CT concepts? Below we present three possibilities: independently existing entities, representational artifacts, and clinical situations.

SNOMED CT concepts are instantiated by objects that exist independently of the clinical context

We will call this the *standard interpretation*, as it is the most straightforward one and corresponds to the view commonly defended by the realist approach to ontologies. This stance postulates the existence of real objects and processes as independent of the circumstances of their observation[5].

Under this viewpoint, the concept ***Infusion pump*** would be instantiated by each and every individual infusion pump, independent of its involvement in any clinical process. It would not designate the mental concept or construction plan of an infusion pump. In the same line, the concept ***Tonsillectomy*** would be instantiated by every really occurring surgical removal of a tonsil. ***Heart operation planned*** would denote a plan of a heart operation as the result of a real physician's decision, and ***Pulmonary Valve Stenosis*** would morphologically altered state of an existing pulmonary valve in a real patient.

SNOMED CT concepts are instantiated by representational artifacts as contained in an electronic patient record

We will call this the *EHR interpretation*. Under this view it does not matter whether some thing really exists or not. The only criterion is a mention in a documentation artifact such as an electronic patient record (EHR). This can be nicely shown by the postcoordinated SNOMED CT concept ***Denied tonsillectomy***. It is not instantiated by a real tonsillectomy but by an EHR entry on tonsillectomy, an information object which may be further refined by qualifiers such as *Denied*, *Planned*, *Scheduled.* Under this point of view, an EHR entry "denied tonsillectomy" is indeed subsumed by the entry "tonsillectomy", so that this sentence holds true. How such an entry is interpreted by the EHR used in terms of what things in reality it denotes is not relevant here. But it is obvious that on the level of real objects and processes a denied tonsillectomy can never be a kind of tonsillectomy, so that the sentence is false at the level of real objects.

A similar line of reasoning applies to the example ***Heart operation planned***. Although this concept is

not a subconcept of *Heart operation* (which would parallel the above example), its standard interpretation leads to contradictions: According to its definition, *Heart operation planned* implies the sentence:

∃ *associated_procedure.Operation on heart.*

Following the description logics semantics this means that for each instance of *Heart operation planned* there must be at least one instance of *Operation on heart*. This contention can easily be disproved as planned procedures are not always executed. If *Heart operation planned*, on the contrary, is interpreted as to be instantiated by EHR objects, the sentence becomes true, using the same argument we used to justify the subsumption relation between *Tonsillectomy* and *Denied tonsillectomy*.

SNOMED CT concepts are instantiated by patients or clinical situations.

Typical examples that suggest this third flavor of interpreting SNOMED CT concepts is suggested by the way SNOMED CT formalizes composed clinical findings and procedures. The standard interpretation conflicts with the fact that all elements of a combined finding, such as the complex heart malformation called *Tetralogy of Fallot* are introduced as its taxonomic parents. As a result, *Tetralogy of Fallot* is subsumed by the concepts *Septal defect* and *Cardiomegaly*, among others. Furthermore, SNOMED CT separated the findings from the morphology using so-called relationship groups. According to[6], role groups are expressed in DL as an anonymous relation called *rg*.

Role groups order the elements of a complex concept definition and prevent it from ambiguous associations. If we re-interpret *rg* as *has_part* as proposed by[7] there is little to criticize from an ontological point of view. However, role groups also appear in definitions where the reason is not obvious, e.g.

Pulmonic Valve Stenosis ≡ ∃ *rg*.
(∃ *assoc_morphology.Stenosis* ⊓
∃ *finding_site.Pulmonary Valve Structure*)

Let us rephrase this equivalence, taking description logics semantics seriously:

"*Every pulmonic valve stenosis has some part which exhibits at least one stenosis somewhere at a pulmonary valve; and everything having some part which exhibits at least one stenosis somewhere at a pulmonary valve is a pulmonic valve stenosis*".

Whereas the first phrase sounds somewhat circular, the second one expands the concept of pulmonic valve stenosis to an extent that each and every condition which is characterized, among other things, by a stenotic pulmonary valve, is subsumed by the concept *Pulmonic Valve Stenosis.* It is no wonder that in the SNOMED CT hierarchy, this concept does not only subsume *Congenital Stenosis of Pulmonary Valve*, but also *Pulmonic Valve Stenosis With Insufficiency*, *Tetralogy of Fallot* and *Pentalogy of Fallot*.

Coming back to the question of ontological commitment: if we understand by the extension of the concept *Pulmonic Valve Stenosis* the pathological structure as it exists in a patient, then we can't but reject the view that a *Tetralogy of Fallot* is a kind of *Pulmonic Valve Stenosis*. What should be criticized here is that implication is mistaken for subsumption: Of course, for every *Tetralogy of Fallot* there is some *Pulmonic Valve Stenosis*. However, this does not mean that *Tetralogy of Fallot* is related to *Pulmonic Valve Stenosis* by taxonomic subsumption.

The puzzle can be solved if we substitute the standard interpretation by what we will call here the *epidemiological interpretation*. Under this assumption, disorders and finding concepts do not extend to states or processes but to their participants or bearers, i.e. to patients. Hence, *Pulmonic Valve Stenosis* and *Tetralogy of Fallot* are to be read as "patients with a pulmonic valve stenosis" and "Fallot patients". Then the subsumption statement becomes true: Every Fallot patient is also a patient with a *Pulmonic Valve Stenosis.*

The picture is also consistent if we assume that these SNOMED concepts extend to clinical situations[8] rather than to particular disorders or states. Consequently, we can argue that every clinical situation that includes a *Tetralogy of Fallot* also includes a *Pulmonic Valve Stenosis*, paralleling the argument that the set of patients with *Tetralogy of Fallot* forms a subset of the bearers of a *Pulmonic Valve Stenosis.*

Even the EHR interpretation makes sense here, as it is plausible that all records annotated by *Tetralogy of Fallot* should by considered as being annotated by *Pulmonic Valve Stenosis*.

Discussion and Conclusions

As much as we may find good explanations for the discussed types of SNOMED CT modeling decisions we raise our concern in view of the fact that the different ontological commitments are completely implicit and the choice is up to the user. As long there is no agreement on which SNOMED CT concepts extend to objects in clinical reality, to patients, to situations, or to documentation objects, different users may want to express different things by using the same expressions, and misinterpretations may lead to erroneous conclusions. To give just one example: If the same concept is instantiated to

express plans (which always bear the possibility of not being realized) on the one hand and to express actions that have been realized, hospital statistics will become unreliable.

SNOMED CT provides the means to represent situative scenarios that include not only plans and negative contexts but also other contextual "moods" like "suspected" "at risk" or "unknown". This crosses the "ontology - epistemology divide" extends the boundary of what a clinical terminology should represent and therefore overlaps with the realm of information models. The resulting problems with double negations have been intensively discussed in the context of TERMINFO[10].

We therefore defend the position that SNOMED CT should always subscribe to what we named standard interpretation, as it makes no background assumptions and is compatible with the approaches pursued by many other biomedical ontologies, e.g. the ones of the OBO foundry. SNOMED concepts should clearly extend to objects in clinical reality, *viz.* the anatomical structures, the diseases, and the procedures as they occur in patients.

Wherever patients, situations, documentation objects or plans are referred to, this should be made clear in the concept name. For queries that target situations or patients as bearer of disorders but not the disorders themselves, SNOMED CT's postcoordination syntax allows to express this, e.g. by $\exists$ *bearer-of. Pulmonic Valve Stenosis* or $\exists$ *bearer-of. Tetralogy of Fallot.* Using a right-identity rule such as *bearer-of* * *has-part* $\sqsubseteq$ *bearer-of* would then allow to infer that every Fallot patient has a stenosis of the pulmonic valve even if the problematic assertion *Tetralogy of Fallot* $\sqsubseteq$ *Pulmonic valve stenosis* were removed from SNOMED CT. For the encoding of epistemic aspects of the EHR, such as scheduled or cancelled procedures, the consistent use of an information model (e.g. HL7 or openEHR) should be preferred over the idiosyncratic use of logic-based formalism in a clinical terminology.

Furthermore, SNOMED CT should ensure that any qualifier that can be or is attached to concepts is a pure restriction of the concept it qualifies, and not a modification of this concept, as is the case in "*Priority: Denied*". If SNOMED CT aims to provide a way to encode such an information (rather than leaving this task to be solved by an information model), it must be represented in a more consistent way. *Denied tonsillectomy* would then not be a subclass of tonsillectomy, but a subclass of *Denial*. Allowing post-coordinating *Denial* with a procedure then provides a workable way to specify denied procedures using SNOMED CT only.

Acknowledgements

The presented work was executed in the context of the DebugIT project, funded by the EU 7th FP (ICT-2007.5.2-217139).

References

1. International Health Terminology Standards Development Organisation (IHTSDO). Systematised Nomenclature of Medicine – Clinical Terms (SNOMED CT) http://www.ihtsdo.org.
2. Cornet R and de Keizer N. Forty years of SNOMED: A literature review. BMC Medical Informatics and Decision Making. 2008 Oct 27;8 Suppl 1:S2.
3. Gruber TR. A translation approach to portable ontology specifications Knowledge Acquisition. 1993; 5(2), 199–220.
4. Baader F, Calvanese D, McGuiness DL, Nardi D and Patel-Schneider PF. The Description Logic Handbook Theory, Implementation, and Applications (2nd Edition). Cambridge: Cambridge University Press, 2007.
5. Ceusters W and Smith B. A realism-based approach to the evolution of biomedical ontologies. AMIA Annual Symposium Proceedings 2006; 121–125.
6. Spackman KA, Dionne R, Mays E and Weis J. Role grouping as an extension to the description logic of ONTYLOG, motivated by entity modeling in SNOMED. AMIA Annual Symposium Proceedings 2002; 712–716.
7. Schulz S, Hanser S, Hahn U and Rogers J. The semantics of procedures and diseases in SNOMED CT. Methods of Information in Medicine. 2006; 45 (4): 354–358.
8. Rector AL and Brandt S. Why do it the hard way? The case for an expressive description logic for SNOMED. Journal of the American Medical Informatics Association. 2008;15 (6): 744–751.
9. Bodenreider O, Smith B and Burgun A. The Ontology-Epistemology Divide: A Case Study in Medical Terminology. Proc. of the Int. Conf. on Formal Ontology and Information Systems (FOIS 2004), Turin, 4-6. November 2004; Amsterdam: IOS-Press; 2004; 185–195.
10. Krog R, Markwell D, Dolin RH, Davera G, Cheetham E, Hamm R, Spackman K, Rector A, Huff S and Ryan S. Using SNOMED CT in HL7 Version 3; Implementation Guide, Release 1.0 http://lists.hl7.org/read/attachment/84028/1/ terminfo_20060307.doc.

Developing Ontology Support for Human Malaria Control Initiatives

Olawande Daramola, Segun Fatumo
Department of Computer and Information Sciences, Covenant University, Ota, Nigeria.

Abstract

Malaria is one of the most common infectious diseases and an enormous public health problem in Sub-Sahara Africa, Asia and parts of America. In this paper, we discuss the development of the Human Malaria Control Ontology (HMCO) which contains general information on Malaria and epidemiological information that can help in the formulation of effective malaria control policies. The HMCO is aimed at providing interoperability support for the knowledge management of malaria control initiatives, and serve as an open semantic web infrastructure for malaria research and treatment.

Introduction

Ontology is a formal explicit representation of the conceptualization of a domain that provides a platform for the sharing and reuse of knowledge across heterogeneous platforms. An ontology contains semantic descriptions of the features of a domain using concepts and relationship abstractions in a way that is readable by both man and machine. In recent times, the use of ontology have gained increasing relevance in the biomedical domain in that it enables researchers to stay abreast of current biomedical knowledge and promotes the understanding of such information. They also facilitate the sharing and reuse of biomedical knowledge across heterogeneous platforms for the delivery of medical services and implementation of health-related policies [1].

Malaria is one of most worrisome vector-borne diseases that affect humans. It is caused by the parasite Plasmodium falciparum. Malaria is endemic in the tropical regions and sub-tropical regions of the world which are mainly in the South East Asia, Middle East Asia, Central and South America, and Sub-Sahara Africa. Millions of malaria cases are reported each year, killing over one million per year in Sub-Sahara Africa[2]. Generally, the control of a vector-borne disease such as malaria pose a critical challenge due to a number of reasons: 1) the complex nature of its transmission which involves three entities, which are the host (human), vector (female anopheles mosquito) and pathogens (Plasmodium specie); 2) the complicated epidemiology through the vector; and 3) social problems (poverty), geography, and resistance of pathogens to insecticides. All of these challenges compel the need to complement existing biomedical approaches of tackling the spread of vector-borne diseases with readily accessible, interoperable, and semantically-rich knowledge management support. This challenge motivated our pursuit of developing an ontology-based support for human malaria control. The Human Malaria Control Ontology (HMCO) contains information on human malaria that can be leveraged for the formulation of human malaria control initiatives in Sub-Sahara Africa. In terms of benefits the HCMO is expected to: 1) provide an interoperable platform for accessing malaria epidemiology information over the web; 2) provide information support for malaria control research and formulation of malaria control policy initiatives; and 3) Create an interoperable platform for the sharing and reuse of knowledge on malaria.

The outline of the rest of this paper is given as follows. In Section 2 an overview of related work on medical ontologies is presented. Section 3 is a short description of the design and implementation of the HMCO. Section 4 is a discussion of the possible application of the HMCO, while the paper is concluded in Section 5 with an outlook of future work.

Related Work

Medical Ontologies have played useful roles in facilitating the re-use, dissemination and sharing of patient information across disparate platforms. Also, they have been used in semantic–based statistical analysis of medical data. Examples of medical ontologies include GALEN[3], UMLS[4], MeSH[5], ON9[6], Tambis[7], The Systematized Nomenclature of Medicine[8,9], Foundational Model of Anatomy[9], MENELAS ontology[9], Gene Ontology[10] and LinKBase[11]. The NBCO's Bio-portal[12] consist of more than 50 bio-ontologies that span several aspects of bio-medicine including diseases, biological processes, plant, human, bio-medical resources etc. However, none of the ontologies in the bio-portal is specifically dedicated to malaria control. The work by Hadzic and Chang[1] was based on providing interoperability support for research in, and diagnosis of human disease using ontology-based approach. A prototype Generic Human Disease Ontology (GenDO) that contains common general information regarding human diseases was created which captured the information in four dimensions. However the dimension of diseases control was not included.

The Infectious Diseases Ontology (IDO)[13] is designed to make infectious diseases-relevant data derived from different sources comparable and computable. It also provides coverage of entities that are common to many infectious diseases. The Vector-borne disease ontology[14] is an ongoing project that is designed to provide an integrated interoperable platform for the sharing and reuse of knowledge about a group of vector-borne diseases in which the MalIDO (Malaria IDO) is a first step. The MalIDO incorporates several information dimensions such as gene models for A.gambie, Anatomy of mosquito, insecticide-resistance, and physiological processes of mosquito. As a contribution our work in the HMCO specifically focuses on the creation of an interoperable platform that gives access to epidemiological information on malaria in Sub-Sahara Africa that can be used for the formulation of malaria control policies.

Description of the HMCO

The HMCO captures information on human malaria in 7 dimensions. These are: (1) Malaria vectors (2) Malaria types, (3) Malaria parasites, (4) Malaria Symptoms, (5) Malaria treatment (prevention, therapy), (6) Epidemiology data on malaria, and (7) Malaria Control. The design of HMCO was based on the Open Biomedical Ontologies (OBO) foundry principles[15], while OBO foundry naming conventions were also adopted significantly in naming its concepts. It was implemented as an OWL ontology using the Protégé 3.4 Ontology tool. The conceptual taxonomy of the HMCO consists of 97 class abstractions that cover the seven dimensions of our interest. Nine disjoint subclasses comprising v*ector, treatment, continent, type, parasite, epidemiology-info, symptom, year_data,* and *control* were modelled as constituents of the superclass h*uman_malaria* using 'belongTo' object property. The subclasses for the three classes: *symptom, treatment, and continent* were modelled as OWL value partitions (viz. each of the classes was represented as comprising disjointed subclasses that cover all known instances of each class). Concepts relationships among classes (concepts) in the HMCO class hierarchy were represented using object property abstractions that define the nature of association between the classes. These include associations between *parasite* and *vector* ('hasVector'), *type* and *symptom* ('hasSymptom'), *type* and *malaria_therapy* ('hasTherapyDrug'), *type* and *malaria_prevention* ('hasPreventionDrug'), *type* and *parasite* ('isCausedby'), *parasite* and *malaria_therapy* ('isCuredby'), *parasite* and *malaria_prevention* ('isPreventedby), *vector* and *continent* ('isFrom), *epidemiology_info* and *year_data* ('hasEpidemydata') etc. The class hierarchy and description of entities in the HMCO is shown in Table 1 (appendix). Also, specific object properties and datatype properties in the HMCO have appropriate cardinality restrictions imposed on them in order to effectively capture the semantics of relationships among the classes in the HMCO. Figure 1 is a view of the class hierarchy of the HMCO.

Application of the HMCO

Currently, the HMCO knowledgebase is being populated with available data (at present we have epidemiological data on malaria from the year 1999 to 2003 for 16 countries in Sub-Sahara Africa), while the data gathering process is still ongoing. At the completion of the first version of the HMCO, it is expected to serve as a web-based repository for accessing epidemiology information on malaria in Sub-Sahara Africa. As a first step to attaining this, a prototype semantic web application has been built that can be used to query the HMCO knowledgebase.

The Java programming language implementation technology was engaged in building the prototype semantic web application using the NetBeans Java IDE. The Web GUI that facilitates client interaction with the HMCO knowledgebase was implemented using Macro Media Flash and Dream Weaver web design tools, and Java Server Pages (JSP). The business logic for querying the HMCO was implemented as an Enterprise Java Beans (EJB) component that is invoked from a Java Servlet class running on Sun Application Web Server 9.0. The EJB makes use of Protégé ontology Java APIs to access the HMCO knowledgebase for information retrieval. The Pellet 1.5 Descriptive Logics (DL) reasoned was used as the OWL DL reasoner[16] to facilitate semantic web reasoning (entailment, subsumption, and ABox reasoning) on the classes and individuals in the HMCO.

Conclusion

In this paper, a description of the Human Malaria Control Ontology (HMCO) has been presented. The HMCO offers as its contribution, an interoperable platform for accessing epidemiological information on malaria as viable knowledge management infrastructure for malaria control policy formulation and research in Sub-Sahara Africa. Though still an ongoing work, a preliminary test of the usability of the HMCO has been undertaken with promising results. Subsequently, the HMCO will incorporate other dimensions of information on malaria by importing relevant ontologies like the IDO and the Vector-borne diseases ontology (particularly the MalIDO) and will be presented for submission to the bio-ontology portal for open access and evaluation.

- human_malaria
  - control
    - drainage
    - mosquito_net
    - insecticide
  - epidemilogy_info
    - Drc_Data; Kenya_Data; Angola_Data; Nigeria_Data; Mozambique_Data; Namibia_Data; Ghana_Data; Sudan_Data; CAR_Data; Mali_Data; Uganda_Data; Malawi_Data; Tanzania_Data; Zambia_Data
  - Parasite
    - P. falciparum
    - P. ovale
    - P. malariae
    - P. vivax
  - Symptom
    - rigor
    - vomiting
    - convulsion
    - haemoglobinuria
    - anemia
    - fever
    - arthralgia
    - shivering
    - abnormal_posturing
    - retina_damage
  - treatment
    - malaria_prevention
      - hydroxyxhloriquine; chloroquine; proguanil; doxycycline; mefloquine; atavoquone- proguanil
    - malaria_therapy
      - artesunate-sulfadoxine-pyrimethamine; artesunate-amodiaquine; artesunate-mefloquine; artemether-lumefantrine; quinine; cotrifazid; sulfadoxine-pyrimethamine; primaquine
  - type
    - severe_malaria
    - chronic_malaria
  - vector
    - A.gambiae; A.freeboni, A.culicifacious; A. fluviatilis; A. minimus; A. phillipinesis; A. stephensi; A. leucosphyrushave
  - Continent
    - Africa; Antarctica; Asia; Australia; Europe; North_America; South_America
  - year_data
    - _1990 ... _2008

Figure 1. A view of the class hierarchy of the HMCO

Acknowledgements

This research was undertaken under the Covenant University Postgraduate Research Award.

References

1. Hadzic M and Chang E. Medical ontologies to support human disease research and control International Journal of Web and Grid Services. 2005; 1:139–150.
2. Snow RW, Guerra CA, Noor AM, Myint HY and Hay SI. (2005). The global distribution of clinical episodes of Plasmodium falciparum malaria. Nature 434 (7030): 214–7.
3. Rector A, Solomon W, Nowlan W and Rush T. A Terminology Server for Medical Language and Medical Information Systems. Methods of Information in Medicine. 1995; 34:147–157.
4. Pisanelli D, Gangemi A and Steve G. An Ontological Analysis of the UMLS Methatesaurus. Proceedings of AMIA Conference, 1998.
5. Stuart N, Aronson A, Doszkocs T, Wilbur J, Bodenreider O, Chang F, Mork J and McCray A. Automated Assignment of Medical Subject Headings. Poster presentation at: AMIA Annual Symp., Washington, DC, 1999.
6. Pisanelli D, Gangemi A and Steve G. WWW-available Conceptual Integration of Medical Terminologies: The ONIONS Experience. Proceedings of AMIA Conference, 1997.
7. Stuckenschmidt H, van Harmelen F, Serafini L, Bouquet P and Giunchiglia F. Using C-OWL for the Alignment and Merging of Medical Ontologies, proceedings of the First International Workshop on Formal Biomedical Knowledge Representation (Co-located with KR 2004), Whistler, Canada 2004.
8. SNOMED International, a division of the College of American Pathologists (CAP) (http://www.snomed.org).
9. Bodenreider O and Burgun A. Biomedical Ontologies, Medical Informatics: Advances in Knowledge Management and Data Mining in Biomedicine. Springer-Verlag. 2005; 211–236.
10. Gene Ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nature Genet. 2000; 25: 25-29.
11. LinKBase: http://www.landcglobal.com/pages/linkbase.php
12. National Center for Biomedical Ontology Bio-Portal: http://www.bioontology.org (accessed April 3, 2009).

13. Infectious Diseases Ontology: http:// www.infectiousdiseaseontology.org (accessed March 3, 2009).
14. Vector-Borne Disease ontology: http://www.infectiousdiseaseontology.org/IDO_ Extensions.html (accessed April 3, 2009).
15. OBO Foundry Principles: http://www.obofoundry.org/crit.shtml (accessed April 3, 2009).
16. Pellet DIG Reasoner: http://pellet.owldl.com/

	Class name	**Description**	**Subclasses**
1	human_malaria	Main ontology superclass	control, epidemilogy_info, parasite, symptom, treatment, type, vector, continent, year_data
2	control	Defines the concepts under malaria control	severe_malaria, chronic_malaria
3	epidemiology_info	Captures information on epidemiology data of 16 countries in Sub-Sahara Africa	Drc_Data, Kenya_Data, Angola_Data, Nigeria_Data, Mozambique_Data, Namibia_Data, Ghana_Data, Sudan_Data, Central African Republic (CAR)_Data, Mali_Data, Uganda_Data, Malawi_Data, Tanzania_Data, Zambia_Data
4	parasite	Defines the different types of human malaria pathogens	P. falciparum, P. ovale, P. malariae, P. vivax
5	symptom	Defines the different types of malaria symptoms as a set of disjointed classes.	rigor, vomiting, convulsion, haemoglobinuria, Anemia, Fever, arthralgia, shivering, abnormal_posturing (children), retina_damage (children)
6	treatment	Defines the types drugs for prevention and treatment of human malaria	malaria_prevention, malaria_therapy
7	type	Defines the types of human malaria	severe_malaria, chronic_malaria
8	vector	Defines the different types of human malaria vectors	A.gambiae, A.freeboni, A.culicifacious, A. fluviatilis, A. minimus, A. phillipinesis, A. stephensi, A leucosphyrushave
9	continent	Defines continental regions to which specific malaria vectors belong. Instances of this class the maps to a property of the vector class	Africa, Antarctica, Asia, Australia, Europe, North_America, South_America
10	year_data	Defines information on malaria endemics on a yearly basis for 19 years.	_1990 ... _2008

Table 1. An Overview of Classes in the HMCO

Towards an Ontological Representation of Resistance: The Case of MRSa

Albert Goldfain[1], Lindsay G. Cowell [2], Barry Smith[3]
[1]Blue Highway, Syracuse, NY, USA
[2]Duke University Medical Center, Durham, NC, USA
[3]University at Buffalo, Buffalo, NY, USA

Abstract

*This paper addresses a family of issues surrounding the biological phenomenon of resistance and its representation in realist ontologies. Resistance terms from various existing ontologies are examined and found to be either overly narrow, inconsistent, or otherwise problematic. We propose a more coherent ontological representation using the antibiotic resistance in Methicillin-Resistant Staphylococcus aureus (*MRSa*) as a case study.*

Introduction: IDO, SaIDO, and MRSa

The phenomenon of resistance is an important feature of biological reality, encompassing phenomena such as the resistance of an individual to specific diseases, the resistance of disorders to specific treatments, and the resistance of certain pathogens to certain drugs. As such, resistance is a phenomenon that needs to be captured in biomedical ontologies.

The Infectious Disease Ontology (IDO) consortium is developing a set of interoperable ontologies that together are intended to provide coverage of the infectious disease domain. At the core of the set is IDO itself, which provides a representation of all of these types of entities, drawn from both the biomedical and the clinical domains that are relevant to infectious diseases in general. Domain-specific extensions (e.g., pathogen-specific extensions) of this core IDO complete the set by providing ontology coverage of entities relevant to specific sub-domains of the infectious disease field. IDO is itself an extension of the Basic Formal Ontology (BFO).

The *Staphylococcus aureus* Infectious Disease Ontology (SaIDO) is an extension of IDO concerning *Staph aureus* (Sa) infection. Sa can be partitioned into two subtypes: Methicillin-Susceptible Sa (MSSa) and Methicillin-Resistant Sa (MRSa). The latter subtype is a defined class that is distinguished by its resistance to methicillin (and other β-lactam antibiotics). Due to its rapid evolution in the face of antibiotic selective pressures, MRSa has become the paradigm of resistance (a so-called "superbug"), and has drawn significant attention from NIAID/NIH, CDC, and biomedical researchers throughout the developed world.

Subtypes of Sa can also be specified by assigning bacterial strains to clonal complexes based on genotypic differences. Variants can differ in their degree of resistance and in the types of drug to which they are resistant, forming a continuum, in terms of which Sa can be (and is) categorized. This provides one powerful reason to produce an ontologically correct representation of resistance.

In this communication, we consider the issues arising from the representation of resistance in realist ontologies and specifically, in IDO. We will focus our attention on the antibiotic resistance of MRSa to methicillin as a case-study.

Ontological Issues Stemming from Resistance

An important principle for realist ontology development is to avoid as far as possible the use of negative differentia (e.g., 'nonphysical', 'not part of the heart') in formulating definitions. This "positivity design principle" enforces the use of terms which capture information about the entities represented in the ontology rather than information about the state of our knowledge at some given time.[1]

At some level, however, resistance seems to require a negative aspect for its description. After all, a continuant is resistant precisely when something does *not* happen. John's resistance to marriage entails a host of processes that do *not* happen (for example, John does not buy an engagement ring, does not get a marriage license, and so forth). In the case of MRSa, resistance to methicillin entails that a process of cell wall formation is not interfered with. The key is that the implicit negativity of resistance is only a semantic feature of the description *at some level*. The biological phenomenon of resistance is manifested at various levels of biological reality: genes, cells and their parts, organisms, and populations. Negative descriptions at a macro-scale here mask the positive and active aspects of resistance at the micro-scale. A comprehensive ontological treatment must, accordingly consider resistance at different levels of granularity.

In BFO-based ontologies, the **lacks** relation can be used to capture negative findings at one scale of biological description while avoiding the problems of using negative predicates or characteristics.[2] In

describing resistance, we will have a need to say that an independent continuant does not exhibit a dependent continuant. As we will see below, this amounts to an independent continuant lacking a certain disposition.

Resistance is referred to by several disciplines: epidemiologists describe the spread of resistance in a population, the medical community speaks of patient resistance to disease and pathogen resistance to drugs, and geneticists make reference to the genes that confer resistance when certain alleles are present. The IDO suite of ontologies must capture all of these discipline-specific aspects of resistance and the relations between them.

Resistance in Existing Ontologies

We surveyed the treatment of resistance in existing ontologies.

Gene Ontology (GO). A general treatment of resistance is outside the scope of the GO, as resistance is not a biological process, molecular function, or cellular component. Within the sub-ontology of biological processes, however, GO contains the term 'response to drug', with synonyms 'drug resistance' and 'drug susceptibility/resistance'.

> [GO:0042493] Response to Drug: A change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of a drug stimulus.

This treatment is inadequate because the narrower term "drug resistance" is made a synonym of the broader term "response to drug". Resistance arises spontaneously as the result of genetic diversification. The presence of the drug provides a fitness advantage to those cells or viral particles that have the resistance conferring gene or mutation, thus they outcompete the susceptible individuals. The resistance is not a direct response to the drug stimulus, although the manifestation of resistance may be a consequence of exposure to the drug. A response to a drug is a *process*, whereas resistance is a *continuant*. This error (although GO usually is very good at preventing this confusion), arises from an inadequate analysis of resistance. Finally, the GO definition of drug resistance seems to hinge on a 'change in state', but cells which do not change state are manifesting a 'response to a drug' just as much as those which do.

NCI Thesaurus. The NCI Thesaurus has the following entry for 'resistance':

> [C19391] Resistance: Natural or acquired mechanisms, functions, activities, or processes exhibited by an organism to maintain immunity to, or to resist the effects of, an antagonistic agent, e.g., pathogenic microorganism, toxin, drug.

The primary problems with this treatment of resistance are that: i) the definition is circular, since it uses 'resist' in defining 'resistance', and ii) the term 'resistance' is a child of "resistance process", making resistance a process, as in the GO, and excluding many types of resistance, because the definition of 'resistance process' is biased towards multicellular organism resistance mediated by host defense mechanisms.

SNOMED-CT. SNOMED-CT contains the entry 'drug resistance (disorder)' with two defining relationships:

> Drug Resistance Is a Drug-Related Disorder
>
> Drug Resistance has Causative Agent (Attribute)
>
> Drug or Medicament.

With a parent term like 'drug-related disorder', it is clear that this definition is given from the perspective of the patient. From the perspective of a pathogen (*qua* organism) or tumor, for example, drug resistance is not a disorder, but rather a benefit. Also, the definition specifies that drug resistance is caused by a drug, but resistance is caused typically by the presence of a gene or mutation. It is only the manifestation of resistance that results from the presence of the drug. Finally, as with other terms in SNOMED, only necessary but not sufficient conditions for drug resistance are provided. Good definitions should spell out both.

Infectious Disease Ontology (IDO). IDO includes the term 'protective resistance', the definition of which attempts to address some of these problems:

> Protective resistance is a disposition that inheres in an organism by virtue of the fact that the organism has a part (e.g., a gene product), the disposition of which is to ensure a physiologic response of a certain degree to a potentially damaging entity *P*, or to prevent the completion of some process caused by *P*, thereby protecting the organism from or mitigating the damaging effects of *P*.

In the next section, we describe the ontological case study that helped lead us to this definition.

Towards a More Robust Ontological Treatment

To better understand the representational demands posed by resistance (and to expose the problems raised by this and similar phenomena from an ontological point of view), it will be useful to go through a detailed example. We choose drug resistance for a single combination of pathogen, antibiotic, and resistance-mechanism types. In this section we will sketch the outlines of a formal

representation of the resistance of MRSa to methicillin as conferred by PBP2a, a penicillin binding protein (PBP) and a product of the gene *mecA*. Both methicillin and penicillin are β-lactam antibiotics and, for the purposes of our formalization, a PBP can be considered to be a *methicillin* binding protein. Chambers[3] gives a concise description of this form of resistance: "[M]ethicillin resistance in staphylococci is due to expression of PBP2a, a novel, low-affinity PBP for which there is no homologue in methicillin-susceptible strains". We formalize this information as a set of triples expressing the relevant ontological relationships. We also include a series of inference rules that would lead a logic-driven reasoner to deduce from the triples that MRSa is resistant to methicillin. Alongside the statistical techniques employed in biology, it will one day be desirable for automated reasoners to compute antibiotic resistance from logical formalizations. Using ontologies as predictive tools will guide treatment decisions and support automated drug discovery.

The terms used in our representation will be derived from IDO, GO, and the Protein Ontology. The relations used are drawn from the OBO Relation Ontology (RO) and its extensions (see http://www.obofoundry.org/ro/). Naïvely, we could introduce a new relation **resistant_to** and represent the entire situation as MRSa **resistant_to** methicillin. However, this would hide the complexity of the mechanisms of resistance working at a smaller scale and elminate many important inferences about resistance. Also, it is important to avoid a proliferation of relations in the OBO Foundry, since restriction to a small set of relations promotes reuse and interoperability of the constituent ontologies.

A more faithful representation requires at least the following components (where **is_a** and **has_part** are used for relations between both continuant and occurent universals):

[1] bacterium **is_a** organism

[2] MRSa **is_a** bacterium

[3] synthesis_of_peptidoglycan **is_a** process and **has_participant** Penicillin_Binding_Protein (PBP)

[4] PBP **has_function_realized_as_process** synthesis_of_peptidoglycan

[5] Bacterial_cell_wall **is_location_of** PBP

[6] Canonically, synthesis_of_peptidoglycan **results_in_development_of** bacterial_cell_wall

[7] formation_of_bacterial_cell_wall **is_a** process

[8] PBP2a **is_a** PBP

[9] methicillin_PBP_binding_process **is_a** binding process that **has_participant**s methicillin and PBP

[10] affinity_to_methicillin **disposition_of** some PBP to undergo a methicillin_PBP_binding_process that is **realized** in the presence of a methicillin.

[11] methicillin_PBP_binding_process **negatively_regulates** synthesis_of_peptidoglycan.

[12] PBP2a **lacks** affinity_to_methicillin

[13] *mecA* **is_a** gene

[14] MRSa **has_part** *mecA*

[15] *mecA* **generically_specifies** PBP2a_production

[16] PBP2a_production **results_in_formation_of** PBP2a

These triples will be used along with several rules of inference and derived facts (labeled *IRn* and D*n* respectively in what follows). For readability, all variables are italicized and initial universal quantifier symbols are suppressed. First, we specify that **is_a** and **has_part** (for both continuants and occurrents) are transitive, allowing us to derive some basic taxonomic facts about the domain:

(IR1) x **is_a** y & y **is_a** $z \rightarrow x$ **is_a** z

(IR2) x **has_part** y & y **has_part** $z \rightarrow x$ **has_part** z

(D1) MRSa **is_a** organism

The parts of an organism are the products of the organism's expressed genes, and these products are located in the appropriate places:

(IR3) (*o* **is_a** organism & *g* **is_a** gene & *o* **has_part** *g* &
g **generically_specifies** *proc* &
proc **results_in_formation_of** *prod* &
o **has_part** *locp* & *locp* **is_location_of** *prod*) →
o **has_part** *prod* **located_in** *locp*

(D2) MRSa **has_part** PBP2a **located_in** bacterial_cell_wall

The inference rule (IR3) makes a few simplifying assumptions. Since not all genes are expressed, we are only modeling the situation in which *g* is an expressed gene. We also assume that the process *proc* leading to *prod* is active, and that the single gene *g* generically specifies *proc* (rather than a set of genes).

If a continuant lacks a disposition to undergo a process in some situation, and that process negatively regulates a second process which has the continuant as a participant, then the continuant participates in the second process in that situation:

(IR4) *p* **lacks** disposition to undergo *proc1* **realized** in situation *s*
& *proc1* **negatively_regulates** *proc2*
& *proc2* **has_participant** *p* →
In situation *s*, *p* **participates_in** *proc2*

(D3) In the presence of methicillin, PBP2a **participates_in** synthesis_of_peptidoglycan.

This lack of a disposition (i.e., the affinity to methicillin) has a categorical basis in the fact that methicillin binds to PBPs and prevents them from carrying out their function. However, PBP2a lacks this affinity, so the presence of methicillin does not

prevent the essential sub-processes of cell-wall construction in MRSa.

If an organism has a continuant as a part and that part participates in a process in some situation, then the process unfolds in the organism in that situation. Finally, if a process unfolds in an organism in some situation and the process results in the development of a continuant which (canonically) is a part of the organism, then the organism has the continuant as a part in that situation.

(IR5) In situation *s*, *p1* **participates_in** *proc* & *p1* **located_in** *p2* & o has_part p2 → *proc* **unfolds_in** *o* in situation *s*.
(D4) synthesis_of_peptidoglycan **unfolds_in** MRSa in the presence of methicillin.
(IR6) In situation *s*, *proc* **unfolds_in** *o* & Canonically, *proc* **results_in_development_of** *p* → *p* **part_of** *o* in situation *s*
(D5) Bacterial_cell_wall **part_of** MRSa in the presence of methicillin.

The canonical cell wall is a rigid configuration of peptidoglycan. From the perspective of MRSa, the canonical cell wall is a healthy one. The assertion (D5) captures the active, and thus positive, microphysical side of the resistance coin.

However the chain of reasoning here presents a puzzle. What does the lack of a disposition in (IR4) amount to? Consider the following pair:

(A) Continuant C lacks disposition D to undergo process P in situation S

(B) Continuant C undergoes P in a situation of type S.

Both (A) and (B) can be true at the same time. In fact the conjunction of (A) and (B) implies that (B) happens for a non-dispositional reason (i.e., (B) is not, in the corresponding case, a manifestation of the disposition D). Even if John lacks the disposition to feel hungry when in the presence of sushi, he may still feel hungry in such a situation because he has been fasting for three days. We need a way to say that PBP2a *necessarily* lacks affinity to methicillin in order to permit the relevant cell-wall formation.

Mereological Issues

If we take resistance to be a specifically dependant continuant that inheres in an independent continuant, then we must still answer some mereological questions: Is the resistance of PBP2a (i.e., of a part) identical to the resistance of the cell (i.e., of the including whole)? Furthermore, is cell resistance identical to the resistance of a portion of tissue in which the cell resides or the containing host organism or, for that matter, of the containing population? The ontology of resistance must address which scales of biological reality resistant continuants occupy, and the identity of resistance across scales.

Another issue that should be addressed at different scales of biological reality is the way in which facts at each scale are used to *explain* the phenomenon of resistance. At the genetic scale, MRSa having *mecA* and MSSa lacking *mecA* are explanatory. At the cellular level (D5) is explanatory.

Conclusion

We have seen that resistance is an important multi-scale and multi-domain phenomenon, often with a one-to-many relationship between a resistant organism and the underlying mechanisms of resistance. Several desiderata for an ontological representation were found lacking in existing ontologies. Our preliminary formalization of resistance honors both a *positivity design principle* and a *principle of non-proliferation of relations*, both of which are sound principles for the design of effective ontologies. Some puzzles remain (e.g., an account for the lack of a disposition), but further study of resistance will have great benefits for biomedical ontologies.

Acknowledgements

This work was funded by the National Institutes of Health through Grant R01 AI 77706-01. Smith's contributions were also funded through the NIH Roadmap for Medical Research Grant 1 U 54 HG004028 (National Center for Biomedical Ontology).

References

1. Spear A. "Ontology for the twenty-first century: An introduction with recommendations". 2006. *http://www.ifomis.org/bfo/manual.pdf*
2. Ceusters W, Elkin P and Smith B. "Negative findings in electronic health records and biomedical ontologies: A realist approach". *Int J Med Inform*. 2007;76(3):s326–s333.
3. Chambers H. "Penicillin-binding protein–mediated resistance in pneumococci and staphylococci". *J of Infectious Disease*. 1999; 179(2):S353–359.

Providing a Realist Perspective on the eyeGENE Database System

Werner Ceusters
New York State Center of Excellence in Bioinformatics & Life Sciences, Buffalo, NY, USA

Abstract

One of the achievements of the eyeGENE Network is a repository of DNA samples of patients with inherited eye diseases and an associated database that tracks key elements of phenotype and genotype information for each patient. Although its database structure serves its direct research needs, eyeGENE has set a goal of enhancing this structure to become increasingly well integrated with medical information standards over time. This goal should be achieved by ensuring semantic interoperability with other information systems but without adopting the incoherencies and inconsistencies found in available biomedical standards. Therefore, eyeGENE's current pragmatic perspective with focus on data and information, rather than what the information is about, should shift to a realism-based perspective that includes also the portion of reality described, and the competing opinions that clinicians may hold about it. An analysis of eyeGENE's database structure and user interfaces suggests that such a transition is possible indeed.

Introduction

The eyeGENE database is a repository of genotype and phenotype information of patients with inherited eye diseases collected through the National Ophthalmic Disease Genotyping Network, an initiative launched by the National Eye Institute in 2006.[1] The database design used the innovate approach of defining the structure of phenotype information by means of metadata, so that new diagnoses and questions concerning clinical findings could be added or modified by the eyeGENE administrator at any time. The goal was to allow collection of a large number of samples with a minimal data entry burden to the clinician and genetics testing labs, and to provide an easy overview of key data for a researcher who may wish to study details of an attached eye image or otherwise study the patient's data in more depth.

To avoid this system becoming yet another information silo, eyeGENE set a further goal of integrating the eyeGENE data with applicable medical information standards over time. It can be expected that adopting currently available and emerging medical information standards will provide an additional layer of benefits in more easily collecting, sharing and analyzing data in the future.

As a first step, an extensive study was performed on existing and emerging standards relevant to clinical research data, including the identification of gaps and overlaps.[2] This study revealed that this goal is confounded by deficiencies in many standards pertinent to clinical data registration, which suffer from reductionist views on reality which are constrained by what can be seen through the lenses of either information systems[3] or terminologies and ontologies that adhere to what is called '*concept representation*'.[4] Without appropriate remediation, semantic interoperability between systems adhering to such standards will be on a less than fully logically sound foundation and will suffer limitations over time.

Objectives

As witnessed by the success of the OBO-Foundry a growing number of scholars adheres to a realist view on reality and to an implementation along these lines both in ontologies[5] and information systems.[6] The goal of the study reported on here was (1) to understand the type of view embedded in the eyeGENE database and (2) in case this view would differ from the realist one, to propose a migration path towards the latter.

Material and Methods

We studied the available documentation about eyeGENE's core medical information, including parts of its information model and user interfaces. We looked at some of the clinical questions (and corresponding possible-answer sets) that are asked to eyeGENE users when they enter data in the system, as well as to system generated reports about lab procedures performed on genes. We did not have access to a data-dictionary with data-definitions and corresponding business rules.

We checked in the first place for design choices in the system that would lead the information to be collected not to match with the corresponding structure of reality, the latter under the realist view consisting of:

1. first-order reality, which includes entities such as specific patients, their relatives, the disorders they are suffering from, the lab tests that have been conducted, and so forth;
2. second-order reality, including, for instance, interpretations and opinions on the side of clinicians, including hypotheses and diagnoses;

thus being *about* entities in first-order reality, although not accessible to third parties without additional third-order references;

3. third-order reality, which is composed of information about first- or second-order reality, examples being entries in information systems such as the eyeGENE database.

We also checked for structural and functional issues in eyeGENE that in absence of sufficient background information for disambiguation would lead to difficulties in interpreting data once entered.

Results and Discussion

We found that to meet its goal of future integration with high quality medical information systems over time, the pragmatic design approach initially followed by the eyeGENE developers should be transformed to remove current limitations of (1) conflating the three levels of reality as described above, and (2) not representing faithfully the relevant portions of reality at each level.

An example of a non-faithful representation of first-order reality is eyeGENE's treatment of the patient's demographic information: the user interface lists a number of data entry fields, amongst which the postal code, as 'required'. A motivation for including 'required fields' in data entry forms is to have data as complete as possible. Sometimes, however, as is the case here, these constraints violate what is the case in reality: many countries do not use postal codes at all. If eyeGENE's realm is not limited to patients living in the US, such constraints pose a problem as they force the user to enter fake data, or, when the latter is against the user's principles, prevent him from entering data at all. A strategy often applied is to allow for various sorts of null-values, but that changes the semantics of the data field drastically: it would then not always contain strings that denote postal codes, but strings that denote, for example, that there are no postal codes in the corresponding country, or that the postal code is not known by the user. Confusions of this sort are, for example, abundantly present in HL7.[3]

Another example of a required field in eyeGENE is 'gender' with the two possible values 'female' and 'male'. This might seem to be consistent with first-order reality as each human being can be expected to be either 'male' or 'female'. However, for each of the three possible interpretations of what the word 'gender' here might stand for, matters are not that obvious. *Phenotypic gender* is not either male or female in hermaphrodites, *genotypic gender* comes in many more flavors, while, finally, *administrative gender*, depending on the community in which it is defined, is based not only on scientific grounds but also on political, ethical, and even religious considerations, thereby giving rise to oddities to the effect that the different treatments of the right of gender self-identification makes it possible that the same person has a different administrative gender in Australia and in the US.[7]

The eyeGene database contains many examples not of unfaithful representation of reality but rather of undocumented reductionism. It allows, for instance, the eye fundus to be described as being normal or exhibiting any combination of four types of anomalies. By 'undocumented', we mean that it is left unspecified whether these four types are the *only* possible types in reality, or whether there are many more possibilities of a sort which are not relevant for the purposes for which eyeGENE has been designed, and therefore are not offered as additional alternatives.

An example of a conflation of first-order and second-order reality is in the registration of diagnoses. Clinicians are requested to provide the date of examination and then to select one or more types of diagnoses out of a list of 21. Based on that information and with the goal to collect further data about signs and symptoms, clinical data entry forms specific for each type of diagnosis are generated. These forms are composed out of building blocks some of which, for example to provide details about the patient's 'best corrected visual acuity', can appear in forms related to more than one diagnosis. Once data are provided in the context of one diagnosis, the same data re-appear in the form corresponding to another diagnosis. This setup, although being very pragmatic – it frees the clinician from entering the same data more than once – leads to ambiguities from an ontological perspective.

One ambiguity arises from the mere fact of entering diagnoses without identifying the corresponding disorder *about which* that diagnosis is a diagnosis: disorders are first-order entities on the side of the patient while diagnoses are second-order entities on the side of, for instance, the clinician.[8] Disorders and diagnoses live totally different lives: patients may have a disorder without any diagnosis being made; clinicians may come to one diagnosis while the patient may have either two or more disorders or no disorder at all; distinct clinicians may bring forward different diagnoses for the one disorder the patient has; a clinician may change his diagnosis over time, while the disorder does not change at all, and so forth. The problem becomes obvious when more than one clinical examination form is entered: in absence of identifiers for the disorder, it is not possible to deduce formally in case a diagnosis entered on an earlier form is different from the diagnosis on a later form whether the difference is because the earlier diagnosis is revised, whether there is a second disorder involved, or, if distinct clinicians

entered the forms whether there is disagreement about the correct diagnosis.

Another ambiguity, when multiple diagnoses are specified, is to what the individual clinical signs relate. Although clinical signs provide evidence in favor or against certain diagnoses, a particular clinical sign in some patient is not related to any diagnoses entertained for that patient, but rather to at least one disorder from which that patient suffers.

Recommendations

It is no surprise that the information model of the eyeGENE database exhibits the sorts of mismatches with reality just described: to our best knowledge, ***all*** information systems designed according to the state of the art in information modeling suffer from these incoherencies because of at least two misconceptions.

One is the *erroneous assumption of inherent classification* adhered to in many database design circles according to which entities can be referred to only as instances of pre-specified classes.[9] Under the realist view, in contrast, the position is defended that in information systems entities should exclusively be referred to by means of globally unique and singular identifiers.[6] These identifiers can then to be used in descriptions of various sorts indicating, for instance, what universals are instantiated by the entity referred to, what terms from a terminology or concept-based ontology apply to it, or how the entity relates to other entities.

The other misconception is the tyranny of the use case, what leads some to argue that '*if most people wrongly believe that crocodiles are a kind of mammal, then most people would find it easier to locate information about crocodiles if it were located in a mammals grouping, rather than where it factually belonged*'.[10]

Of course, the incoherencies of the information model and business rules as compared to what is the case in reality are not relevant to the original goals for which eyeGENE has been designed. But they do become a problem when the data have to be pooled with data coming from other information systems that describe partially or in total the same domain from a different perspective and are collected for another purpose. In that case, the second system, if designed following prevailing approaches, will also contain incoherencies with respect to reality, but in different ways than eyeGENE. A comparative analysis of the underlying information models may reveal areas where they are in agreement and other places where they cannot both be correct. But in absence of an external benchmark, there is no means to assess which one is right, not even when both models are in agreement because they both might have it wrong in the same way.

We argue that reality should function as that benchmark, and that realism-based ontology provides the means to reach that goal in similar ways as it is increasingly and successfully used for quality assurance in biomedical terminologies and ontologies.[5] The reason is that no portion of reality depends on the information used to describe it or on the purposes for which such information is collected. This is not to say that such information does not contribute to the evolution of reality at all. On the contrary, as soon as it is generated, that information is part of reality itself (level 3), and so is the system used to manage it. Therefore any attempt to make such system, in our case the eyeGENE database system, coherent with respect to reality, should acknowledge the priorities and objectives that have been taken into account at design time. If, for instance, through realism-based analysis one discovers a reductionist approach (e.g. the eye fundus description described earlier), it would be a bad idea to bother the users of the eyeGENE database with a more complex interface that does not bring them advantage in any way, even if it would help secondary users of the data.

The right way forward, so we argue, is by mapping the information model of eyeGENE to a domain model that itself is not reductionist in nature. Reductionist models are typically created when UML is used as this language forces reality to be viewed through the eyes of an information system, using a (partially graphic) vocabulary which is inadequate to describe reality faithfully. The HL7 RIM is the most dramatic example, dramatic because its acceptance as ISO standard gives it an unjustified aura of excellence.[11]

Note that we see no harm in using an existing information model to scope the corresponding domain model. The procedure, in the context of eyeGENE, would be to study each of its tables, data fields and associated allowed values, as well as any hard- or soft-coded business rules that restrict data-input, with the following goals: (1) to assess what (type of) entity in reality would be denoted by any data instance, (2) to represent how these entities in reality relate to each other as well as to other ontologically relevant entities that are not explicitly addressed in the information model, this being the domain model proper, and (3) to describe formally how the information model has to be interpreted in terms of the domain model. The latter can then be used to inform third party systems with which the eyeGENE database system would exchange data about the implicit restrictions in eyeGENE. It can also be used to identify issues that must be resolved in further releases.

As an example, eyeGENE's information model relates a *PatientDiagnosis* to (1) a *ClinicalEncounter* which itself is related to a *Patient* and a

PhysicianPerson and (2) a *Diagnosis*. The eyeGENE database system limits the latter to 21 types, however, upon closer inspection, not to types of diagnoses, but to types of diseases. The domain model would tell us that there are of course many more types of diseases. The interpretation model would then contain statements clarifying this distinction. With respect to (1), the interpretation model could clarify, for instance, whether the date of the *ClinicalEncounter* is the date that the diagnosis was made, and that this by itself would not allow inferences to be made about when the disease started. To some extent, eyeGENE users can clarify such issues in free text, but this cannot be used for automated processing.

No new formalism is required to achieve such integration. The same sort of bridging axioms that are commonly written to map or merge concept-based ontologies[12], can be applied to explain eyeGENE's information model in terms of the domain model.

Conclusion

The eyeGENE database system is successfully in use since July 2006 and processes 35 samples per month. It is foreseen that this number will grow to 100 by end 2009. To most optimally fulfill its ambitious goal of integration with high quality medical information systems in future developments, the eyeGENE database system can become a model of fulfilling a stated objective in the NIH roadmap to '*require new ways to organize how clinical research information is recorded, new standards for clinical research protocols, modern information technology*'. One expectation, in the context of the patient profile, is that at some future time relevant phenotypic data can be automatically extracted from an electronic medical record using a standard in widespread use. At that point, a larger set of base patient data in more specific detail would be practical to collect. Realism-based ontology combined with adequate identification and reference of entities at each level of reality is one new way that can be explored to turn these data into knowledge.

Acknowledgements

With thanks to David Scheim of the National Eye Institute for many useful comments on the first version of this paper.

References

1. National Eye Institute. eyeGENE – National Ophthalmic Disease Genotyping Network. Jan 2009; http://www.nei.nih.gov/resources/eyegene.asp. Accessed January 26, 2009.
2. Rudnicki R and Ceusters W. *Emerging medical information standards as applicable to clinical research data: A study performed in the context of the project 'Exploring eyeGENE, an International Genotype / Phenotype Database, from a Bioinformatics Perspective'*. Buffalo NY: NYS Center of Excellence in Bioinformatics & Life Sciences; July 16, 2008.
3. Ceusters W and Smith B. What do identifiers in HL7 identify? An essay in the ontology of identity. In: Okada M and Smith B (eds.) Interdisciplinary Ontology; Proceedings of the Second Interdisciplinary Ontology Meeting (*InterOntology 2009*), Tokyo, Japan, February 28 – March 1, 2009;:77–86.
4. Smith B. From Concepts to Clinical Reality: An Essay on the Benchmarking of Biomedical Terminologies. *Journal of Biomedical Informatics.* 2006;39(3):288–298.
5. Smith B, Ashburner M, Ceusters W, *et al*. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology.* 2007;25:1251–1255.
6. Ceusters W and Manzoor S. How to track Absolutely Everyting? In: Obrst L, Ceusters W and Janssen T (eds.) *Ontologies for Intelligence*. Amsterdam: IOS Press; 2009 (in press).
7. Milton SK. Top-Level ontology: The problem with naturalism. In: Guarino N (ed.) *Formal Ontology in Information Systems*. Vol 85–94. Amsterdam, The Netherlands: IOS Press; 1998.
8. Scheuermann RH, Ceusters W and Smith B. Toward an Ontological Treatment of Disease and Diagnosis. *Proceedings of the 2009 American Medical Informatics Association's Summit on Translational Bioinformatics*: San Francisco, California, March 15–17, 2009;: 116–120.
9. Parsons J and Wand Y. Emancipating Instances from the Tyranny of Classes in Information Modeling. *ACM Transactions on Database Systems.* June 2000 2000;25(2):228–268.
10. Huhns MN and Stephens LM. Semantic Bridging of Independent Enterprise Ontologies. In: Kosanke K (ed.) Enterprise Inter- and Intra-Organizational Integration: Building International Consensus. Boston, MA: Kluwer Academic Publishers; 2002:83–90.
11. Aerts J. Ten good reasons why an HL7-XML message is not always the best solution as a format for a CDISC standard. February 10, 2009; http://www.xml4pharma.com/HL7-XML/HL7-XML_for_CDISC_Standards.pdf.
12. Özçep Ö. Towards principles for ontology integration. In: Eschenbach C and Gruninger M. (eds.) Formal Ontology in Information Systems, IOS Press, Amsterdam, 2008;:137–150.

Cross-Product Extensions of the Gene Ontology

Christopher J. Mungall[1], Michael Bada[2], Tanya Z. Berardini[3], Jennifer Deegan[4], Amelia Ireland[1,4], Midori A. Harris[4], David P. Hill[5], Jane Lomax[4]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA, USA; [2]University of Colorado Denver, Department of Pharmacology, Aurora, CO, USA; [3]Carnegie Institute for Science, Stanford, CA, USA; [4]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK; [5]The Jackson Laboratory, Bar Harbor, ME, USA

Abstract

The Gene Ontology is being normalized and extended to include computable logical definitions. These definitions are partitioned into mutually exclusive cross-product sets, many of which referenc=e other OBO Foundry ontologies. The results can be used to reason over the ontology, and to make cross-ontology queries.

Introduction

The Gene Ontology (GO)[1] was conceived of as a means of providing structured annotations for genes and gene products, in terms of molecular function (MF), biological process (BP) and cellular component (CC). The current version of the GO has nearly 27,000 terms and 47,000 relationships. As the GO evolves, the relational graph becomes more tangled, which poses a problem for ontology maintenance, correctness and visualization. It has long been recognized that a normalized approach to ontology development helps with re-use, maintainability and evolution[2,3,4]. The OBO Foundry[5] was initiated in part to provide a means of normalizing the GO, such that for example the GO definition of "oocyte differentiation" could reference the term "oocyte" in the OBO Cell ontology (CL), and an automated reasoner tool could be used to classify this as a kind of "germ cell differentiation", based on the CL classification. This is also an example of a 're-use' pattern, common in software engineering.

Almost all of the terms in the GO have textual definitions, crafted for the human users of the GO. When these textual definitions are rendered in a computable from, we can leverage reasoner technology to automate the more tedious and error-prone aspects of ontology maintenance. We can also use these computable definitions to make cross-ontology queries and better visualize the ontology.

Logical Definitions and Cross Products

We provide computable logical definitions for terms using genus-differentia constructs, of the form "an X is a G that D". Here X is the term we are defining, G is the genus (more general term), and D is the differentia, a collection of characteristics that serve to discriminate instances of X from other instances of G. The differentiae are specified as relationships to other terms, using relations from the Relations Ontology[6]. In OBO Format (the native means of representing the GO) these are specified using `intersection_of` tags, which list the necessary and sufficient conditions for a term. For example:

```
[Term]
id: GO:0032543
name: mitochondrial translation
intersection_of: GO:0006412 ! translation
intersection_of: occurs_in GO:0005739 !
mitochondrion
```

In OWL Manchester Syntax, this is written as an equivalence axiom between the class mitochondrial translation and the description **`translation`** `and` **`occurs_in`** `some` **`mitochondrion`**.

In the example above, the logical definition for the process references a GO cellular component term. Often we will want to reference other OBO ontologies, and this introduces multiple dependencies. We therefore partition the full set of logical definitions for GO into cross-product mapping. A cross-product of two ontologies A x B is the set of biologically meaningful terms that can be constructed by extending A using terms from B as differentia. The GO term in the example above would be mapped to a definition in the BP x CC cross product.

Each cross-product mapping is maintained as an individual resource, independent of the others (see Table 1). Currently they are optional add-ons to the GO. We distinguish between intra-GO cross products and inter-GO cross products, the latter consisting of logical definitions that reference an OBO ontology not under the management of the Gene Ontology Consortium.

A subset of the intra-GO cross products are the self-cross products: terms that can be defined solely by using terms in the same ontology.

Biological processes

The **BP x CC** cross-product set includes definitions for biological process terms that have cellular

component terms as differentia. Sometimes we need to specify the subcellular location in which a process occurs, in which case we use an *occurs_in* relation. Sometimes we are describing the output of a process, such as when a cellular component is assembled or disassembled. The GO also has a rich set of subcellular transport terms, in which case the logical definition needs to be precise about the origin, destination and route of the transported entity.

Many BP terms can be defined using a BP term in the differentia. For example, the different phases of the cell cycle can be subtyped according to whether they are part of mitosis or meiosis. These definitions are grouped into the **BP x BP** set.

GO includes 3 broad categories of regulatory processes – regulation of molecular function, regulation of biological process, and regulation of biological quality – these comprise 3 distinct cross-product sets. The first two are intra-GO; the latter references terms from the PATO ontology of biological qualities[7], together with anatomical ontologies.

The cross products make use of 3 new relations introduced into GO – *regulates*, *negatively_regulates* and *positively_regulates*.

We have created a separate cross-product set for the more complex multi-organism interaction regulation terms. The logical definitions we provide here are necessarily a simplification, as we must go beyond the current expressive capabilities of OBO or OWL in order to represent inter-organism interactions.

Anatomy

GO has many terms describing development at the cellular and gross anatomical level. There are also non-development terms that nevertheless reference types of anatomical entity – for example, "muscle contraction".

We use the species-neutral OBO Cell ontology (CL)[8] for defining terms such as "oocyte differentiation" in the **BP x CL** set. Gross anatomy proves more of a challenge because the main OBO gross anatomy ontologies are specific to a species or taxon. We therefore extracted the implicit anatomical ontology embedded in the GO and used this together with alignments to existing anatomy ontologies to seed a multi-species anatomy ontology called Uberon, which is used in the definition of terms such as "muscle contraction". These definitions are part of the **BP x Uberon** set. Uberon covers only animals; plant development terms are in **BP x PO** (plant anatomy ontology). There are also individual species-species extensions such as **BP x Fly_anatomy**.

Molecules and Proteins

Molecular and chemical entities are represented in the CHEBI ontology[9], with proteins represented in PRO[10]. We use these in 3 cross-product sets, **{BP,MF} x CHEBI**[11] and **BP x PRO**. The Protein Ontology is still relatively new, so this last set is currently relatively small. We also intend to work with the PRO curators to make a **CC x PRO** set.

Cellular Components

Many of the terms in CC can be assigned logical definitions based on parthood relations to other components – for example, "nuclear chromosome" is a chromosome that is *part_of* a nucleus. For other definitions in **CC x CC**, we introduce additional spatial relations, such as *surrounds* and *surrounded by*.

The GO CC ontology has many terms representing complexes, some of which are defined by their constituent parts, others by function. The latter have logical definitions in the **CC x MF** cross-product.

Some cell component terms are differentiated by the cell type of which they are a part – for example, a sarcoplasm *is a* cytoplasm that is *part_of* a muscle cell. We map GO terms such as sarcoplasm to the BP x CL set, most of which use the *part_of* relation. For others, such as neuromuscular junction we use adjacency relations.

Reasoning

The current set of logical definitions can be used by a variety of different reasoners. We use the OBO-Edit[12] reasoner, because it is integrated within the normal editing environment for the GO, and provides incremental reasoning support.

We have not found any reasoner that is capable of reasoning over the union of the GO plus all cross-product sets plus all referenced ontologies. However, we are able to reason over individual cross-product sets and their referenced ontologies individually.

We use reasoning primarily for ontology maintenance, to compute and check the subsumption hierarchy. The GO regulation hierarchy in particular has benefited from this work, with over 2000 missing links added to GO, which could potentially improve the results of term enrichment analyses. We use the reasoner in what we call 'repair mode' – we invoke the reasoner to spot mistakes and fill in missing links in the ontology, always asserting links that can be automatically computed. This ensures that editors can edit the ontology without invoking the reasoner over the union of all logical definitions. This stands in contrast to how the reasoner is used in SO and the Fly anatomy ontology. We also use the reasoner to make inferences about the source ontologies[13].

We are still exploring uses of the cross-product sets beyond ontology construction and maintenance. This includes improved visualization, enhancing term enrichment analyses, annotation inferences and using the CHEBI cross-products to harmonize pathway database representations and GO metabolic processes.

Post-Composition

The GO does not pre-compose terms for all biologically meaningful compositions of terms, as this would lead to a large, unwieldy ontology. The guiding principle is to generate compositional terms where the differentia is important to the biology. We are simultaneously exploring an approach whereby annotators can extend GO terms on-the-fly, i.e. selecting compositions from the cross-product at annotation time. For example, an annotator can select the GO term 'mitochondrial membrane' for a cellular component annotation and extend this using a differentia '*part_of* Purkinje cell', with the differentia term coming from CL. This is logically equivalent to annotating to a term 'mitochondrial membrane of Purkinje cell', but avoids bloating the ontology with the full set of biologically instantiable terms in the CC x CL cross-product.

		XP Name		Size	Examples
Intra-GO			* Biological process	606	**S phase of mitotic cell cycle = S phase** and *part_of* **mitosis**
		regulation	Biological process X self (regulates)	3529	**Regulation of neuroblast proliferation = biological regulation** and *regulates* **neuroblast proliferation**
			Biological process X self (multi-organism)	374	**modulation of intracellular transport in other organism during symbiotic interaction = interspecies interaction between organisms** and *regulates* **intracellular transport** and *during* **symbiosis** and *regulates_process_in* **external organism**
			Biological process X MF (regulates)	201	**Regulation of protein kinase activity = biological regulation** and *regulates* **protein kinase activity**
		Biological process X cellular component		476	**Mitochondrial translation = translation** and *occurs_in* **mitochondrion**
		Biological process X SO		61	**group I intron catabolic process = catabolic process** and *has_input* **group I intron**
		* Cellular component X self		682	**Acrosomal membrane = membrane** and *surrounds* **acrosome**
		Cellular component X molecular function		173	**histone deacetylase complex = protein complex** and *has_function* **histone deacetylase activity**
		* Molecular function X self (regulates)		104	**Lipase activator activity = molecular function** and *regulates* **lipase activity**
		Molecular function X cellular component		48	**Microtubule motor activity = motor activity** and *results_in_movement_along* **microtubule**
Inter-GO	Anatomy	Biological process X cell		544	**Oocyte differentiation = cell differentiation** and *results_in_acquisition_of_features_of* **oocyte**
		Biological process X Uberon		583	**Neural plate formation = anatomical structure formation** and *results_in_formation_of* **neural plate**
		Biological process X quality {X anatomy}		31	**Regulation of cell volume = biological regulation** and *regulates* **(volume** and *quality_of* **cell)**
		Molecular function X Uberon		9	**Structural constituent of bone = structural molecule activity** and *inheres_in* **bone**
		Cellular component X cell		28	**Neuromuscular junction = synpase** and *adjacent_to* **motor neuron axon** and *adjacent_to* **contractile fiber**
	Molecule	CHEBI	Biological process X CHEBI	3077	**L-cysteine catabolic process to taurine = catabolic process** and *has_input* **L-cysteine** and *has_output* **taurine**
			Molecular function X CHEBI	315	**nitrate reductase activity = oxidoreductase activity** and *reduces* **nitrate**
		PRO	Biological process X PRO	37	**Interleukin-1 biosynthesis = biosynthetic process** and *has_output* **interleukin-1**

Table 1. GO logical definitions are partitioned into mutually exclusive cross-product sets. Examples are shown from each of the sets. The second column shows the number of existing GO terms that have been mapped to a logical definition in each set. Asterisks (*) denote self cross-products. In total 10878 terms have been mapped, 41% of all terms in the ontology.

Conclusions

The extended collection of cross-product resources described here represents a significant advance in the evolution of the GO and its integration with other OBO ontologies. The use of these logical definitions, in conjunction with a reasoner has substantially increased the quality of the GO and eased the more prosaic aspects of ontology maintenance. We are still exploring application beyond the ontology itself.

This work also highlights the importance and necessity of the OBO Foundry effort, particularly with respect to efforts to create single orthogonal well-partitioned ontologies each representing a distinct domain of biology.

Methods and Availability

In contrast to some ontology development efforts, in which computable definitions are assigned when terms are created, we have been working retrospectively, constructing logical descriptions for pre-existing terms. To help us with this task we use Obol[14], which heuristically generates proposed logical definitions based using ontology-specific grammars. Ontology editors then vet the definitions, often substantially.

The full extended GO can be obtained on the GO wiki: http://wiki.geneontology.org/index.php/Category:Cross_Products

Comments and contributions are welcome.

Acknowledgments

This work is supported by the NHGRI, via the Gene Ontology Consortium, HG002273.

References

1. Ashburner M, Ball CA, Blake J, Butler H, Cherry J, Corradi J, Dolinski K, Eppig J, Harris M, Hill D, Lewis S, Marshall B, Mungall C, Reiser L, Rhee S, Richardson J, Richter J, Ringwald M, Rubin G, Sherlock G and Yoon J. Creating the gene ontology resource: Design and implementation. *Genome Res,* 2001, *11*, 1425–1433.
2. Hill DP, Blake JA, Richardson JE and Ringwald M. Extension and integration of the gene ontology (GO): Combining GO vocabularies with external vocabularies. *Genome Res,* 2002, *12*, 1982–91.
3. Wroe CJ, Stevens R, Goble CA and Ashburner M. A methodology to migrate the gene ontology to a description logic environment using DAML+OIL. *Pac Symp Biocomput,* 2003, 624–35.
4. Rector AL. Modularisation of domain ontologies implemented in description logics and related formalisms including OWL. *Proceedings of the 2nd international conference on Knowledge capture, ACM,* 2003, 121–128.
5. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, Consortium TO, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone S-A, Scheuermann RH, Shah N, Whetzel PL and Lewis S. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol,* 2007, *25*, 1251–1255.
6. Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL and Rosse C. Relations in biomedical ontologies. *Genome biology* 2005, 6(5):R46.
7. Gkoutos GV, Green EC, Mallon AM, Hancock JM and Davidson D. Using ontologies to describe mouse phenotypes. *Genome Biol,* 2005, *6*, R8.
8. Bard J, Rhee SY and Ashburner M. An ontology for cell types. *Genome Biol,* 2005, *6*, R21.
9. Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcantara R, Darsow M, Guedj M and Ashburner M. ChEBI: A database and ontology for chemical entities of biological interest. *Nucleic acids research* 2008, 36(Database issue):D344–350.
10. Natale DA, Arighi CN, Barker WC, Blake J, Chang TC, Hu Z, Liu H, Smith B and Wu CH. Framework for a protein ontology. *BMC bioinformatics* 2007, 8 Suppl 9:S1.
11. Bada M and Hunter L. Identification of OBO nonalignments and its implications for OBO enrichment. *Bioinformatics,* 2008, *24*, 1448–1455.
12. Day-Richter J, Harris MA, Haendel M and Lewis S. OBO-Edit--an ontology editor for biologists. *Bioinformatics,* 2007, *23*, 2198–2200.
13. Bada M, Mungall C and Hunter L. A Call for an Abductive Reasoning Feature in OWL-Reasoning Tools toward Ontology Quality Control. *5th OWL Experiences and Directions Workshop (OWLED 2008),* 2008.
14. Mungall CJ. Obol: Integrating Language and Meaning in Bio-Ontologies. *Comparative and Functional Genomics,* 2004, *5(7)*, 509–520.

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 · Buffalo, New York, USA

Automated Annotation-Based Bio-Ontology Alignment with Structural Validation

Cliff A Joslyn*, Bob Baddeley*, Judith Blake†, Carol Bult†, Mary Dolan†, Rick Riensche*, Karin Rodland*, Antonio Sanfilippo*, Amanda White*

*Pacific Northwest National Laboratory, Richland, WA, USA
†The Jackson Laboratory, Bar Harbor, ME, USA

Abstract

We outline the structure of an automated process to both align multiple bio-ontologies in terms of their genomic co-annotations, and then to measure the structural quality of that alignment. We illustrate the method with a genomic analysis of 70 genes implicated in lung disease against the Gene Ontology.

Introduction

Ontologies are commonly aligned based on similar annotations[3, 7], requiring validation of the quality of the induced alignment. In this short paper we make describe an approach to automated annotation-based bio-ontology alignment combined with subsequent measurement of the quality of those alignments. We do so using an example from lung disease genomics.

We begin with a list of 70 genes implicated in lung diseases. These are annotated to the Biological Process (BP) and Molecular Function (MF) branches of the Gene Ontology (GO[4]). The Cross Ontology Analytics tool (XOA, http://xoa.pnl.gov,[10, 12]) is then used to generate proximities between pairs of nodes in the BP and MF branches. The XOA scoring allows generation of putative alignments between BP and MF nodes, and then Joslyn *et al.*'s order-theoretical approach[6] is used to measure the structural quality of the generated alignments.

Lung Disease Genomics

The impact of genomics to study classes of diseases has yet to be fully realized. Research about lung diseases, focused on the cancers and other pathologies of specific tissue types, will benefit from systems analysis of cellular pathways and processes implicated in the presentation of disease states[9]. Genomic and proteomic analysis via ontological representations of gene product location and function has enabled the construction of predictive functional networks awaiting experimental validation[5].

We identified a set of 70 genes through our work in lung development and disease to evaluate the contribution of ontological alignments to further refined experimental hypotheses. We identified these 70 genes through expression analysis of mouse lung samples representing different developmental stages. The gene list is available at `ftp://ftp.informatics.jax.org/pub/curatorwork/ICBO09/lung_dev_genes.txt`. This defined set was chosen to be representative of molecular systems implicated in lung development and function.

The 2/26/09 version of mouse annotations (`ftp://ftp.informatics.jax.org/pub/reports/gene_association.mgi`) yields 1937 lines of GO annotations, including 424 distinct BP annotations, of which 388 were experimental annotations from mouse systems. There were 80 distinct MF annotations, 40 with experimental support. Overall, there are 62 genes with experimental BP annotations and 50 genes with experimental MF annotations. 48 genes, included in the results of the previous sentence, had both MF and BP experimental annotations.

Alignment Generation

XOA automatically generates links between BP and MF nodes based on their common annotations. Information theoretical approaches[8] are effective within one hierarchy. But because they require that similarity between two GO codes be computed in terms of the informational content of the most immediately dominating parent GO code, they cannot link GO codes across distinct gene subontologies. The vector space model approach obviates this limitation by computing the similarity between two GO codes as the cosine of vectors that encode the gene annotation associated with the two GO codes[1]. XOA combines these two approaches by turning relational links across GO codes into hierarchical links[12].

We model semantic hierarchies as finite, bounded, partially ordered sets (posets) $\mathcal{P} = \langle P, \leq \rangle$ [2], with nodes $a \in P$ as ontology concepts related by `is-a` links through $\leq$. The XOA similarity between the GO node $a \in P$ and the GO node $a' \in P'$ is then

$$XOA(a, a') := \max \left(\max_{b \in P} \left(\mathrm{sim}(a, b) \cos(a', b) \right), \max_{b' \in P'} \left(\mathrm{sim}(a', b') \cos(a, b') \right) \right),$$

where $\cos(a, a')$ denotes the cosine measure[11] between

GO nodes $a \in P, a' \in P'$ in the GO node $\times$ gene annotation matrix, and $\text{sim}(a,b)$ denotes Resnik's information theoretical similarity measure [8] between GO nodes $a, b \in \mathcal{P}$. An XOA analysis of the GO nodes annotated to our 70 test genes reveals 1970 BP-MF pairs $\vec{l} := \langle a, a' \rangle$ which are significant, with $p \leq 5\%$. Each such pair of **anchors** is a potential link between BP and MF.

An ontology **alignment** is a mapping $f: \mathcal{P} \to \mathcal{P}'$ taking anchors $a \in P$ in a semantic hierarchy $\mathcal{P} = \langle P, \leq \rangle$ to those $a' \in P'$ in another $\mathcal{P}' = \langle P', \leq' \rangle$. But a BP node $a \in P$ which has a high XOA score with an MF node $a' \in P'$ is also likely to have a high XOA score with other MF nodes $b' \in P'$. The complete set of 1970 links $\vec{l}$ yields a many-to-many alignment relation $F \subseteq P \times P'$. We need an alignment function $f: \mathcal{P} \to \mathcal{P}'$ with all left anchors appearing only once, so we sort the links by XOA to select the highest-scoring links $\langle a, a' \rangle$ where a or a' appears. These 36 one-to-one links are shown in Table 1.

Alignment Evaluation

We measure the structural properties of f shown in Table 1 (see [6] for more information). But our primary criterion is that f should not distort the metric relations of concepts, taking nodes that are close together and making them farther apart, or *vice versa*.

For two ontology nodes $a, b \in \mathcal{P}$, their **lower distance** is $d_l(a,b) := |\downarrow a| + |\downarrow b| - 2 \max_{c \in a \wedge b} |\downarrow c|$, where $\downarrow x = \{y \leq x\}$ is the set of all descendants of x, and $a \wedge b$ is the set of greatest lower bounds (glbs) below a and b. If a and b lack a glb, we assume a bottom node $0 \in P$ which is below all the leaves. The dual **upper distance** $d_u(a,b) = |\uparrow a| + |\uparrow b| - 2 \max_{c \in a \vee b} |\uparrow c|$ is also available, where $\uparrow x = \{y \geq x\}$ is the set of all ancestors of x, and $a \vee b$ is the set of least upper bounds (least common subsumers). Upper distance may appear more natural, but is not generally preferable for technical reasons related to the desire for e.g. siblings deep in the hierarchy to be closer together than siblings high in the hierarchy. While in general it may be preferable to use both in combination, in this paper we use lower distance only.

We can measure the change in distance between $a, b \in P$ induced by f as the **distance discrepancy**

$$\delta(a,b) := |\bar{d}_l(a,b) - \bar{d}_l(f(a), f(b))|,$$

where $\bar{d}_l(a,b) := \frac{d_l(a,b)}{\text{diam}_{d(\mathcal{P})}} \in [0,1]$ is the normalized lower distance between a and b in $\mathcal{P}$ given the diameter $\text{diam}_d(\mathcal{P}) := \max_{a,b \in P} d(a,b)$. In this case, we have diam(BP) = 14659, diam(MF)= 8260. Finally, we can measure the entire amount of distance discrepancy at a node $a \in P$ compared to all the other anchors $b \in P$ by summing

$$\delta_f(a) := \sum_{b \in P} \delta(a,b) = \sum_{b \in P} |\bar{d}_l(a,b) - \bar{d}_l(f(a), f(b))|.$$

Note that we use δ_f to indicate that this is an overall discrepancy of a with respect to the entire alignment f. Also note that since f is one-to-one, it is invertible, so $\forall a \in P, \delta_f(a) = \delta_f(f(a))$ and $\forall a' \in P', \delta_f(a') = \delta_f(f^{-1}(a'))$. Thus we can denote $\delta_f(\vec{l}) = \delta_f(a)$ for $\vec{l} = \langle a, f(a) \rangle$, which is also shown in Table 1.

Discussion and Further Work

Fig. 1 shows an abstract representation of a portion of the GO involving the top four scoring XOA links and the top two δ_f links. In general, we are pleased with the quality of the links provided by the XOA scores coupled with the one-to-one link filtering. It is a good sign that the nodes that did come up as significant are ones that make sense in the light of the gene list context (development). With one exception, the top 6 to 8 linked nodes represent molecules and processes associated with cell motility and with known regulators of cellular differentiation, such as the hedgehog signaling pathway. The frequency of nodes associated with motility underscore the importance of cellular migration during differentiation.

The distribution of XOA vs. δ_f is shown in Fig. 2. It can be seen that the XOA scoring method produces a strong alignment, with links having generally low δ_f scores. There are two exceptions which deserve further study to improve the analysis:

BP:GO:0007154 cell communication
MF:GO:0000062 acyl-CoA binding

BP:GO:0000187 activation of MAPK activity
MF:GO:0004672 protein kinase activity

To interpret this, for a given one-to-one link $\vec{l} = \langle a, f(a) \rangle$ between a BP node a and MF node $f(a)$, the XOA score measures the co-annotation of a and $f(a)$, while the δ_f score meaures the distance of $\vec{l}$ from all the other links in virtue of f, that is, the distance of a from all other BP anchors b, and dually the distance of $f(a)$ from all other MF anchors $f(b)$.

The lower distance $d_l(a,b)$ involves the numbers of nodes below a, b, and both of them. Thus from Fig. 1 we can see that both "BP:GO:0007154 cell communication" and "MF:GO:0004672 protein kinase activity" have unusually many nodes below them (341 and 105 respectively). This makes them effectively "far away" from the other nodes in BP and MF, while their corresponding anchor in the other

XOA	δ_f	BP Node	MF Node
10.14	0.070	GO:0006637 acyl-CoA metabolic process	GO:0016290 palmitoyl-CoA hydrolase activity
9.85	0.071	GO:0032927 positive regulation of activin receptor signaling pathway	GO:0050431 transforming growth factor beta binding
9.57	0.072	GO:0050677 positive regulation of urothelial cell proliferation	GO:0042056 chemoattractant activity
9.13	0.072	GO:0007228 positive regulation of hh target transcription factor activity	GO:0005113 patched binding
8.66	0.071	GO:0045723 positive regulation of fatty acid biosynthetic process	GO:0008009 chemokine activity
8.53	0.082	GO:0035023 regulation of Rho protein signal transduction	GO:0005099 Ras GTPase activator activity
8.00	0.079	GO:0048010 vascular endothelial growth factor receptor signaling pathway	GO:0005172 vascular endothelial growth factor receptor binding
7.51	0.087	GO:0050674 urothelial cell proliferation	GO:0005104 fibroblast growth factor receptor binding
7.44	0.076	GO:0016049 cell growth	GO:0005160 transforming growth factor beta receptor binding
7.41	0.233	GO:0048678 response to axon injury	GO:0019899 enzyme binding
7.39	0.103	GO:0007178 transmembrane receptor protein serine/threonine kinase signaling pathway	GO:0004702 receptor signaling protein serine/threonine kinase activity
7.33	0.115	GO:0033144 negative regulation of steroid hormone receptor signaling pathway	GO:0003690 double-stranded DNA binding
6.72	0.177	GO:0009967 positive regulation of signal transduction	GO:0048185 activin binding
6.52	0.080	GO:0007169 transmembrane receptor protein tyrosine kinase signaling pathway	GO:0004714 transmembrane receptor protein tyrosine kinase activity
6.42	0.080	GO:0014044 Schwann cell development	GO:0004675 transmembrane receptor protein serine/threonine kinase activity
6.40	0.103	GO:0045941 positive regulation of transcription	GO:0003713 transcription coactivator activity
6.33	0.075	GO:0048012 hepatocyte growth factor receptor signaling pathway	GO:0005017 platelet-derived growth factor receptor activity
6.31	0.089	GO:0045893 positive regulation of transcription DNA-dependent	GO:0016563 transcription activator activity
6.27	0.071	GO:0042993 positive regulation of transcription factor import into nucleus	GO:0015460 transport accessory protein activity
6.20	0.183	GO:0001558 regulation of cell growth	GO:0019838 growth factor binding
6.08	0.072	GO:0007171 activation of transmembrane receptor protein tyrosine kinase activity	GO:0005161 platelet-derived growth factor receptor binding
6.05	0.071	GO:0030949 positive regulation of vascular endothelial growth factor receptor signaling pathway	GO:0005111 type 2 fibroblast growth factor receptor binding
5.75	0.070	GO:0006919 caspase activation	GO:0019834 phospholipase A2 inhibitor activity
5.57	0.078	GO:0048706 embryonic skeletal development	GO:0005024 transforming growth factor beta receptor activity
5.50	0.415	GO:0000187 activation of MAPK activity	GO:0004672 protein kinase activity
5.46	0.101	GO:0006816 calcium ion transport	GO:0005262 calcium channel activity
5.36	0.776	GO:0007154 cell communication	GO:0000062 acyl-CoA binding
5.30	0.144	GO:0006468 protein amino acid phosphorylation	GO:0004674 protein serine/threonine kinase activity
5.21	0.072	GO:0051795 positive regulation of catagen	GO:0001540 beta-amyloid binding
5.19	0.093	GO:0016481 negative regulation of transcription	GO:0016564 transcription repressor activity
5.17	0.070	GO:0051450 myoblast proliferation	GO:0005021 vascular endothelial growth factor receptor activity
5.04	0.072	GO:0050890 cognition	GO:0019855 calcium channel inhibitor activity
5.01	0.073	GO:0000122 negative regulation of transcription from RNA polymerase II promoter	GO:0003702 RNA polymerase II transcription factor activity
4.85	0.072	GO:0007184 SMAD protein nuclear translocation	GO:0046332 SMAD binding
4.84	0.071	GO:0001707 mesoderm formation	GO:0045545 syndecan binding

Table 1: One-to-one alignment links $\vec{l} = \langle a, f(a) \rangle$ for $p \geq 5\%$, sorted down by XOA score, and showing $\delta_f(\vec{l})$. Underlined links are illustrated in Fig. 1.

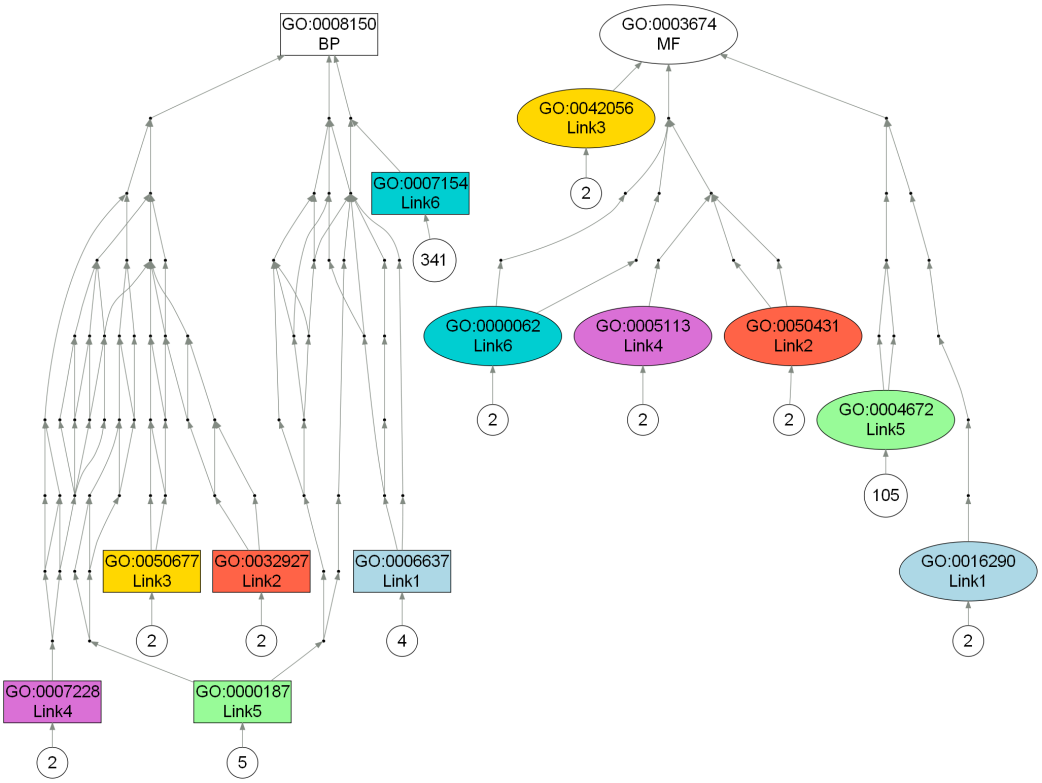


Figure 1: The portion of the BP (left) and MF (right) GO branches involving the top four XOA and top two δ_f links. Only the anchors are shown with their GO IDs (see Table 1 for descriptions). Matching nodes are indicated by color and link numbers. Ancestors are shown, up to the BP or MF root, but all interior nodes are collapsed. Below each anchor is the number of descendant nodes. There are no common nodes below any pair of anchors.

ontology is close to its comrades. This is clear in Fig. 1, and thus our method identifies these links which are clearly significant by XOA, but also distant from the other links.

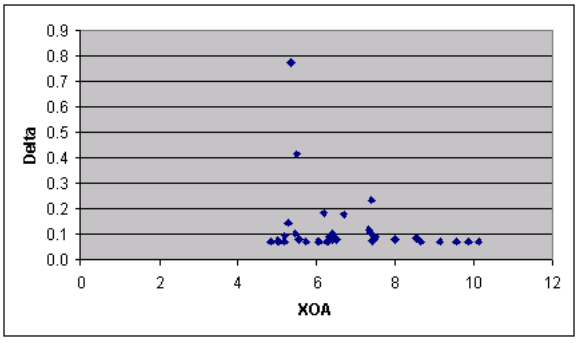


Figure 2: $XOA(a, f(a))$ vs. $\delta_f(a)$.

δ_f provides a measure only about ontology *structure*, and there may be reasons in ontology design or annotation for high δ_f to be preferable, e.g. if it were important that annotations be made high in the structure in some cases. The results would also be different if there were common nodes below pairs of anchors, which is entirely possible in the GO DAG structure with multiple inheritance, especially if the anchors were higher. Finally, note that the number of descendants is correlated both with level in the GO, and the information content (probability) of a node used in semantic similarity calculation. These correlations need to be explored in future work. Further work for a full paper includes:

- There are potential difficulties of mixing experimental and inferential annotations, as reported here, these should be analyzed separately.
- The analytical pipeline needs to be tested for sensitivity at multiple points, especially the filtering to one-to-one links: it is likely that there are links which re-use an anchor which have only a slightly different XOA score, but would produce a preferable mapping according to δ_f. Additionally, the alignment measurement method[6] originally was designed to work on many-to-many alignment relations $F \subseteq P \times P'$, so extensions in this direction may be desirable.
- We have begun analysis on the distribution of δ_f as a function of p-value cutoff.
- Other aspects of the alignment measurement methodology[6] need to be incorporated, including: reconciling the use of upper distance together with lower distance; and the additional use of an **order discrepancy measure**, which rather than being sensitive to the *distances* between links, measures *order violations* (e.g. mapping siblings to parent-child links) implied by an alignment.

Acknowledgements

This work supported under NIH/NINDS grant R01NS057484-03 and NIH/NHGRI grant HG002273.

References

1. Bodenreider OM, Aubry M, Burgun A. "Non-lexical approaches to identifying associative relations in the gene ontology", *Pacific Symp. Biocomputing*. 2005; 104-115
2. Davey BA, Priestly HA. *Introduction to Lattices and Order*, Cambridge: Cambridge UP; 1990
3. Jérôme E, Shvaiko, P. *Ontology Matching*, Hiedelberg: Springer-Verlag; 2007
4. Gene Ontology Consortium. "The Gene Ontology: Tool For the Unification of Biology". *Nature Genetics*. 2000; 25(1):25-29
5. Guan Y *et al.* "A Genomewide functional network for the laboratory mouse". *PLoS Comutational Biology*. 4 e1000165 [PMID: 18818725]. 2008
6. Joslyn CA, Donaldson A, Paulson P: (2008) "Evaluating the Structural Quality of Semantic Hierarchy Alignments", *Int. Semantic Web Conf. (ISWC 08)*. Available from: http://dblp.uni-trier.de/db/conf/semweb/iswc2008p.html#JoslynDP08
7. Kirsten T, Andreas T, Rahm E: (2007) "Instance-Based Matching of Large Life Science Ontologies", in S Cohen-Boulakia, V Tannen ed. *DILS 2007, Lecture Notes in Bioinformatics*, 4544. p. 172-187. Heidelberg: Springer-Verlag.
8. Lord PW, Stevens R, Brass, A, Goble CA. "Investigating Semantic Similarity Measures Across the Gene Ontology: the Relationship Between Sequence and Annotation". *Bioinformatics*. 2003; 10:1275-1283
9. Raj JU *et al.* "Genomics and proteomics of lung disease: conference summary". *Am. J. Physiol Lung Cell Mol Physiol*. 2007; 293:L45-L51, PMID: 17468134
10. Riensche RM, Baddeley BL, Sanfilippo A, Posse C, Goplan B. "XOA: Web-Enabled Cross-Ontological Analytics". *IEEE Congress on Services*. 2007.
11. Salton GA, Wong A, Yang CS "A Vector space model for automatic indexing". CACM. 1975; 18(11):613-620.
12. Sanfilippo A, Posse C, Gopalan B, Riensche R, Beagley N, Baddeley B, Tratz S, Gregory M "Combining Hierarchical and Associative Gene Ontology Relations With Textual Evidence in Estimating Gene and Gene Product Similarity". *IEEE Trans. Nanobio.*. 2007; 6(1):51-59

Metarel: An Ontology to Support the Inferencing of Semantic Web Relations within Biomedical Ontologies

Ward Blondé[1,2], Erick Antezana[1], Bernard De Baets[2], Vladimir Mironov[3], Martin Kuiper[3]
[1]Department of Plant Systems Biology, VIB, Ghent, Belgium
[2]Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Belgium
[3]Department of Biology, Norwegian University of Science and Technology, Trondheim, Norway

Abstract

While OWL, the Web Ontology Language, is often regarded as the preferred language for Knowledge Representation in the world of the Semantic Web, the potential of direct representation in RDF, the Resource Description Framework, is underestimated. Here we show how ontologies adequately represented in RDF could be semantically enriched with SPARUL. To deal with the semantics of relations we created Metarel, a meta-ontology for relations. The utility of the approach is demonstrated by an application on Gene Ontology Annotation (GOA) RDF graphs in the RDF Knowledge Base BioGateway. We show that Metarel can facilitate inferencing in BioGateway, which allows for queries that are otherwise not possible. Metarel is available on http://www.metarel.org.

Introduction

Ontologies have become one of the cornerstones of Knowledge Management (KM) in the Life Sciences.[1] They are increasingly used for annotating and integrating biomedical data, including genomic data, patient data, disease data, molecular data and more. For ontologies to fulfill their intended role, it is mandatory that both the ontologies and the data are modeled with the use of technologies that enable efficient integration and querying. In addition, these technologies should allow inferencing of new knowledge, one of the great promises of Knowledge Representation (KR).

The Semantic Web provides such technologies, the most important ones being the Resource Description Framework RDF and the Web Ontology Language OWL.[2,3] The life sciences, in particular the domain of systems biology, are expected to be among the early adopters of these technologies.[4] While a number of successful applications of the Semantic Web technologies in the life sciences have been reported (GenoQuery, LinkHub, Thea-online, BioDash), the field is still in its infancy and a number of technical hurdles need to be overcome.[5-9] For example, while OWL allows semantically rich knowledge representation, querying large knowledge bases represented in OWL poses a tremendous computational tractability challenge.[10] Here we explore how a highly optimized RDF implementation can be used to alleviate some of the hurdles while still supporting rich inferencing.

Rendering Class Level Relations in RDF

The two languages used widely for bio-ontologies, OBOF and OWL, differ markedly in the way they express the semantics of relations between classes.[11,12] OWL expresses such relations by defining properties (relations between instances) in a property hierarchy. Relations between classes are created by adding extra fillers on the properties, which allows for number restrictions and grouping properties as necessary and sufficient conditions for defining classes. These fillers make the links between classes indirect in OWL/RDF, the RDF representation of OWL. OBOF, on the other hand, assumes all classes as defined by definition tags and relations are never considered as sufficient conditions. This approach has allowed to make direct links for relations between classes (see Figure 1).

The modeling with direct links in RDF, illustrated in Figure 1A, has a number of advantages over OWL/RDF: it is less verbose, it requires less computational power for loading and querying, and it is more intuitive. Moreover, the number of instances documented in biological ontologies is very small compared to the number of classes, which makes the treatment of relations between classes especially important in this domain. The direct links for relations between classes in OBOF can be readily modelled in RDF by putting them in the central place of an RDF triple (the predicate).

Interestingly, these relations (like is_a, part_of, etc.) have their own URIs (Unique Resource Identifiers) and they can appear in the first or the last place of a triple (resp. subject or object) as well. They can connect with any other URIs by using metarelations

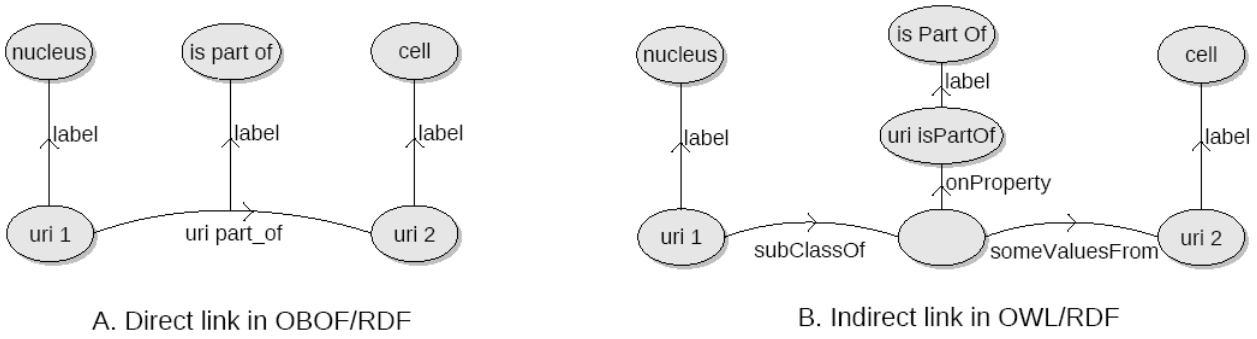


Figure 1. OBOF/RDF (1A) relies on direct links between classes. OWL/RDF (1B) uses anonymous classes, as blank nodes, to link classes.

as predicates. To exploit the opportunities this opens, we created Metarel, a meta-ontology for relations that can support RO, the Relationship Ontology of the OBO Foundry.[13,14] With Metarel, OBO ontologies can be translated to RDF in a format with direct links, without any loss of expressivity.

Metarel was manually created with OBO-Edit and it has an export to RDF, created with ONTO-PERL.[15] We list some of its most important features: 1. It allows to create meta-relations between relations; 2. It distinguishes 'all-some' relations from other types of relations; 3. It can indicate unambiguously which pairs of relations are each others inverses; 4. It has a place in its hierarchy where formally defined instance level relations can be attached; 5. It classifies reflexive and transitive relations in meta-classes; 6. It contains constructs for composites of relations; 7. It contains a meta-class for relations that are relevant towards inferencing (e.g. exclude *anatomically_related_to*, but not *dorsal_to*).

Metarel is used in the RDF knowledge base BioGateway.[16,17] which has OBO relations. By linking all the relations in BioGateway to Metarel we obtain BioMetarel, essentially a bioscience specification of Metarel.

Inferencing with OBO Foundry Relations

To investigate the efficacy of BioMetarel to facilitate inferencing in BioGateway we started from the semantics of its OBO relations. First of all, we emphasize that only relations at the class level are defined in RO. Their definitions refer to relations at the instance level, but those have neither unique identifiers nor definitions on their own. Therefore we will only infer new relations at the class level. As discussed in [14], the relations in OBOF have an 'all-some' semantics. This means that if *e.g. A part_of B*, then for all the instances *a* of class *A* there is some instance *b* of class *B* for which *a* is part of *b*. The validity of any inferences from this semantics depends on the extent to which annotators have applied this rule correctly in producing knowledge statements for OBO. We found five sound mechanisms to infer new relations with OBO semantics:

1. A reflexive closure creates a relation link *A R A* for every class *A* and for every reflexive relation *R*. A query that asks for all the parts of the Golgi apparatus, will also return 'Golgi apparatus', because *part_of* is reflexive.
2. A transitive closure creates a relation link *A R C*, from any class *A* to any class *C*, for every transitive relation *R*, if the relation links *A R B* and *B R C* exist already. E.g. 'nucleolus *part_of* nuclear lumen' and 'nuclear lumen *part_of* nucleus', creates 'nucleolus *part_of* nucleus'.
3. The inferencing of relations that have priority over the subsumption relation 'is_a' creates a relation link *A R C* if the links *A R B* and *B is_a C* exist, as well as for *A is_a B* and *B R C*, whenever *R* has an all-some semantics. E.g. 'BRAF1_HUMAN *has_function* diacylglycerol binding' and 'diacylglycerol binding *is_a* lipid binding', creates 'BRAF1_HUMAN *has_function* lipid binding'.
4. The inferencing from the relation hierarchy creates a relation link *A R B* if the link *A S B* exists, and if *S* is a subrelation of *R*. E.g. 'AKIP _HUMAN *negatively_regulates* mitosis' creates 'AKIP_HUMAN *regulates* mitosis'.
5. A compositional closure creates a relation link *A R C* if the links *A S B* and *B T C* exist and if

```
BASE   <http://www.semantic-systems-biology.org/>
PREFIX ssb:<http://www.semantic-systems-biology.org/SSB#>
INSERT INTO GRAPH <25.H_sapiens_tc> {
  ?class1 ?resulting_relation ?class3.
}
WHERE {
  GRAPH <25.H_sapiens_tc> {
    ?class1 ?first_relation ?class2.
  }
  GRAPH <gene_ontology_edit_tc> {
    ?class2 ?second_relation ?class3.
  }
  GRAPH <biometarel> {
    ?composite ssb:first_relation ?first_relation.
    ?composite ssb:second_relation ?second_relation.
    ?composite ssb:resulting_relation ?resulting_relation.
  }
}
```

Figure 2. A SPARUL update query that computes the compositional closure of GOA annotations for Homo Sapiens. In this update, a composition like '*located_in* plus *part_of* results in *located_in*' is extracted from BioMetarel and operated over the Gene Ontology and the GOA data

the relations *S*, *T* and *R*, not necessarily all different, form a composite (role chain). E.g. 'NARF_HUMAN *located_in* nuclear lumen' and 'nuclear lumen *part_of* nucleus' creates 'NARF_HUMAN *located_in* nucleus'.

Somewhat counterintuitively, relations with an all-some semantics can not have inverses. Consequently, they can not be symmetric either, as this would imply they are their own inverses. Consider for example the statement 'feather part_of animal'. As every feather is part of a bird, and every bird is an animal, this can be considered a sound statement. The inverse statement 'animal has_part feather', meaning that every animal has some feather as part, is clearly nonsense. To indicate e.g. that every feather is part of some bird, and also every bird has some feather as part, annotators should use two statements.

Creating the Closures

As an application, we created a relational closure over the relations in the Gene Ontology Annotations (GOA) and the Gene Ontology (GO) in BioGateway.[18,19] By this we mean the explicit creation of all the relations that are relevant in queries from users, and that can be inferred from the documented relations in BioGateway and from the semantics of relations. We used the RDF update language SPARUL for computing and adding these relations as RDF triples.[20] The method consists of four steps:

1. *Creating Biorel.obo*. This file expands RO.obo with all the relations that are used in the OBO Foundry, and with some extra tags for transitivity that were missing in RO.obo.
2. *Creating BioMetarel*. We merged the exports Biorel.rdf and Metarel.rdf to the BioMetarel RDF graph, with SPARUL. The enhancement of the semantic content for inferencing occurred in this step.
3. *Creating the closure of the Gene Ontology.* This was done by recursively running the SPARUL queries, (effectively applying the inferencing mechanisms in the previous section) on the BioMetarel and the GO graphs, until no further inferences could be made. The closure graph of GO contains 1.2 million triples, whereas the original GO graph contained only 0.57 million triples.
4. *Creating the closure of the Gene Ontology Annotations.* We created the compositional closure and the priority over *is_a*, for all the GOA graphs in BioGateway. The closure graph for *Homo sapiens*, for example, has 4.0 million triples, compared to 3.3 million triples for the normal graph.

The preconstructed closures allow many useful queries with SPARQL, the RDF query language, that are otherwise not possible.[21] Examples include finding the proteins that are located in the same protein complex and finding all the proteins with a given function or involved in a given process. All the original sources are kept separately, to allow querying on the original annotations. SPARQL queries that try to obtain similar results without closures are very complicated and computationally unperformant.

Conclusion

We have shown how the relation ontology Metarel can be used to perform extensive inferencing in biomedical ontologies represented in RDF, a Semantic Web standard. For this, we integrated the OBO Foundry relations in the hierarchy of Metarel, and the ensuing biological relationship ontology BioMetarel was used to recursively inference in the RDF store BioGateway. Triples constructed by inferencing were propagated by operating SPARUL update queries over BioMetarel and the relevant biomedical ontologies. Such inferences allow more powerful queries, and essentially increase the value of RDF for Knowledge Management significantly.

Acknowledgements

We thank Peter Dawyndt for discussions about Metarel, Nirmala Seethappan for maintenance of BioGateway, and Kenny Billiau, Bjørn Lindi, Dany Cuyt and Luc Van Wiemeersch for IT support.

References

1. Shadboldt N, Hall W and Berners-Lee T. The semantic web revisited, IEEE Intelligent Systems, 2006, 21:96–101.
2. RDF resource description framework, 2004. http://www.w3.org/RDF/.
3. OWL web ontology language guide, 2004. http://www.w3.org/TR/owl-guide/.
4. Ruttenberg A, Clark T, Bug W, *et al*. Advancing translational research with the semantic web. BMC Bioinformatics, 2007, 8 Suppl 3, p. S2.
5. Antezana E, Kuiper M and Mironov V. Biological knowledge management: The emerging role of the Semantic Web technologies. Brief Bioinform, 2009, p.24
6. Lemoine F, Labedan B and Froidevaux C. GenoQuery: A new querying module for functional annotation in a genomic warehouse. Bioinformatics, 2008, 24:322–329.
7. Smith A, Cheung K, Yip K, *et al*. LinkHub: A semantic web system that facilitates cross-database queries and information retrieval in proteomics. BMC Bioinformatics, 2007, 8 Suppl 3, p. S5.
8. Pasquier C. Biological data integration using semantic web technologies. Biochimie, 2008, 90:584–594.
9. Neumann E and Quan D. Biodash: A semantic web dashboard for drug development. In Pacific Symposium on Biocomputing. World Scientific, 2006, pp 176–187.
10. Weithöner1 T, Liebig T, Luther M and Böhm S. What's wrong with OWL benchmarks? In Proc. of the 2nd Int. Workshop on Scalable Semantic Web Knowledge Base Systems, 2006, pp. 101–114.
11. Smith B, Ashburner M, Rosse C, *et al*. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. Nat Biotech, 2007, 25(11):1251–1255.
12. Golbreich C and Horrocks I. The OBO to OWL mapping, GO to OWL 1.1! In OWLED, 2007, p35.
13. Metarel. http://www.metarel.org.
14. Smith B, Ceusters W, Klagges B, *et al*. Relations in biomedical ontologies. Genome Biology, 2005, 6:R46.
15. Antezana E, Egaña M, De Baets B, Kuiper M and Mironov V. ONTO-PERL: An api supporting the development and analysis of bio-ontologies. Bioinformatics, 2008, pp. 885–887.
16. BioGateway. http://www.semantic-systems-biology.org/biogateway.
17. Antezana E, Blondé W, Egaña M, Rutherford A, Stevens R, De Baets B, Mironov V and Kuiper M. Structuring the life science resourceome for semantic systems biology: Lessons from the BioGateway project. SWAT4LS, 2008, CEUR-WS 435.
18. Camon E, Magrane M, Barrell D, *et al*. The Gene Ontology Annotation (GOA) database: Sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Research, 2004, 32:262–266.
19. The Gene Ontology Consortium. Gene Ontology: Tool for the unification of biology. Nature Genet., 2000, 25:25–29.
20. SPARUL. http://jena.hpl.hp.com/~afs/SPARQL-Update.html.
21. SPARQL. http://www.w3.org/TR/rdf-sparql-query/

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

Using Multiple Reference Ontologies: Managing Composite Annotations

John H. Gennari[1], Maxwell L. Neal[1], Jose L.V. Mejino Jr.[2], Daniel L. Cook[1,2,3]

Departments of [1]Medical Education & Biomedical Informatics, [2]Biological Structure, and [3]Physiology & Biophysics, University of Washington, Seattle, WA, USA

Abstract

There are a growing number of reference ontologies available across a variety of biomedical domains and current research focuses on their construction, organization and use. An important use case for these ontologies is annotation—where users create metadata that access concepts and terms in reference ontologies. We draw on our experience in physiological modeling to present a compelling use case that demonstrates the potential complexity of such annotations. In the domain of physiological biosimulation, we argue that most annotations require the use of multiple reference ontologies. We suggest that these "composite" annotations should be retained as a repository of knowledge about post-coordination that promotes sharing and interoperation across biosimulation models.

Connecting Multiple Reference Ontologies

We define a reference ontology as a carefully-constructed ontology that aims to completely cover a specific realm or domain of knowledge[1,2]. By definition, such an ontology must be both broad and deep in its domain, and designed for reusability across multiple sorts of users and use cases. In biology, one goal of the OBO resource (http://obofoundry.org/) is to encourage the development of non-overlapping reference ontologies so that users can unambiguously access terms from such ontologies. In biology, an exemplar reference ontology is the Foundational Model of Anatomy (FMA)[2].

Ontologies are most effective when they are designed with specific use cases in mind. For many, the motivating use case has been *annotation*: users need to add unambiguous semantic metadata about their raw data, whether that data is from genomic research, clinical findings, or images. To date, the conception of these annotations has been relatively simple. For example, a gene expression level from some experimental result will be annotated in-line with a Gene Ontology (GO) id, or possibly a direct URI to the relevant GO term.

Annotations (even simple ones) provide a compelling justification for ontology development. Annotations allow external users, or even computer systems to explore and automatically align data and results across multiple sources. This use of annotations requires users to carry out two sorts of tasks: (1) annotating source data against ontologies, and (2) searching and integrating data from sources that use those ontologies for annotation. As others have pointed out, these tasks fit well into the intelligent information retrieval capabilities of the semantic web[3].

In this paper, we argue that this relatively simple use of annotation and ontologies can become very complex if annotations include multiple ontologies. Our domain of interest is in biosimulation, where researchers build models for understanding pathology or physiology. We show that when researchers annotate such models, they need to use multiple orthogonal ontologies. We present our preliminary architecture for these *composite annotations*, and describe prototype tools and ideas for the two user tasks described above: Annotating biosimulation models and then searching and integrating those models.

As we show, these annotations provide a solution to one case of the post- vs. pre-coordination problem: there are too many properties of too many biological entities to attempt to pre-coordinate all combinations. Instead, via composite annotations, users can post-coordinate concepts as needed, and store those combinations of terms across ontologies that are useful and relevant for their tasks. Without retaining this knowledge, ontology developers and end users are faced with a combinatorial problem—a cross product of terms across many large orthogonal ontologies.

The Biophysical Semantics of Biosimulation

For several years, our research group has been developing systems and ontologies for use with physiological biosimulation models. Recently, researchers have aimed at building a complete Physiome[4], a flexible integration of component models into large-scale or special-purpose biosimulations for application to clinical and investigatory problems. Toward this goal, a number of libraries of biosimulation models have been made available, notably BioModels (an SBML collection, http://www.ebi.ac.uk/biomodels/), the CellML repository (http://www.cellml.org), and the JSim library (http://physiome.org/jsim/).

The fundamental challenge for integrating and understanding biosimulation models is that although these models are based on classical physics and formally expressed in mathematics, the semantics of these models—the meaning of variables and equations—is usually only implicit in model computational code (e.g. naming conventions) or annotated using *ad hoc* in-line code comments. Although current best practices in biosimulation modeling include adherence to some annotation standards[5], these have not yet been widely adopted. We certainly applaud the use of OBO standards such as the FMA, ChEBI, GO, and the OBO Cell Type ontology. However, if annotations for biosimulation models are in-line, maintaining and searching over these annotations can be a challenge.

In addition, all of the above ontologies are for biological structure and physical entities. For physiological modeling, it is important to also represent the principles by which such entities participate in processes. Recently, we have developed the Ontology of Physics for Biology (OPB)[6], an ontology of the physical properties and physical laws by which biological processes occur. As such, it is orthogonal to strictly structural representations (e.g., FMA, ChEBI) in that it represents the physical properties that reside in structural entities. Thus, in biosimulation models, the elements of interest necessarily include both reference to structural entities of biology (e.g., blood, muscle, or smaller entities such as glucose or oxygen) as well as ***properties*** of those entities (e.g., flow, mass, or chemical concentration). Below, we provide specific examples of these composite annotations.

Example Composite Annotations

As a simple example, consider a common concept used in many cardiovascular biosimulation models: Aortic blood pressure. This concept may be mapped to differently named variables (Aop, AP, PAorta, etc) in different models. To integrate models that share this concept, these variables would have to be annotated with both the anatomical entity (blood-in-aorta) as well as the physical property that is modeled: fluid pressure. This is a simple example, because it involves just two reference ontologies, the FMA and the OBP, and because fluid pressure is a property of the FMA entity blood-in-aorta.

As a slightly more complex example, consider the concentration of oxygen in the blood of the aorta. This entity (which might be used by many different biosimulation models) needs three ontologies: ChEBI, for oxygen, the OPB, for chemical concentration, and the FMA, for blood in the aorta. If we omit any of these three ontologies, our representation is inaccurate or even erroneous. If we are not explicit about chemical concentration then we might be discussing (for example) the flow of oxygen in the aorta. If we omit the aorta, we might be discussing concentration of oxygen in the vena cava. Finally, we need ChEBI for oxygen as there are many other chemicals of interest in the aortic blood (e.g., calcium ion concentration).

Finally, annotations become most complex in models that are multi-scale. Consider a model that includes glucose concentration in beta cells. It may matter a great deal whether that concentration is cytoplasmic, extracellular, arterial, or venous. Potentially, such a concept might need five reference ontologies: cell component (e.g., GO cell component), cell type (e.g. the OBO CellType), as well as the FMA, the OPB, and ChEBI.

Effectively, composite annotations are recording "cross-products of interest" over the participating reference ontologies. Thus, one could imagine a set of tuples for pathway level biosimulation that were {OPB x ChEBI x FMA} or perhaps {OPB x ChEBI x GOCellComponent}. (In this conference, Mungall et al. also discuss a similar cross-product idea for the GO.[7]) However, the vast majority of such tuples would be nonsensical or not of interest for a particular model or group of biosimulation researchers (e.g., momentum of oxygen in the skull bone). In addition, our composite annotations need internal structure—formal terms that describe the relationship between, for example, blood and the aorta ("contained-in"). The research questions we raise deal with how to create, store, and retrieve for reuse, these sort of composite annotations.

Managing Annotations: SemSim for Biosimulation

For a single biosimulation model, we have developed an approach to composite annotation we call a "SemSim model" (for Semantic Simulation)[8,9]. SemSim models are OWL-based ontologies that capture the computational and semantic aspects of a biosimulation model, and they include a set of annotations for that particular biosimulation model. At most, there is one annotation per variable and equation in the source code. For variables, these are *composite annotations*, where each annotation has the structure we diagram in Figure 1.

Biosimulation model variables, such as "PAorta", are annotated by first mapping them to physical properties, such as pressure, flow, concentration, etc. These properties are defined in the OPB, and referenced in the composite annotation. It is these properties that take on numeric values during any given simulation run. As Figure 1 shows, these properties are then connected to the physical entities (via "has property" links) which then point to entities in structural reference ontologies. If there is more

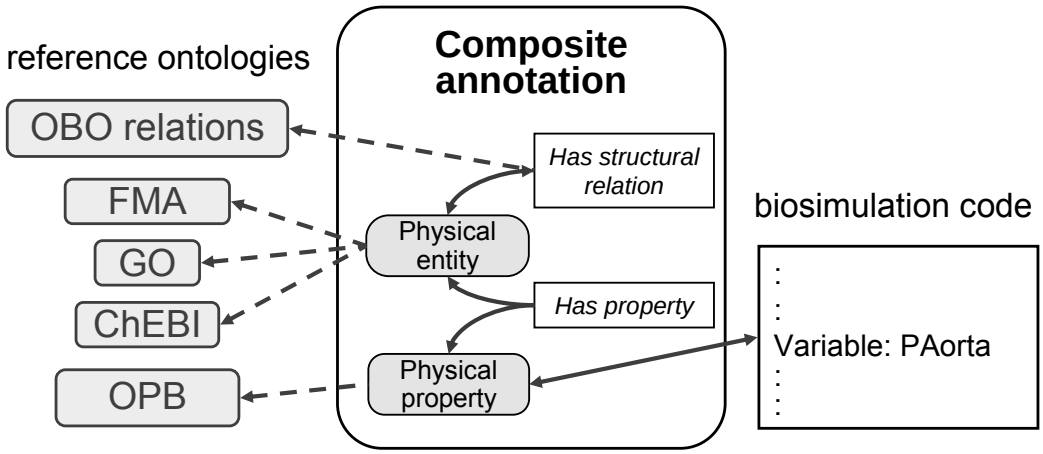


Figure 1. The structure of our composite annotations, which connect variables in simulation code to a set of reference ontologies. A SemSim model is a collection of these annotations, for a set of variables as used in a specific piece of biosimulation code.

than one entity and more than one structural ontology (e.g., oxygen in blood), then these are connected via "has structural relation" links. As we discuss later, where possible, we use the OBO relations ontology for such linking terms (part_of, contained_in, etc.).

In contrast to current annotation practices, our SemSim models are external entities to both the reference ontologies and the biosimulation code. One advantage of this design is that the source code can remain un-modified, an important feature when annotating legacy code. A second advantage is that we can easily collect the set of all annotations as a repository for search and reuse.

Creating Composite Annotations

For biosimulation researchers, the structure shown in Figure 1 should be largely invisible. Thus, we have developed prototype tools that hide this representational complexity and help users author and create composite annotations from biosimulation source code. For creating annotations, our prototype tool, SemGen, parses the source code to find instances of variables, and then prompts the user for search terms to use with particular ontologies. The system then queries these ontologies to find exact matches and IDs for the terms, and finally generates the composite annotation as part of a SemSim model.

As we develop a larger repository of annotations for biosimulation models, our SemGen system can better assist users. For, example, if a model uses a variable that captures "cytoplasmic glucose concentration in pancreatic beta cells", then this annotation could require five searches across five participating reference ontologies. However, if some other user has already created a similar or related annotation, then the SemGen system can return a list of these as soon as the user enters any one of these terms. E.g., as soon as "glucose" is entered, the system could return a list of all prior glucose annotations, and one of these may be a close or perfect match for the user.

Because there are relatively few biosimulation models available, the number of useful composite annotations for models is small, at least compared to the cross product of the cardinality of the reference ontologies. Thus, annotators help us carry out post-coordination of terminologies: the composite annotations are created only on an as-needed basis, and then stored in a repository for reuse.

Using Annotations to Search and Merge Models

As we alluded to earlier, there are two sorts of user tasks for annotations. In addition to creating composite annotations, users need to search annotations and their models, and then perhaps merge or adapt models created by others. Reusing and adapting others' models is common in biosimulation engineering, but currently, this work is manual, costly, error-prone, and typically requires extensive communication and collaboration between bioengineers.[9]

In prior publications, we have presented early results that show how our SemSim approach would make model merging semi-automatic.[8,9] Although promising, this preliminary work avoids some of the broader indexing and retrieval challenges for a repository of composite annotations. In particular, for semantic web use cases, composite annotations need (a) a unique name or URI, and (b) indices for appropriately efficient retrieval. We can assume that each reference ontology term (such as "FMA: blood in aorta") already has a URI. Thus, although unwieldy, one could use URIs for composite annotations that simply consist of concatenations of the URIs of each reference ontology term.

We believe that users may want to search the annotation repository in a variety of ways. Thus, it seems likely that these annotations will need to be indexed with all of their component terms. To continue with the glucose example, users may want to begin with glucose, or pancreatic beta cells, or "cytoplasmic glucose" and therefore, all of these should be indexed, so that the system can retrieve the full term regardless of how the user searches.

Managing Orthogonal Ontologies: OBO Relations

The management of multiple ontologies for annotating biosimulation models is just a specific example of managing multiple orthogonal ontologies. This issue is faced by the OBO set of ontologies, and partially addressed by the OBO Relation Ontology. This ontology provides the formal relations needed to describe ***how*** the structural entities in a composite annotation relate to each other. For example, for cytoplasmic glucose concentration in beta cells, we can say precisely that we are referring to the cytoplasm (GO CellComponent) that is "part of" (OBO relation) the pancreatic B cell (OBO CellType).

Thus, the OBO relations ontology provides the ability to appropriately link entities across OBO ontologies that pertain to structural entities. However, this ontology does not include relationships appropriate for connecting non-structural ontologies such as the OPB. How should the notion of "pressure" be related to the concept of "blood"? In our SemSim approach, we currently use the generic "has property" relation for such links.

Pragmatically, our initial work has focused on managing and building composite annotations. We certainly use the OBO relation ontology where appropriate, but as a first goal, building a corpus of useful composite annotations will be a significant contribution, and can ease the task of biosimulation model integration.

Discussion and Conclusions

In this paper, we describe composite annotations to represent entities of interest to biosimulation modelers. In addition, we propose that these annotations can be used as a way of storing knowledge about post-coordination, so that useful terms such as "concentration of oxygen in blood of aorta" can be easily retrieved or created on-the-fly. Elsewhere, we demonstrated the value of such annotations for merging biosimulation models, and here, we raise issues and propose possible solutions for building a semantic web repository of such composite annotations.

In support of the Physiome vision, the biosimulation research community is working to integrate models to build larger and more complex models (with the expectation that such models are more predictive and useful). We argue that reference ontologies and tool support could provide significant assistance with this work. However, a key first step to integrating models is a solid understanding of the semantics of model variables and equations. We propose that a repository of composite annotations could both make annotation of additional models easier, as well as allow researchers and systems to find variables that share common semantics across biosimulation models.

Acknowledgements

This work was partially funded by NIH grants #R01 HL087706-01 and #T15 LM007442-06. We also thank Michal Galdzicki for contributions to these research ideas.

References

1. Brinkley JF, *et al*. *A framework for using reference ontologies as a foundation for the semantic web.* AMIA Annu Symp Proc, 2006: p. 96–100.
2. Rosse C. and Mejino JLV. *A Reference Ontology for Bioinformatics: The Foundational Model of Anatomy.* Journal of Biomedical Informatics, 2003. 36: p. 478–500.
3. Sahoo SS, *et al*. *An ontology-driven semantic mashup of gene and biological pathway information: Application to the domain of nicotine dependence.* Journal of Biomedical Informatics, 2008. 41(5): p. 752–765.
4. Hunter PJ and Borg TK. *Integration from proteins to organs: the Physiome Project.* Nat Rev Mol Cell Biol, 2003. 4(3): p. 237–43.
5. Le Novere N, *et al*. *Minimum information requested in the annotation of biochemical models (MIRIAM).* Nat Biotechnol, 2005. 23(12): p. 1509–15.
6. Cook DL, *et al*. *Bridging Biological Ontologies and Biosimulation: The Ontology of Physics for Biology.* AMIA Annu Symp Proc, 2008: p. 136–140.
7. Mungall C, *et al*. *Cross product extensions of the gene ontology*. in *Proceedings of the International Conference on Biomedical Ontology*. 2009. Buffalo, NY.
8. Gennari JH, *et al*. *Integration of multi-scale biosimulation models via light-weight semantics.* Pac Symp Biocomput, 2008. 13: p. 414–425.
9. Neal ML, *et al*. *Advances in semantic representation for multiscale biosimulation: A case study in merging models.* Pac Symp Biocomput, 2009: p. 304–315.

MIREOT: The Minimum Information to Reference an External Ontology Term

Mélanie Courtot[1], Frank Gibson[2], Allyson L. Lister[3], James Malone[4], Daniel Schober[4, 5], Ryan R. Brinkman[1], Alan Ruttenberg[6]

[1]Terry Fox Laboratory, BC Cancer Agency, Vancouver, BC, Canada; [2]Abcam plc, Cambridge, UK; [3]CISBAN and School of Computing Science, Newcastle University, Newcastle upon Tyne, UK; [4]The European Bioinformatics Institute, Cambridge, UK; [5]Institute of Medical Biometry and Medical Informatics (IMBI), University Medical Center, Freiburg, Germany; [6]Science Commons, Cambridge, MA, USA

Abstract

While the Web Ontology Language (OWL) provides a mechanism to import ontologies, this mechanism is not always suitable. First, given the current state of editing tools and the issues they have working with large ontologies, direct OWL imports have sometimes proven impractical for day-to-day development. Second, ontologies chosen for integration may be under active development and not aligned with the chosen design principles. Importing heterogeneous ontologies in their entirety may lead to inconsistencies or unintended inferences. In this paper we propose a set of guidelines for importing required terms from an external resource into a target ontology. We describe the guidelines, their implementation, present some examples of application, and outline future work and extensions.

Introduction

While the Web Ontology Language (OWL[1]) provides a mechanism to import ontologies (owl:imports), current limitations in tools and reasoners can sometimes make such a solution impractical on a day-to-day basis. First, some OWL tools (*e.g.,* Protégé, SWOOP) can neither load or reason over very large ontologies, such as the NCBI Taxonomy[2] or the Foundational Model of Anatomy[3], making direct OWL imports of such ontologies impractical. Second, different resources may have been constructed using different design principles, which may not align. Importing such ontologies as a whole could lead to inconsistencies or unintended inferences.

To address these issues, we have developed a set of guidelines for importing terms from multiple ontology resources, avoiding the overhead of importing the complete ontology from which the terms derive.

The Minimum Information to Reference an External Ontology Term (MIREOT) guidelines were created to aid the development of the Ontology of Biomedical Investigations (OBI[4]). OBI uses the Basic Formal Ontology (BFO[5]) as upper-level ontology and is part of the Open Biomedical Ontologies (OBO) Foundry[6]. One of the fundamental principles of the OBO Foundry is to reuse, where sensible, existing ontology resources, therefore avoiding duplication of effort and ensuring orthogonality. MIREOT allows us to do so by providing a way to import external terms from ontologies not yet using BFO as an upper ontology, or not yet using OWL DL.

Policy

In deciding upon a minimum unit of import, our first step was to consider the practice of other ontologies.

For example, in the Gene Ontology (GO[7]), the intended denotation of classes remains stable. Even when the ontology is repaired or reorganized, the effects of such changes do not change the intended meaning of individual terms. Rather the changes are towards more carefully expressing the logical relations between them. When a term's definition changes meaning, the term is deprecated[8]. We can therefore consider a term as stable, in isolation from the rest of the ontology, and use terms (i.e. individual classes in isolation from the ontology) as basic unit of import. The current implementation of MIREOT has been limited to import of terms from other Foundry ontologies, which adhere to a similar deprecation policy.

The minimum amount of information needed to *reference* an external class is the source ontology URI and the external term's URI. Generally, these items remain stable and can be used to unambiguously reference the external class from within the importing target ontology. The minimum amount of information to *integrate* this class is its position in the hierarchy, specifically the URI of its direct superclass in the target ontology.

Taken together, the following minimal set is enough to consistently reference an external term:

- **Source Ontology URI.** The logical URI of the ontology containing the external term to be imported.

- **Source Term URI.** The logical URI of the specific term to import.
- **Target Direct Superclass URI.** The logical URI of the direct asserted superclass in the target ontology.

To ease development of the target ontology we also recommend, although do not require, that additional information about the external class be added such as its label and textual definition.

Implementation

An implementation of the MIREOT guidelines was performed in the context of the OBI project, and can be decomposed into a two-step process:

1. Gather the minimum information for the external class.
2. Use this minimum information to fetch additional elements, like labels and definitions.

Once the external term is identified for import, the first step is to gather the corresponding minimum information set.

This set is stored in a file that we call *external.owl* (http://tinyurl.com/b7vvlt). A Perl script, *add-to-external.pl*[9] is used to automatically append the minimum information set to the *external.owl* file. This script takes as arguments the identifiers of the external class to be imported and its parent class in the target hierarchy, in this case in the OBI hierarchy.

In addition, a mapping mechanism between the prefix used in the identifier and the external source ontology URI is built into the script. Curators therefore need only specify the ID of the external class to import and the ID of the class it should be imported under, within the target ontology.

```
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix owl: <http://www.w3.org/2002/07/owl#>
prefix obi: <http://purl.obofoundry.org/obo/>
prefix obo: <http://www.geneontology.org/formats/oboInOwl#>

construct
{
  _ID_GOES_HERE_ rdf:type owl:Class.
  _ID_GOES_HERE_ alias:preferredTerm ?label.
  _ID_GOES_HERE_ rdfs:label ?label.
  _ID_GOES_HERE_ alias:definition ?definition.
}
where
{
  {   _ID_GOES_HERE_ rdfs:label ?label. }
  UNION
  {   _ID_GOES_HERE_ obo:hasDefinition ?blank.
      ?blank rdfs:label ?definition}
}
```

Figure 1. Template SPARQL query. For convenience, we use alias:preferredTerm and alias:definition to reference our annotations properties IAO 0000111 and IAO 0000115[11] respectively.

Additional elements can be obtained programmatically via SPARQL[10] CONSTRUCT queries, as described in Figure 1 (see also http://tinyurl.com/bss9mw). These queries specify which extra information about the class to gather, such as the definition and preferred label, and how to map these into the corresponding OBI annotation properties.

For example, in the current OWL rendering of OBO files, definitions are individuals and the rdfs:label of those individuals record the text of the definitions. Within the OBI implementation of the MIREOT guidelines, only annotation properties that map directly to our own metadata are mapped: new properties (*e.g.*, curation status annotation property, definition editor or definition source), if not specified in the source ontology, are not created.

The external term is directly imported from the external resource, with the status and definition as defined by the external resource. Finally, a script, *create-external-derived.lisp*[9], iterates through the minimum information stored in *external.owl*. Depending on the source ontology URI of each of our imported terms, it then selects the correct SPARQL template and substitutes the relevant ID. The queries are then executed against the Neurocommons SPARQL endpoint[12].

This supplementary information, which is prone to change as the source ontologies evolve, is stored in a second file, *externalDerived.owl* (http://tinyurl.com/bmb3f4). This file can be removed on a regular basis, e.g., before release of OBI. It is then rebuilt via script based on external.owl, allowing us to keep imported information up-to-date. The two files, external.owl and *externalDerived.owl*, are then imported by *obi.owl*, providing the necessary information to OBI editors while at the same time keeping it independent from OBI proper classes.

In the following sections we present two different cases of application of the MIREOT guidelines.

Use Case One – Cell Class

We replaced the OBI class *cell* with that from the Cell Type (CL) ontology[13]. CL is part of the OBO Foundry effort, and we would like to use the *cell* class as defined by this resource, instead of creating our own duplicated class. The following invocation of the *add-to-external.pl* script

perl add-to-external.pl CL:0000000 IAO:0000018

will add the class *cell* (CL:0000000) as subclass of the class *material entity* (IAO:0000018), and set the source ontology URI as http://purl.org/obo/owl/CL. Once imported, the cell class can be used as would be any other OBI class. For example, the process "electroporation" is defined as:

is_a cell permeabilization

has_specified_input some cell

has_specified_output some

(cell and has_quality some electroporated))

utilizes_device some power supply

More generally, additional axioms may be used to relate members of the class to other entities in the ontology.

Use Case Two – Taxonomic Information

The cell use case highlights what is likely to be the most common import scenario, i.e. a simple import of one external term, making it available for direct use in the target ontology. However, in some cases, we may require more, and to account for this MIREOT has been devised to be flexible.

OBI currently uses the NCBI taxonomy for its species terms. We can easily imagine that somebody would want to query a dataset asking the question "give me all experiments in mammals". In this case, we would need to know that human and mouse are subclasses (even indirect) of mammals in the NCBI taxonomy. Therefore, when mapping towards an NCBI term, it is needed to get the class itself and all its superclasses up to one of a set of top-level classes in the taxonomy.

```
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix owl: <http://www.w3.org/2002/07/owl#>
prefix obi: <http://purl.obofoundry.org/obo/>
prefix tax: <http://purl.org/obo/owl/NCBITaxon#NCBITaxon>

construct
 { ?super rdf:type owl:Class.
   ?super rdfs:subClassOf ?parent.
   ?super alias:preferredTerm ?label.
   ?super rdfs:label ?label.
 }
 where
  {
     {  #direct superclass.
        { _ID_GOES_HERE_ rdfs:subClassOf ?super. }
        { ?super rdfs:subClassOf ?parent.
          ?super rdfs:label ?label.
        }
     }
   UNION
        # now harvest the class annotations
      { ?super rdfs:subClassOf ?parent.
        ?super rdfs:label ?label.
       FILTER (?super=_ID_GOES_HERE_)
      }
   FILTER (!((?parent=alias:bacteria) || (?parent=alias:eukaryota) ||
     (?parent=alias:viruses) || (?parent=alias:archaea) ||
     (?parent = alias:cellularOrganism) || (?super=alias:bacteria) ||
     (?super=alias:eukaryota) || (?super=alias:viruses) ||
     (?super=alias:archaea) || (?super = alias:cellularOrganism)))
}
```

Figure 2. Template SPARQL query. For convenience, we use alias:preferredTerm and alias:definition to reference our annotations properties IAO 0000111 and IAO 0000115[11] respectively.

When the *create-external-derived.lisp* script parses the external.owl file and encounters an NCBI taxonomy ID, it will therefore invoke a specific SPARQL query (Figure 2). As per the mechanism described above, the minimum information about the imported external class (*e.g., Mus musculus*) is defined in *external.owl*, whereas the additional information (rank information - genus, kingdom, phylum, etc.) is stored in *externalDerived.owl*.

Discussion

The MIREOT standard is a trade-off between complete consistency checking and heavyweight importing versus lightweight importing but partial consistency checking. We are aware of and accept that by copying only parts of an ontology there is the risk that inferences drawn may be incomplete or incorrect: correct inference using the external classes is only guaranteed if the full ontologies are imported. When deciding to import an external term we review the textual definition and, if needed, talk with the original editor. As we are importing from OBO Foundry ontologies we have a community process for monitoring change, a shared understanding of the basics of our domain, and the intention to eventually share the same upper-level ontology. Therefore, we expect that terms will be deprecated if there is a significant change in meaning, and are flexible enough to adjust and update our import of terms as the other ontologies start enhancing their logical definitions.

Another consideration using this approach is the status of assertions made on external terms. In adding axioms such as the subclass axiom when importing the external term, the aim is to only assert true statements. If additional restrictions are required (for example in OBI, cell is the bearer of the role reagent role or specimen role), those should be stored in the target ontology: the *external.owl* and *externalDerived.owl* are meant to include only the imported information. We anticipate that some of the statements added by the target ontology may migrate to the source ontologies at some point in the future; a fruit of the collaborative nature of OBO Foundry ontology development.

Future Work

The current implementation of the MIREOT guidelines relies on command-line scripts, making it sometimes uncomfortable to use for our curators. Ideally, a Protégé[14] plug-in could be developed to improve the interaction between the curators and the tool and the implementation of the MIREOT guidelines.

In the future, we also expect to provide an option in the OBI distribution that replaces *external.owl* with *imports.owl*, a file of imports statements generated by extracting the ontology URIs mentioned in external.owl.

The MIREOT guidelines are currently being implemented by other ontologies, like the Vaccine Ontology (VO[15]), and we ultimately hope that combined feedback will allow us to perfect the mechanism.

Acknowledgements

In memory of our friend and colleague William Bug, Ontological Engineer.

The OBI consortium is (in alphabetical order): Ryan Brinkman, Bill Bug, Helen Causton, Kevin Clancy, Christian Cocos, Mélanie Courtot, Dirk Derom, Eric Deutsch, Liju Fan, Dawn Field, Jennifer Fostel, Gilberto Fragoso, Frank Gibson, Tanya Gray, Jason Greenbaum, Pierre Grenon, Jeff Grethe, Yongqun He, Mervi Heiskanen, Tina Hernandez-Boussard, Philip Lord, Allyson Lister, James Malone, Elisabetta Manduchi, Luisa Montecchi, Norman Morrison, Chris Mungall, Helen Parkinson, Bjoern Peters, Matthew Pocock, Philippe Rocca-Serra, Daniel Rubin, Alan Ruttenberg, Susanna-Assunta Sansone, Richard Scheuermann, Daniel Schober, Barry Smith, Larisa Soldatova, Holger Stenzhorn, Chris Stoeckert, Chris Taylor, John Westbrook, Joe White, Trish Whetzel, Stefan Wiemann, Jie Zheng.

The authors' work is partially supported by funding from the NIH (R01EB005034), the EC EMERALD project (LSHG-CT-2006-037686), the BBSRC (BB/C008200/1, BB/D524283/1, BB/E025080/1), the EU FP7 DebugIT project (ICT-2007.5.2-217139), and the Michael Smith Foundation for Health Research.

References

1. Web Ontology Language (OWL), http://www.w3.org/2004/OWL/.
2. Wheeler DL, Barrett T, Benson DA, *et al*. Database resources of the national center for biotechnology information. Nucleic acids research, 33(Database issue): D39–45, Jan 2005.
3. Golbreich C, Zhang S and Bodenreider O. The Foundational Model of Anatomy in OWL: Experience and perspectives. Web semantics (Online), 4(3):181–195, 2006.
4. OBI Ontology, http://purl.obofoundry.org/obo/obi.
5. Grenon P, Smith B and Goldberg L. Biodynamic ontology: Applying BFO in the biomedical domain. Studies in health technology and informatics, 102:20–38, 2004.
6. Smith B, Ashburner M, Rosse C, *et al*. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology, 25(11):1251–1255, Nov 2007.
7. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic acids research, 32(90001):D258–D261, Jan 2004.
8. Go editorial style guide, http://www.geneontology.org/GO.usage.shtml.
9. OBI scripts, http://obi.svn.sourceforge.net/viewvc/obi/trunk/src/tools/.
10. SPARQL Query Language for RDF, http://www.w3.org/TR/rdf- sparql- query/.
11. The Information Artifact Ontology (IAO), http://code.google.com/p/information-artifact-ontology/.
12. Neurocommons sparql endpoint, http://sparql.neurocommons.org/.
13. Bard J, Rhee SY and Ashburner M. An ontology for cell types. Genome biology, 6(2):R21, 2005.
14. The Protégé Ontology Editor and Knowledge Acquisition System, http://protege.stanford.edu/.
15. The Vaccine Ontology, http://www.violinet.org/vaccineontology/.

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

Towards Context-Driven Modularization of Large Biomedical Ontologies

Pinar Oezden Wennerberg, Sonja Zillner
Siemens AG, Munich, Germany

Abstract

Formal knowledge about human anatomy, radiology or diseases is necessary to support clinical applications such as medical image search. This machine processable knowledge can be acquired from biomedical domain ontologies, which however, are typically very large and complex models. Thus, their straightforward incorporation into the software applications becomes difficult. In this paper we discuss first ideas on a statistical approach for modularizing large medical ontologies and we prioritize the practical applicability aspect. The underlying assumption is that the application relevant ontology fragments, i.e. modules, can be identified by the statistical analysis of the ontology concepts in the domain corpus. Accordingly, we argue that most frequently occurring concepts in the domain corpus define the application context and can therefore potentially yield the relevant ontology modules. We illustrate our approach on an example case that involves a large ontology on human anatomy and report on our first manual experiments.

Introduction

Medical research and clinical practice deal with complex and heterogeneous data, which poses challenges to realizing applications such as medical image and text search. Thus, it becomes necessary to support these applications with machine processable, explicit medical knowledge e.g., about human anatomy, radiology or diseases. This knowledge can be acquired from biomedical domain ontologies and can be used in the application, for example in semantic medical image and text search. Semantic medical image search is, indeed, the context of this work that lies within the THESEUS MEDICO research project.

Based on experience throughout the MEDICO project, we have observed that biomedical ontologies are typically very complex and comprehensive. This characteristic makes it difficult to use them straightforwardly in the target application due to efficiency reasons. At the same time, not all of the knowledge contained is relevant for the application context. In most cases, there is a specific set of ontology concepts and relations that sufficiently provide the required information. Using only parts of the ontology that are relevant for the application in mind allows for a significant improvement in the efficiency.

In MEDICO one use case focuses on patients suffering from lymphoma in the neck area. Lymphoma is a type of cancer in lymphocytes and it is a systematic disease with manifestations in multiple organs. During its diagnosis and treatment imaging is done several times based on the use of different imaging modalities (e.g. CT scan, X-Ray, MRI). This makes the lymphoma use case particularly relevant for a flexible medical image and text search application.

In this paper we describe our first ideas on an approach for modularizing large biomedical ontologies and illustrate it on the example of a large and comprehensive ontology about human anatomy. It is based on identifying statistically most relevant ontology concepts from domain corpora, in our case a corpus on lymphomas. We argue that these concepts can be used to distinguish the parts of ontologies that are most relevant for the application context. These parts can potentially yield the ontology modules that provide sufficient knowledge for the purposes of the software application. The modules identified in this way will additionally be discussed with the clinical experts for quality assessment and relevance.

The rest of this paper is structured as follows. Next section presents the related work. We then proceed with describing the relevant sources with a focus on the domain corpus and explain the statistical analysis process. In the Module Identification subsection we discuss the first (manually) identified modules supporting our ideas and assumptions. These are displayed in form of sub-hierarchies accessible through the UMLS tree browser. The paper concludes with first observations, discussions and future work.

Related Work

In most cases the application scenario, the level of detail and the complexity of the medical knowledge determines the way how the modules should be identified. In other words, there is no well-defined or broadly accepted definition for the “one and only way” to modularize ontologies. On the contrary, many different approaches and techniques for ontology modularization have been implemented.[2,3,4] Most views agree that there is no universal way to

modularize ontologies and that the choice of a particular technique should be guided by the requirements of the considered application. Spaccapieta[5] and d'Aquin[6] provide a good overview of concepts and methods for achieving scalability through modularization of ontologies.

In general, ontology modularization can be addressed automatically or user-driven, but in both cases the modularization of the ontology is a challenging task. For example, ontology modularization approaches that guarantee logical consistency[7] may deliver too large fragments and can be slow in performance. On the other hand graph-based approaches[8] are more efficient but they do not guarantee the logical completeness. Finally, manually created ontology fragments[9] do naturally have the required level of granularity but they are expensive in terms of time and resources and are open to human errors.

The technique introduced in this paper has the objective to enable a semi-automatic identification of ontology modules and it does not prioritize completeness. Rather, we account for the practical applicability of the extracted modules to improve the efficiency of the application. Nevertheless, the extracted modules shall be discussed with clinical experts for quality assessment.

Materials and Methods

Foundational Model of Anatomy (FMA)[10] ontology is the most comprehensive machine processable resource on human anatomy. It covers 71,202 distinct anatomical concepts (e.g., 'Neuraxis' and its synonym 'Central nervous system') and more than 1.5 million relations instances from 170 relation types. In addition to the hierarchical is-a relation, concepts are connected by seven kinds of part-of relationships (e.g., 'part of', 'constitutional part of', 'regional part of' etc.) The version we currently refer to is the version available in March 2009.

PubMed Lymphoma Corpus

The lymphoma corpus is based on medical publication abstracts on lymphoma from PubMed[11] scientific abstracts database. Its purpose is to provide specific domain knowledge about lymphoma, as this is one major use case of MEDICO. To establish the corpus we first extracted a set of the lymphoma relevant concepts from the NCI Thesaurus and then used these to identify from PubMed most frequently reported lymphomas, which are 'Non-Hodgkin's Lymphoma', 'Burkitt's Lymphoma', 'T-Cell Non-Hodgkin's Lymphoma', 'Follicular Lymphoma', 'Hodgkin's Lymphoma', 'Diffuse Large B-Cell Lymphoma', 'Aids Related Lymphoma', 'Extranodal Marginal Zone B-Cell Lymphoma of Mucosa-Associated Lymphoid Tissue', 'Mantle Cell Lymphoma', 'Cutenous T-Cell Lymphoma'. For each lymphoma type we compiled a set of XML documents that are generated from the PubMed abstracts. The text sections of the XML files were run through the TnT part-of-speech parser to extract all nouns and adjectives in the corpus.

The reason for including adjectives is based on our observations with the concept labels. Especially for the anatomy domain, the adjectives carry information that can be significant for medical decisions, for example, when determining whether an image is related to the *right* or to the *left* ventricle of the heart. Therefore, throughout the paper, when we talk about concepts, we refer to both adjectives and nouns. Then a relevance score (chi-square) for each noun and adjective was computed by comparing their frequencies in the domain specific corpus with those in the British National Corpus (BNC).[12] The resulting corpus consists of 71.973 files.

Statistical Analysis of Concepts

The objective of the statistical analysis is to identify a set of concepts that are most relevant for the application and for the use case. The process starts with converting the ontology into a flat list of concepts after some filtering is applied to the concept labels in the same way as explained in our previous work[13]. The statistically most relevant concepts are then identified on the basis of chi-square scores computed for nouns and adjectives. Ontology concepts that are single words and that occur in the corpus, correspond directly to the noun/adjective that the concept is build up of. For example, the noun 'ear' from the corpus corresponds to the FMA concept 'Ear'. Thus, the statistical relevance of the ontology concept is the chi-square score of the corresponding noun/adjective.

In the case of multi-word ontology concepts, the statistical relevance is computed on the basis of the chi-square score for each constituting noun and/or adjective in the concept name, summed and normalized over its length. Thus, relevance value for 'Lymph node', for example, is the summation of the chi-square scores for 'Lymph' and 'node' divided by 2. In order to take frequency into account, we further multiplied the summed relevance value by the frequency of the term. This assures that only frequently occurring terms are judged as relevant. A selection from the list of most relevant FMA concepts in the corpus is shown below (Table 1). We only focused on the Mantle Cell Lymphoma collection of the PubMed corpus, however currently we are extending the profiles to the rest.

FMA	Score
1. normal cell	240175,31
2. cell morphology	197495,31
3. stem cell	193389,88
4. plasma cell	190968,82
5. cell membrane	189984,02
6. cell surface	189981,54
7. lymphoid tissue	152765,58
8. lymph	99856,00
9. immunoglobulin	53361,00
10. inguinal lymph node	34943,38

Table 1. 10 most relevant FMA terms in the PubMed corpus

Identification of Potential Modules

Module identification starts with locating the statistically most relevant ontology concepts in the ontology hierarchy. The work reported in this paper was done manually. For the first experiments, we examined the context of the three concepts; 'Inguinal lymph node', 'Plasma cell' and 'Plasma membrane'. To locate the concepts in the hierarchy we used the UMLS Knowledge Server and selected the FMA view. We then searched for the three concepts using exact match. The hierarchical contexts are displayed below (Figure 1, Figure 2, Figure 3).

The locations of the 'Inguinal lymph node', 'Plasma cell' and 'Plasma membrane' in the FMA hierarchy display the 'Anatomical structure' as their next common parent. Therefore, it is most likely to be the root of the ontology module. The sum of the shortest paths from each concept (i.e. 'Inguinal lymph node', 'Plasma cell' and 'Plasma membrane') to 'Anatomical structure' will, in this case, be appended to it as its children. The sub-hierarchy consisting of 'Anatomical structure' as root and 'Cardinal organ part', 'Cell' and 'Cardinal cell part' as its children (and the children's descendants) may then be the potential ontology module. Consequently, the expectation from this module would be that it contains sufficient information about anatomy that relates to lymphoma.

First Observations and Discussion

The concept labels reveal lexical overlaps. This suggests that further interrelations can be discovered by comparing the descendant concept labels at the lexical level. In this way, we expect to be able to find lexical correspondences that potentially convey further useful hierarchical information.

One drawback we have observed is that the ontology modules can be rather large. This means that it would be hard to identify the focus of the module. One possible way to avoid too large and too generic ontology modules may be by allowing only a certain number of concepts that were identified as statistically relevant and then by locating only these in the hierarchy. We currently investigate this.

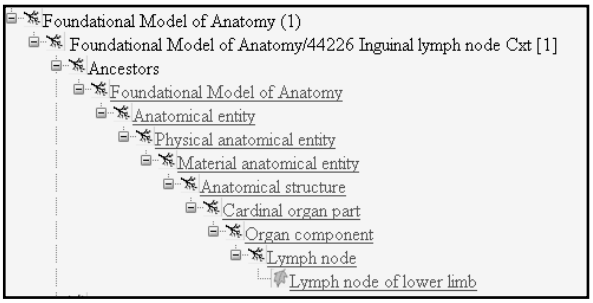


Figure 1. Hierarchical context of 'Inguinal lymph node' in the FMA view of the UMLS tree browser.

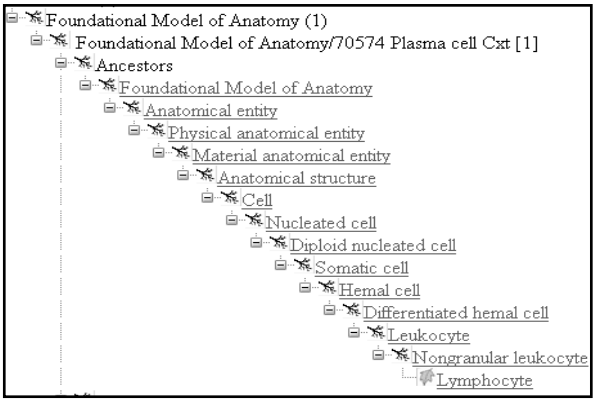


Figure 2. Hierarchical context of 'Plasma cell' in the FMA view of the UMLS tree browser.

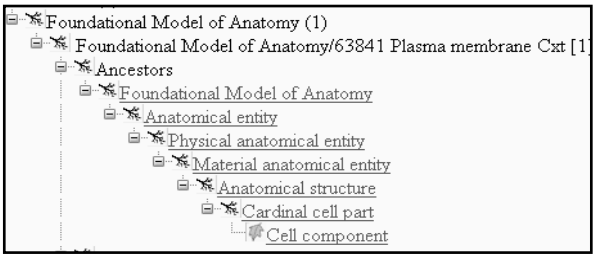


Figure 3. Hierarchical context of 'Plasma membrane' in the FMA view of the UMLS tree browser.

Future Work

As next the semi-automatic identification of the ontology modules will be realized. UMLS Knowledge Source Server and other tools from the National Library of Medicine[14] can support this process. Once this is achieved, it becomes relevant to identify the correspondences between the sub-hierarchies. Lexical methods e.g., string similarity and overlap detection, can be used to discover correspondences between the concept labels. A long term but an important research question concerns finding an effective strategy to identify the optimal size for each module. It is essential to be able to

determine when to terminate appending children to the module hierarchy. This is a challenging task, as optimal size, logical completeness and consistency usually require compromising.

Acknowledgements

This research has been supported in part by the THESEUS Program in the MEDICO Project, which is funded by the German Federal Ministry of Economics and Technology under the grant number 01MQ07016. The responsibility for this publication lies with the authors. We are also thankful to our clinical partner Dr. Alexander Cavallaro of the University Hospital Erlangen for his expert contribution.

References

1. UMLS Knowledge Source Server. Available online: http://umlsks.nlm.nih.gov/uPortal/
2. d'Aquin M, Schlicht A, Stuckenschmidt J and Sabou M. Ontology Modularization for Knowledge Selection: Experiments and Evaluations. 2007: In Proc. of the 18th Int. Conference.
3. Noy N and Musen M. Specifying Ontology Views by Traversal. 2004: In: Proc. of the Int. Semantic Web Conference.
4. Stuckenschmidt J and Klein M. Structure-Based Partitioning of Large Concept Hierarchies. 2004: In Proc. of the Int. Semantic Web Conference.
5. Spaccapieta S. Report on Modularization of Ontologies. 2005: Knowledge Web Deliverable 2.1.3.1.
6. d'Aquin M, Sabou M and Motta E. Modularization: A Key for the Dynamic Selection of Relevant Knowledge Components. 2006: In Proc. of the Workshop on Modular Ontologies.
7. Cuenca G.B, Horrocks I, Kazakov Y and Sattler U. Just the right amount: extracting modules from ontologies. In Proc. of the 16th international conference on WWW, 2007:717-726 NY, ACM.
8. Jimenez-Ruiz E, Berlanga R, Nebot V and Sanz I. Ontopath: A language for retrieving ontology fragments. In Meersman R and Tari Z, (eds.), OTM Conferences (1) 2007;4803: 897-914.
9. Zillner S, Hauer T, Rogulin D, Tsymbal A, Huber M and Solomonides T. "Semantic Visualization of Patient Information." 2008: In Proceedings of the 21st IEEE Int. Symposium on Computer-Based Medical Systems.
10. Mejino JL, Rubin DL and Brinkley JF. FMA-RadLex: An application ontology of radiological anatomy derived from the foundational model of anatomy reference ontology. 2008. Proc. of AMIA Symp:465–469.
11. PubMed Central. Available online: http://www.ncbi.nlm.nih.gov/pubmed/
12. British National Corpus. Available online: http://www.natcorp.ox.ac.uk/
13. Oezden Wennerberg P, Buitelaar P and Zillner S. *Towards a Human Anatomy Data Set for Query Pattern Mining Based on Wikipedia and Domain Semantic Resources*. 2008: In Workshop on Building and Evaluating Resources for Biomedical Text Mining at LREC.
14. National Library of Medicine, http://www.nlm.nih.gov

Debugging Mappings between Biomedical Ontologies: Preliminary Results from the NCBO BioPortal Mapping Repository

Jyotishman Pathak, Christopher G. Chute
Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN, USA

Abstract

The ability to provide semantic mappings between multiple large biomedical ontologies is considered as a very important, albeit labor-intensive and error-prone task. To facilitate such a process, several approaches for collaborative ontology mapping building and sharing have been proposed in the recent past. However, despite the improvements in community-wide mappings development, more often the mapping rules are redundant, incoherent, and at times, incorrect. In this paper, we present an approach for identifying such "erroneous mappings" using Distributed Description Logics. Specifically, we illustrate how logical reasoning can be used to discover semantic inconsistencies caused by erroneous mappings, and provide preliminary results of experiments based on the National Center for Biomedical Ontology BioPortal mapping repository.

Introduction

The ability to specify semantic mappings between biomedical ontologies is an important research agenda in the medical informatics community. Several approaches have been proposed for alignment between ontologies ranging from entirely manual[1], to semi-automatic[2,3], to fully-automatic[4] mapping techniques, many of which have met with varying degrees of success. More recently, with the growing number of ontologies in the biomedical domain, and hence the increasing requirement for their alignment, community-based approaches to create mappings have been proposed that allow users and domain experts to specify semantic correspondences in a collaborative manner[5,6]. However, despite these advancements, an important limitation of the existing efforts is the lack of ability to identify, debug, and invalidate semantically inconsistent mappings (or erroneous mappings). As mentioned by Noy et al.[5], such a requirement is vital because in many cases a concept definition may change with a new version of the ontology, and thereby making an existing mapping invalid, or users may add new or delete existing mappings that result in the aligned ontologies becoming logically inconsistent.

Toward this end, we propose a technique for identifying erroneous mappings between biomedical ontologies. In particular, we exploit the underlying semantics of the mappings as well as the mapped ontologies based on Distributed Description Logics (DDL)[7] to pinpoint mappings that are logically inconsistent[8]. Our basic assumption is that a mapping that correctly states the semantic correspondences between the ontology concepts should not cause inconsistencies in the mapped ontologies. The advantage of using DDL is that it allows us to detect such inconsistencies, which can then be regarded as symptoms caused by erroneous mappings. For example, Figure 1 shows two equivalent mappings between concepts A and B in ontology O_1 (source) with concepts C and D in ontology O_2 (target), respectively. Furthermore, B is asserted as a subConcept of A in O_1, whereas both C and D are asserted as disjoint from each other in O_2. Assuming that both the mappings are valid as well as ontologies O_1 and O_2 are logically consistent, one can infer (via global interpretation) that the concept D should be a subConcept of C in O_2. However, since they are asserted as disjoint in O_2, thereby causing a logical inconsistency, implies that at least one of the mappings is erroneous—identification of which is our objective. Specifically, the main contributions of the proposed work are:

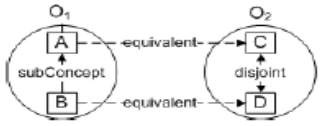


Figure 1: Inconsistent Mappings Example

- We leverage DDL[7] and ontology mapping repair techniques[8] to describe a formal framework for identifying erroneous biomedical ontology mappings.
- We illustrate the applicability of our approach by experimenting with the NCBO BioPortal mapping repository[5] and provide preliminary results.
- We provide an open-source prototype implementation of our software based on the DRAGO distributed reasoning system: http://code.google.com/p/bioontologies-mapping-debugger.

Background

Distributed Description Logics (DDL)[7] is a knowledge representation formalism for representing sets of ontologies and semantic relations between them. It provides a mechanism for referring to ontologies and for defining rules that connect "concepts" in different ontologies. This is achieved using the notion of importing and reusing concepts between ontologies and enabling reasoning with multiple ontologies connected by directional semantic mapping (called the bridge rules). In particular, DDL extends the notion of interpretation introduced above to fit the distributed nature of the model and to reason about concept subsumption across ontologies.

More formally, let I be a set of non-empty indices such that $\{O_i\}_{i\in I}$ is a set of ontologies. Concepts and axioms are represented with the index of the ontology they belong to such that i:C denotes a concept in ontology O_i and j:C⊆D represents that concept C is a sub-concept of D in ontology Oj , where i:C and j:C are different concepts. Semantic relations between concepts in different ontologies are represented via axioms, called bridge rules that are of the following form: (1) i:C → j:D (into-rule); and (2) i:C ← j:D (onto-rule); where, C and D are concepts in ontologies O_i and O_j , respectively. Furthermore, the derived bridge rule i:C≡j : D can be defined as a conjunction of the into- and onto-bridge rules. These rules do not represent the semantic relations stated from an external observation point of view such as the Web. Instead, a rule i to j expresses relations between i and j viewed from j-th subjective point of view. Specifically, an into-bridge rule i:C→j:D states that, from j-th point of view, the concept C in i is less general that its "local" concept D. Equivalently, the onto-relation i:C←j:D expresses the more generality relation. In general, note that the into-rule (i:C→j:D) is not necessarily an inverse of the onto-rule (i:C←j:D) since these rules reflect a subjective point of view. Thus, a "distributed ontology" D_{OR} can now be defined as a tuple, $(\{O_i\}_{i\in I}, \{R_{ij}\}_{i\neq j\in I})$, where $\{O_i\}_{i\in I}$ is the set of ontologies, and $\{R_{ij}\}_{i\neq j\in I}$ is the set of bridge rules between those ontologies.

An important aspect of DDL is that for the fundamental reasoning services of verification of consistency and concept satisfiability, in addition to the ontology itself, the reasoning depends on other ontologies to which it has semantic mappings. This is due to the ability of the bridge rules to transitively propagate knowledge across ontologies in the form of subsumption axioms as illustrated in Figure 2.

The main objective of our work is to leverage DDL[7] and existing techniques for repairing ontology mappings[8] to provide a formal framework for identifying erroneous mappings between biomedical ontologies. In what follows, we formalize ontology mappings (with respect to DDL) and outline steps for identifying erroneous mappings.

$$\frac{i:A\subseteq B, i:A\longrightarrow j:G, i:B\longrightarrow j:H}{j:G\subseteq H}$$

Figure 2: Subsumption Propagation of DDL Bridge Rules

Mappings and Correspondences: At an abstract level, a mapping between source O_i and target ontologies O_j is defined via a set of correspondences, where each correspondence represents a semantic relation between concepts in O_i and O_j.

Definition 1 (Semantic Correspondence): Given ontologies O_i and O_j , a semantic correspondence can be represented (minimally) by a 3-tuple <C,C′,r>, such that C∈F(O_i), C′∈F(O_j), and r is a semantic relation, where F is a function that identifies elements in O_i and O_j. Furthermore, in this work, we restrict r to the set {≡, ⊆, ⊇}, essentially limiting to equivalence and subsumption. Given a set of semantic correspondences, we can define the notion of a mapping as a collection of such correspondences.

Definition 2 (Ontology Mapping): Given ontologies O_i and O_j, M is a mapping between O_i and O_j , iff for all correspondences <C,C′,r>∈M, we have C∈F(O_i), and C′∈F(O_j).

To formalize ontology mappings in terms of DDL presented earlier, we encode the semantic correspondences as bridge rules. In particular, each correspondence <C,C′,r> between a pair of ontologies O_i and O_j is translated into a bridge rule via a translation function T as follows:

$$T(\langle C,C',\subseteq\rangle)\equiv i:C\longrightarrow j:C' \wedge j:C'\longrightarrow i:C$$
$$T(\langle C,C',\supseteq\rangle)\equiv i:C\longrightarrow j:C' \wedge j:C'\longrightarrow i:C$$

Inconsistent Mappings. A mapping M of a distributed ontology ℑ can be defined as inconsistent with respect to a particular concept i:C if it becomes unsatisfiable modulo the mappings

Definition 3 (Mapping Consistency): Given a distributed ontology ℑ, the mapping M between ontologies Oi,Oj∈ℑ is consistent with respect to a concept i:C iff concept C is unsatisfiable in O_i implies that i:C is also unsatisfiable in ℑ. Otherwise, M is inconsistent with respect to i:C. By extrapolation, M is consistent with respect to O_i iff for all i:C, M is consistent with respect to i:C; otherwise M is inconsistent with respect to O_i.

For example, based on Figure 1, M = {O_1:A ≡ O_2:C, O_1:B ≡ O_2:D}. Furthermore, by applying

distributed reasoning it can be inferred that O_2:D⊆C should hold. However, at the same time both C and D are defined as disjoint concepts in O_2, thereby making M inconsistent with respect to D since it cannot be satisfied in the global interpretation. Algorithm 1 follows directly from Definition 3 which also states that the inconsistency of one ontology, or some sub-group of connected ontologies, does not automatically render the entire distributed ontology inconsistent. Arguably, the goal is to determine an erroneous mapping set and identify which of the semantic correspondences involved can be removed to maintain consistency. In particular, we want to determine a "minimal erroneous mapping set" which has the property that none of its subset is an erroneous mapping set.

Algorithm 1 Identification of Mapping Inconsistency

```
1: procedure ISCONSISTENT((ℑ = ({O_i}_{i∈I}, {ℛ_ij}_{i≠j∈I})), i)
2:   for all concepts i : C ∈ T_i do
3:     if (T_i ⊭ C ⊑ ⊥) and (ℑ ⊨ i : C ⊑ ⊥) then
4:       return false
5:     end if
6:   end for
7:   return true
8: end procedure
```

Evaluation

Materials. We evaluated our methods proposed above using the NCBO BioPortal mappings repository. As stated in Noy and Musen[5], the inability to impose any quality control on the mappings that the users submit is a limitation of the existing BioPortal infrastructure, and our work provides preliminary steps in addressing this requirement.

At the time of our evaluation, the repository contained approximately 30,000 mappings between various biomedical ontologies, and a majority of these mappings were between the Open Biomedical Ontologies (OBO) and Web Ontology Language (OWL 1.0) ontologies. Since our technique for inconsistency detection has been implemented on top of the DRAGO distributed reasoning system, which is an OWL-DL based reasoner, we transformed all the mapped OBO ontologies into OWL ontologies via the OBO-in-OWL Protege plugin. Furthermore, the mappings in the BioPortal repository do not use "true" logical equivalence (e.g., owl: equivalentClass), but rather the notion of "similarity"[5]. Since such a weaker definition of equivalence is not modeled in DDL, we transformed each "similar" mapping into an equivalence (≡), into (⊆), and onto (⊇) bridge rules for experimentation.

Results

Table 1 shows the results of our evaluation. From the mapping repository, we chose only those mapped ontologies which had at least 2 or more mappings specified between them. We also did not include mappings involving the Foundational Model of Anatomy (FMA) and International Classification of Diseases (ICD-9) because the current release of DRAGO (version 2.1) does not support nominals (e.g., owl:oneOf, owl:hasValue constructs) present in FMA, and there is no ClaML (Classification Markup Language used to represent ICD-9) to OWL transformer available, respectively. Furthermore, the columns L-Satisfiable and D-Satisfiable in Table 1 represent the total number of classes found satisfiable in the target ontology that are determined by the local axioms of the ontology (localized reasoning) and by propagation of the axioms via mappings (distributed reasoning), respectively.

Discussion

Result Analysis. For mappings between OBO ontologies, no inconsistencies were found. We believe this can be attributed to the fact that none of the evaluated OBO ontologies had disjoint class axioms, and hence none of the mappings were conflicting. Similarly, for mappings between OBO and OWL ontologies, no inconsistencies were observed even though the two original OWL ontologies that were evaluated, Nano Particle Ontology (NPO) and NCI-Thesaurus (NCI-T), had 12,265 and 171 disjoint class axioms, respectively. We believe that the lack of mapping inconsistency can be attributed to: (i) for many mappings, the classes from the disjoint class axioms were not involved, and (ii) for those mappings where such classes were involved, the mappings were logically correct. For example, NPO and ChEBI had the mappings npo:Gold≡chebi:CHEBI 29287 and npo:Carbon≡chebi:CHEBI 27594, such that npo:Gold is disjointWith npo:Carbon, and the classes CHEBI 29287 and CHEBI 27594 (with labels gold and carbon, respectively) had no associations between them. Consequently, there was no conflict in the mappings as well. Finally, due to performance issues, we were not able to evaluate mappings between original OWL ontologies (namely, Galen and NCI-T).

Limitations and Further Work. As mentioned earlier, in this work we limited our scope to one-to-one concept mappings, and further considered only equivalence and subsumption mappings. However, in reality, it is possible to specify arbitrary mappings e.g., disjoint) between any ontological entities (e.g., relationships) and the ability to consider such mappings to find inconsistencies becomes vital.

Source Ontology	Target Ontology	Mapping Type	# Mappings	# L-Satisfiable	# D-Satisfiable
Cereal Plant Trait (OBO)	Plant Environmental Conditions (OBO)	$\equiv, \sqsubseteq, \sqsupseteq$	3	506 (n=506)	506 (n=506)
Phenotypic Quality (OBO)	NCI-Thesaurus (OWL)	$\equiv, \sqsubseteq, \sqsupseteq$	4	66726 (n=66726)	66726 (n=66726)
Nano Particle Ontology (OWL)	ChEBI (OBO)	$\equiv, \sqsubseteq, \sqsupseteq$	4†	21377 (n=21377)	21377 (n=21377)
Cell Type (OBO)	Fungal Gross Anatomy (OBO)	$\equiv, \sqsubseteq, \sqsupseteq$	10	71 (n=71)	71 (n=71)
Molecule Role (OBO)	ChEBI (OBO)	$\equiv, \sqsubseteq, \sqsupseteq$	21	21377 (n=21377)	21377 (n=21377)
Zebrafish (OBO)	Mouse Adult Gross Anatomy (OBO)	$\equiv, \sqsubseteq, \sqsupseteq$	145	2877 (n=2877)	2877 (n=2877)
Galen (OWL)	NCI-Thesaurus (OWL)	$\equiv, \sqsubseteq, \sqsupseteq$	271	N/A	N/A
Mouse Adult Gross Anatomy (OBO)	NCI-Thesaurus (OWL)	$\equiv, \sqsubseteq, \sqsupseteq$	2870	66726 (n=66726)	66726 (n=66726)
Human Disease (OBO)	NCI-Thesaurus (OWL)	$\equiv, \sqsubseteq, \sqsupseteq$	6883	66726 (n=66726)	66726 (n=66726)

† One of the mappings between Nano Particle Ontology and ChEBI was deemed invalid because the class in the source ontology did not exist.

Table 1: `BioPortal` Mapping Evaluation Results

Furthermore, in the current evaluation, we took a snapshot of the mapping repository, thereby not considering how different versions of an ontology will affect the associated mappings. In future, we plan to evaluate how mapping consistency and satisfiability results vary with the evolution of the ontologies. Another limitation of our work is the complexity of the reasoning procedure. DDL subsumption reasoning has been shown to be NEXPTIME[7], thereby significantly impacting the efficiency of the consistency checking process. For example, evaluating the mappings between GALEN and NCI-Thesaurus was not feasible as the program runs out of memory (with a maximum Java heap space of 4GB). Hence, our objective is to leverage approximate reasoning services that apply correct but incomplete heuristics for performance gain[9].

Complementary to our work, the problem of identifying erroneous mappings has been addressed using the notion of a "global ontology"[10]. Consequently, reasoning is done with respect to the global ontology which, in certain cases, can result in increased complexity compared to distributed reasoning that exploits the structure provided by semantic relations for the propagation of reasoning through the local ontologies. However, there are no studies verifying this hypothesis, and our goal is to adapt our approach for such an investigation. Finally, our work raises the issue of evaluating "similarity" mappings between simple ontologies because, for example, in the absence of disjoint class axioms in both source and target ontologies, the mappings, although logically consistent, may still represent incorrect knowledge. We believe this can be partially addressed by leveraging the subsumption propagation of DDL (Figure 2) to create a distributed hierarchy which can be evaluated for correctness and accuracy, although such a proposal warrants further research.

References

1. Vikstrom A, Aner YS, Strender LE and Nilsson GH. Mapping the Categories of the Swedish Primary Health Care Version of ICD-10 to SNOMED CT Concepts. BMC Medical Informatics and Decision Making. 2007;7(9).
2. Noy NF and Musen MA. PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment. 17th National Conference on Artificial Intelligence. AAAI Press. 2000:450–5.
3. Fung KW, Bodenreider O, Aronson AR, Hole WT and Srinivasan S. Combining Lexical and Semantic Methods of Inter-terminology Mapping Using the UMLS. In: 12th World Congress on Health (Medical) Informatics. vol. 129. IOS Press; 2007. p. 605–609.
4. Doan A, Madhavan J, Dhamankar R, Domingos P and Halevy A. Learning to Match Ontologies on the Semantic Web. VLDB Journal. 2003;12(4):303–319.
5. Noy NF, Griffith N and Musen MA. Collecting Community-Based Mappings in an Ontology Repository. 7th International Semantic Web Conference. (LNCS 5318) 2008:371–386.
6. Correndo G, Alani H and Smart PR. A Community based Approach for Managing Ontology Alignments. In: 3rd International Workshop on Ontology Matching. vol. 431. CEURWorkshop Proceedings; 2008. p. 61–72.
7. Borgida A and Serafini L. Distributed Description Logics: Assimilating Information from Peer Sources. J. of Data Semantics. 2003;1:153–184.
8. Meilicke C, Stuckenschmidt H and Tamilin A. Repairing Ontology Mappings. In: 22nd AAAI Conference on Artificial Intelligence. AAAI Press; 2007. p. 1408–1413.
9. Meilicke C and Stuckenschmidt H. Applying Logical Constraints to Ontology Matching. In: 30th Annual German Conference on AI. Springer-Verlag, LNCS 4667; 2007. p. 99–113.
10. Cardillo E, Eccher C, Serafini L and Tamilin A. Logical Analysis of Mappings between Medical Classification Systems. Artificial Intelligence: Methodology, Systems, and Applications. 2008. 311–21.

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

Towards Ontological Facilitation of Standards-Compliant Data Capture and Reposition

Philippe Rocca-Serra[1*], Chris F. Taylor[1,2*], Marco Brandizi[1], Eamonn Maguire[1], Nataliya Sklyar[1], Susanna-Assunta Sansone[1]
[1]The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
[2]The NERC Environmental Bioinformatics Centre, CEH, Oxford, UK
* These authors contributed equally to this work.

Abstract

In recent years, bioscience communities centered on particular areas of study, or groups of technologies, have generated so-called Minimum Information (MI) checklists specifying the data and metadata that should be captured from the totality of information generated in the course of an investigation. In parallel, ontologies, formats, data capture tools and databases have been developed that can support the collection, validation, archiving and sharing of MI checklist-compliant data sets. In this paper, we discuss our use of the ontologies as facilitators of much of the above functionality: to semantically enrich data sets both at the point of capture and subsequently; as part of a rule-based content validation system; and to increase the speed and resolution of database queries.

Introduction

The general trend to enrich all manner of information by attaching machine-readable tags to as much of it as possible manifests in many forms: extended markup languages; folksonomy-style labels on blog entries; ontology terms appended to scientific papers [e.g.,1,2]. In keeping with the trend, some biological databases [e.g.,3,4] have employed terminological artifacts of varying degrees of sophistication (hereafter commonly referred to as ontologies) to support more precisely targeted querying with fewer false negatives.

However, there is a looming logistical issue: once funders and journals begin to require that firstly, researchers make more of their data public, and secondly, that they begin to comply with minimum information (MI) checklists[5,6,7], experimental data will begin to flow into databases at a much greater rate and there will not be enough curators to annotate it all. The only feasible solution is better annotation at source (i.e., by data generators), assisted by some form of automated content validation.

MI checklists, such as MIAME[8], MIAPE[9] and MIGS[10], each specify the information that should be provided when reporting a particular type of biological/biomedical investigation (the examples given addressing microarray-, proteomics- and genomics-based investigations respectively). By requiring a thorough and regularized description of an experimental workflow (for example, by requiring an explicitly specified set of sample characteristics or instrument parameters) MI checklists promote transparency and data accessibility, and support more thorough quality assessment, increasing the value of data set, and by extension the competitiveness of the originators and the host database(s).

However, because these various MI checklists have been developed independently, focusing on one specific technology, they overlap in scope. Furthermore, arbitrary decisions on wording and structure guarantee significant incompatibilities. Until recently there were no mechanisms to coordinate checklist development. Even attempting to establish the range and number of checklists was challenging, and tracking their evolution rather laborious.

Nowadays, researchers are able to perform biological/biomedical investigations where the same sample is run through the full range of technologies, in combination. In this specific case, it is critical that the MI checklists and ontologies developed around a specific technology or biological/biomedical domain are designed to be interoperable and fit neatly into a jigsaw, with users being able to take the pieces that are relevant to report their investigation.

Recently, representatives of a number of checklist development projects began the MIBBI (Minimum Information for Biological and Biomedical Investigations) project[11]. MIBBI has two broad goals: (i) the Portal, to provide straightforward access to checklists and their developers, acting as a 'one-stop shop' in a manner analogous to the Open Biomedical Ontology portal[12]; (ii) the Foundry, to foster the development of new, orthogonal suite of checklist modules, just like the OBO Foundry[13] does for the ontologies.

These new integrated modules will act as drivers for the development of data capture software, exchange

formats, and repositories that can contain the information they specify, where none already exists.

Towards Ontology-Aware and Standards Compliant Tools

An important feature of these new MIBBI modules will be their association with mappings that make an explicit link between particular line items and (sets of) terms in (one or more) ontologies. These mappings will assist software and database developers providing the ontology aware, checklist supportive data capture tools and repositories that will provide the main mechanism for annotating standards-compliant data sets using ontology terms. Such tools will guide users compiling investigation reports; for example, by highlighting those fields mandated by the appropriate MIBBI modules, and by offering a simple validation mechanism to ensure that all those fields contain data of the correct type. However, to support ontology-based annotation of data sets, a mapping is required that explicitly links particular data fields to (parts of) individual ontologies.

When implemented in software, mappings simplify the process of finding and using appropriate ontology terms for users, and ensure that the terms used in across the swathes of reports in repositories of research data come from comparable sources. Such consistency in annotation is important for supporting efficacious querying across sets of reports.

An ontology-aware data capture tool can also perform more sophisticated validation than simply checking data types according to an XML Schema definition or content according to a series of regular expressions[14]. In essence, ontology-based validation uses the MI checklist-to-ontology mapping alongside the mapping from the MI checklist to the data format (if validation is performed by a standalone tool) or database schema with which the data are classified. Valid entries contain ontology terms deemed 'appropriate' by the mapping (and by extension, do not contain inappropriate terms). Of course, mappings need to be updated as ontologies grow and evolve.

Conceptual Design and Preliminary Work

In this section we present our initial work and the overall conceptual design for an integrated, ontology and standards aware data capture and management system, built around MIBBI modules. This work is part of the Investigation / Study / Assay (ISA) Infrastructure[15], a new infrastructure to commonly represent, store and serve experimental metadata (including experimental design, sample source(s) and treatment(s), preparation of a sample for analytical assay, the processes and instruments used throughout, and sample-data file relations). The ISA infrastructure is based upon ISA-Tab[16], a general purpose, common framework with which to communicate experimental metadata. The infrastructure includes several open source Java software components[17], which can work independently, or as a unified system, including:

- ISAcreator, which draws on spreadsheets for its look and feel, to capture and edit experimental metadata;
- ISAconfigurator to manage the ISA-Tab fields displayed in ISAcreator; for example, by making them mandatory and/or requiring the use of ontologies);
- The BioInvestigation Index (BII) database for storing and querying ISA-Tab formatted metadata

Figure 1 shows these components, their interrelations, inputs and outputs.

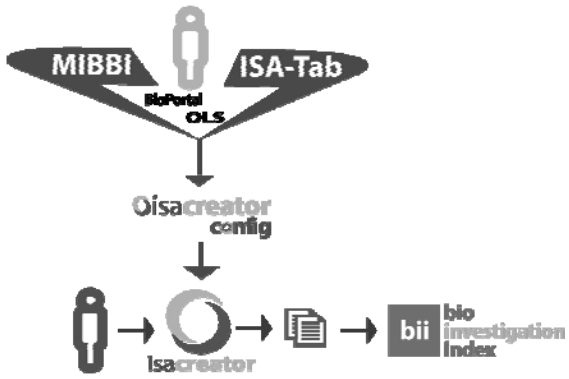


Figure 1. A 'power user' (top of figure) creates a configuration file to match ISA-Tab fields to the requirements of the MIBBI modules relevant to a particular workflow, tying in a specific set of ontologies along the way. ISAcreator then loads the configuration and guides the user (bottom left) in describing the experimental metadata for submission to the BII database, which can be queried using ontology terms.

Module content is taken from MIBBI and mapped by a 'power user' (i.e., curator) into ISA-Tab elements and one or more external ontologies, using the ISAconfigurator. The resulting configuration file is then uploaded in ISAcreator and corresponding interface components are automatically generated, enforcing mandatory fields and/or use of ontology terms, searched and selected from OBO Foundry ontologies accessed in real time via Ontology Lookup Service (OLS)[18] and BioPortal[19] public portals. With the appropriate mappings in place (i.e., those that correspond to the MIBBI modules appropriate to the particular biological/biomedical

investigation from which the experimental metadata is being captured), the ISAcreator tool is ready to be distributed to users (i.e., data generators) and used as an electronic lab-book.

The MIBBI mappings in ISAcreator allow the interface to guide submitters to supply all required MI requirements, and the ontology mappings will assist them in semantically enriching the description of the experimental metadata (N.B., adding ontological annotations to a data set is not normally a requirement made by funders or journals, so the process needs to be as efficient and painless as possible for users).

On request, a validation routine checks not only for a syntactically valid document (i.e., all the right data types in all the right places), but also, to a limited degree, for a *semantically* valid document. It is hoped that this double mechanism will assist data generators in compiling submissions that require little further attention from database curators, most of the gross errors having been caught.

Upon completion of a syntactically and semantically valid investigation report, ISAcreator then outputs the experimental metadata as an ISA-Tab file along with the associated data files in a compressed 'ISArchive', ready for upload to the BII database[20] or any other system that imports ISA-Tab files. Once the investigation is loaded in the database, ontology terms then serve a second, more familiar role as searchable tags on the experimental metadata. This allows precise queries to be formulated with some confidence that in most cases a search for a spade will find all manner of large digging implements, rather than falling foul of the usual quasispecies of variously-typo-ridden variants.

Conclusions

Two inexorable trends – to submit richly annotated biological/biomedical investigations, and to submit them in greater numbers – present software and database developers with a significant challenge: given that database curators are not able to manually check and reannotate the many complex submissions to come, greater automation of various processes is required.

Although some ontology-aware tools exist [21, 22], more sophisticated solutions such as ISAcreator are needed that implement mappings from reporting MI guidelines to data formats and ontologies, reducing the error load of new submissions to databases. The ontology terms that were so useful for compiling and validating files (either by the submitter on completion, or by the curator on receipt) then become useful in the database setting in their traditional role – as hooks with which to retrieve data.

Acknowledgements

We gratefully thank the BBSRC (BB/G000638/1, BB/E025080/1), the EU Network of Excellence NuGO (NoE 503630), the EU Carcinogenomics (PL037712), the NERC Environmental Bioinformatics Centre and the EMBL-EBI for funding the ISA infrastructure, the MIBBI and the ISA-Tab activities.

References

1. Mons B, Ashburner M, Chichester C, van Mulligen E, Weeber M, den Dunnen J, van Ommen GJ, Musen M, Cockerill M, Hermjakob H, Mons A, Packer A, Pacheco R, Lewis S, Berkeley A, Melton W, Barris N, Wales J, Meijssen G, Moeller E, Roes PJ, Borner K and Bairoch A. Calling on a million minds for community annotation in WikiProteins. Genome Biol. 2008;9(5):R89. Epub 2008 May 28.
2. Fink JL and Bourne PE. Reinventing Scholarly Communication for the Electronic Age. 2007: http://www.ctwatch.org/quarterly/print.php?p=83
3. The Arabidopsis Information Resource: http://www.arabidopsis.org/
4. Mouse Phenome Database: http://www.jax.org/phenome
5. Editorial: Standard operating procedures. 2006. Nat Biotechnol. 2006. Nov;24(11):1299
6. Editorial: Democratizing proteomics data. Nat Biotechnol. 2007. Mar;25(3):262.
7. Editorial: Standardizing data. Nat Cell Biol. Oct;10(10):1123-4 (2008).
8. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J and Vingron M. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet. 2001 Dec;29(4):365-71.
9. Taylor CF, Paton NW, Lilley KS, Binz PA, Julian RK Jr, Jones AR, Zhu W, Apweiler R, Aebersold R, Deutsch EW, Dunn MJ, Heck AJ, Leitner A, Macht M, Mann M, Martens L, Neubert TA, Patterson SD, Ping P, Seymour SL, Souda P, Tsugita A, Vandekerckhove J, Vondriska TM, Whitelegge JP, Wilkins MR, Xenarios I, Yates JR III and Hermjakob H. The minimum information

about a proteomics experiment (MIAPE). Nat Biotechnol. 2007 Aug;25(8):887-93

10. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, Ashburner M, Axelrod N, Baldauf S, Ballard S, Boore J, Cochrane G, Cole J, Dawyndt P, De Vos P, DePamphilis C, Edwards R, Faruque N, Feldman R, Gilbert J, Gilna P, Glöckner FO, Goldstein P, Guralnick R, Haft D, Hancock D, Hermjakob H, Hertz-Fowler C, Hugenholtz P, Joint I, Kagan L, Kane M, Kennedy J, Kowalchuk G, Kottmann R, Kolker E, Kravitz S, Kyrpides N, Leebens-Mack J, Lewis SE, Li K, Lister AL, Lord P, Maltsev N, Markowitz V, Martiny J, Methe B, Mizrachi I, Moxon R, Nelson K, Parkhill J, Proctor L, White O, Sansone SA, Spiers A, Stevens R, Swift P, Taylor C, Tateno Y, Tett A, Turner S, Ussery D, Vaughan B, Ward N, Whetzel T, San Gil I, Wilson G and Wipat A. The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol. 2008 May;26(5):541-7.
11. Taylor CF, Field D, Sansone SA, Aerts J, Apweiler R, Ashburner M, Ball CA, Binz PA, Bogue M, Booth T, Brazma A, Brinkman RR, Michael Clark A, Deutsch EW, Fiehn O, Fostel J, Ghazal P, Gibson F, Gray T, Grimes G, Hancock JM, Hardy NW, Hermjakob H, Julian RK Jr, Kane M, Kettner C, Kinsinger C, Kolker E, Kuiper M, Novère NL, Leebens-Mack J, Lewis SE, Lord P, Mallon AM, Marthandan N, Masuya H, McNally R, Mehrle A, Morrison N, Orchard S, Quackenbush J, Reecy JM, Robertson DG, Rocca-Serra P, Rodriguez H, Rosenfelder H, Santoyo-Lopez J, Scheuermann RH, Schober D, Smith B, Snape J, Stoeckert CJ Jr, Tipton K, Sterk P, Untergasser A, Vandesompele J and Wiemann S. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat Biotechnol. 2008 Aug;26(8):889-96.
12. The Open Biomedical Ontologies collection: http://www.obofoundry.org/
13. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ; OBI Consortium, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL and Lewis S. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007 Nov;25(11):1251-5.
14. The Proteomics Standards Initiatives' semantic validator: http://www.psidev.info/index.php?q=node/304
15. ISA Infrastructure: http://isatab.sourceforge.net
16. Sansone SA, Rocca-Serra P, Brandizi M *et al*. The First RSBI (ISA-TAB) Workshop: "Can a Simple Format Work for Complex Studies?". OMICS. Jun;12(2):143-9 (2008).
17. ISA Tools: http://isatab.sourceforge.net/tools.html
18. Ontology Lookup Service: http://www.ebi.ac.uk/ontology-lookup
19. BioPortal: http://bioportal.bioontology.org
20. The BioInvestigation Index prototype instance at EBI: http://www.ebi.ac.uk/bioinvindex.
21. Phenote: http://www.phenote.org
22. Pride Harvest: http://www.ebi.ac.uk/pride/proteomeharvest

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

Extending the Foundational Model of Anatomy with Automatically Acquired Spatial Relations

Manuel Möller[1], Christian Folz[2], Michael Sintek[1], Sascha Seifert[3], Pinar Wennerberg[4]
[1]DFKI GmbH, Kaiserslautern, Germany
[2]University of Applied Sciences Kaiserslautern, Germany
[3]Siemens AG, Corporate Technology, Erlangen, Germany
[4]Siemens AG, Corporate Technology, Munich, Germany

Abstract

Formal ontologies have made significant impact in bioscience over the last ten years. Among them, the Foundational Model of Anatomy Ontology (FMA) is the most comprehensive model for the spatio-structural representation of human anatomy. In the research project MEDICO we use the FMA as our main source of background knowledge about human anatomy. Our ultimate goals are to use spatial knowledge from the FMA (1) to improve automatic parsing algorithms for 3D volume data sets generated by Computed Tomography and Magnetic Resonance Imaging and (2) to generate semantic annotations using the concepts from the FMA to allow semantic search on medical image repositories. We argue that in this context more spatial relation instances are needed than those currently available in the FMA. In this publication we present a technique for the automatic inductive acquisition of spatial relation instances by generalizing from expert-annotated volume datasets.

Introduction

Semantic medical image search as approached by Advances in medical imaging have greatly increased the amount of images produced in clinical facilities. At the same time, modern hospital information systems have also become more complex. Today's clinical facilities typically contain *hospital information systems* (HIS) for storing patient billing and accounting information, *radiological information systems* (RIS) for storing radiological reports, and *picture archiving and control systems* (PACS) for archiving medical images.

It has become challenging for clinicians to query and retrieve relevant previous patient data due to the volume of information, the complexity and heterogeneous nature of today's information systems. In particular, former patient images are useful for analyzing images of a current examination since they help in understanding any progression of pathologies or development of recent abnormalities, e.g., in the context of lymphoma.

The research project MEDICO[1] aims to fuse techniques for automatic image segmentation and text annotation with semantic web techniques. The goal is to allow cross-lingual and modality-independent search and retrieval across medical images, clinical findings and reports. This requires processes for automatic annotation of images and documents with concepts from formal ontologies to allow retrieval to be performed on an abstract level. Thus, searching becomes independent of the concrete data representation and can leverage on the information modeled in formal ontologies, e.g., for query expansion as described in a recent ESWC publication[2].

Medical imaging equipment nowadays generates huge amounts of data either as 2D images (e.g., X-ray) or 3D volumes which are stacks of 2D image slices generated by techniques such as Computed Tomography (CT) or Magnetic Resonance Imaging (MRI). Many approaches for medical image parsing already incorporate information on the spatial distribution of anatomical entities in the human body during the automatic detection. But with only a few exceptions this background knowledge is a fixed part of the source code of the algorithms, e.g., by using statistical distribution functions. Thus, the whole application has to be recompiled in order to alter or extend this knowledge.

On the other hand knowledge about human anatomy has already been modeled in formal ontologies which represent computable artifacts. In MEDICO we use the Foundational Model of Anatomy ontology (FMA)[3] as the main source for knowledge about human anatomy. It includes a well-founded formalism for expressing qualitative spatial relations. In this context we evaluated the existing spatial relation instances in the FMA. Throughout this document, we consider a "spatial relation" to be the modeling of the relation. In contrast, we use the term "spatial relation instance" to refer to the relation between two anatomical concepts. Thus, spatial relation instances denote relations on the class level. By looking more closely on the FMA we found that the overall number of spatial relation instances as well as the coverage of different body regions and biological systems is very limited. We present methods to add and evaluate missing spatial relation

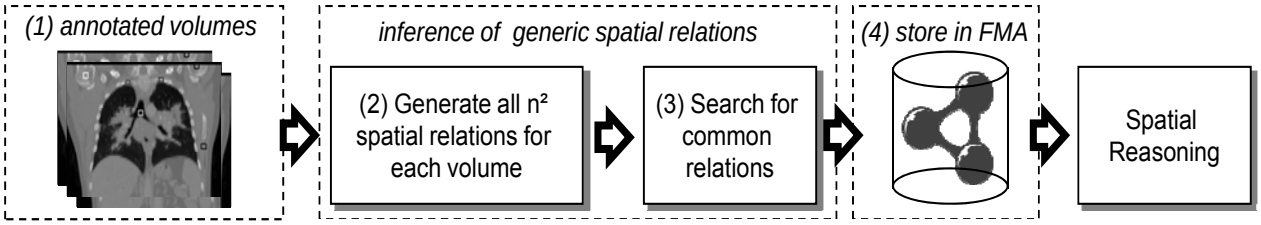


Figure 1. Workflow for Inductive Learning of Spatial Relation Instances

instances to the FMA to generate a critical mass of spatial knowledge sufficient to support automatic image parsing algorithms.

Related Work

MEDICO is based on established Semantic Web standards such as OWL[4] and RDF[5]. The MEDICO Ontology Hierarchy[6] was designed with the rationale to reuse or adapt existing ontologies instead of creating them from scratch. We build on an existing OWL translation of the FMA[7] as the primary source of anatomical knowledge.

The recent research publications in the area of medical image parsing indicate a strong tendency to store spatial relation instances as statistical distributions functions. Here, Gaussian or Bayesian models are often used.[8] A recent study presented a fast and robust approach for full-body organ segmentation.[9] This approach detects organs by taking into account nearby anatomical landmarks to improve precision.

Hybrid approaches for automatic ontology-based image segmentation and semantic annotation have been presented before. Another system for semantic annotation of brain MRI images was presented.[10]

Recently, Hudelot et al. have published an ontology for the representation of fuzzy spatial relations.[11] This work could be used as a future extension of our approach as soon as 3D volumes of organs are available instead of the 3D points which we currently use.

To our knowledge an approach for the automatic acquisition of spatial relations instances at large numbers is still missing.

Approach and Applied Methods

We started with a review of the spatial relation instances available in the current FMA (version 2.0). This revealed that their number is very limited. We counted 1153 instances of the `attributed continuous with` relation which contain directional information between pairs of anatomical entities. Natasha Noy (who provided the translation of the FMA to OWL on which we base our work) pointed out in an email from 2008-11-10 that there is more spatial relation information available via properties such as `contains` and `tributary of`, but –as a matter of fact –those lack directional information.

We also applied an automatic approach which checked all available spatial relation instances for consistency. With "consistency" we mean, for example, that if concept A is on the left side of concept B, we would expect the latter (concept B) to be on the right side of concept A. However, we found that this was not always true. We checked this together with the FMA authors who confirmed and corrected these inconsistencies.

The annotated volume data sets available for our tests contained points in 3D annotated with keywords. The annotated volume data sets available to us contained points without any spatial extension representing highest/lowest points of anatomical entities such as at vertebrae of the spine. To represent them we extended the FMA with classes representing these points. In total we added 253 classes. We also added `subClassOf` and `regional_part_of` assertions to integrate them with the existing FMA structure. Fig. 1 gives a systematic overview of the workflow applied to acquire new spatial relations instances.

(1) Basically, the workflow allows arbitrary input formats. The only requirement is that they contain points in 3D marked with landmarks which are found at this position. From our partners in the MEDICO consortium we received landmark annotations in XML and plain text formats. Thus, the first step was to implement converters which map the proprietary input formats to a common representation of points in 3D and links to concepts in the FMA.

(2) To generate the spatial relations among these landmarks in the format of the FMA we had to perform two steps: (a) Calculate the difference vectors between all pairs of landmarks and (b) map them to directions as they are modeled in the FMA (e.g., `Left`, `Right`, `Superior`). All landmark coordinates were given in millimeters. Consequently, we did not have to account for different slice spacings of the image volumes. Note that in this step the distance information is discarded. Although useful for some applications, this information varies

from patient to patient. For instance, it depends on the patient's individual size and if the stomach is full or empty. Since our aim is to extract features which are *common for all humans*, we reduced the difference vectors to their direction with the assumption that at least the direction of the vectors should generalize well for all humans. The evaluation results in give evidence for the correctness of this assumption. After this step the spatial relation instances were available in tuples of the form `[concept A] [direction] [concept B]`.

(3) To obtain a set of generic spatial relation instances – our *model* – we compared the data of different volume data sets from different patients. Next, we systematically eliminated contradictory tuples. The quality of the resulting model was evaluated using cross validation.

(4) The model was serialized in OWL format, added to our local copy of the FMA and is subsequently available for spatial reasoning. In total we were able to generate about 13,500 spatial relation instances.

Evaluation

For our evaluation we had 30 different volume data sets available with 145 different landmarks annotated on average. These volumes belong to 29 different patients. At first glance this number might seem low; but in fact this corpus required the localization of more than 4000 landmarks in 30 volume data sets which consist of several thousand single images.

Our learning algorithm had two parameters which had influence on the generated model: `minFrequency` determines in how many of the training examples the spatial relation has to occur before it is considered as stable and thus part of the inferred generic model. For example: For fix `X` and `Y` the spatial relation instance `[X] [Left] [Y]` only occurs in less than 20 percent of the training examples. Here we take low numbers for low evidence for this relation and thus do not take it into the model. `minConfidence` determines which fraction of each pair of source and destination concepts has to share the same direction before the pair and its predominant direction are added to the inferred model. This rule is applied when the training corpus contains contradictory spatial relation instances. To give an example: For the concepts `[X]` and `[Y]` there are two distinct classes of tuples in the training corpus: `[X] [Left] [Y]` and `[X] [Inferior] [Y]`. In general, we do not add any of the tuples to the model at all since they are inconsistent. But if more than `minConfidence` of all tuples belong to the same class, we still add a representative of this class, taking their distribution as support for their universality.

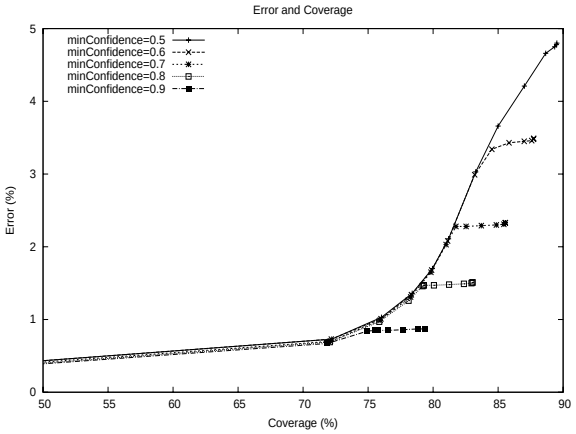


Figure 2. Relation between coverage/error/minConfidence

We performed a 4-fold cross validation on our corpus to evaluate the influence of the parameters discussed above. The results are visualized in Fig. 2. The curve shows the typical trade-off between recall and precision: the more spatial relations of the training models we included (*coverage*) the higher the error gets. Our evaluation also shows that the `minConfidence` parameter had a big influence both on coverage and error rate in the resulting model. The impact of the parameter `minFrequency` was comparable. Based on the dataset available to us we were able to generate approximately 13,500 spatial relation instances. This covers 85% of all spatial tuples appearing in the training volumes with an error rate in the test corpus of only 0.87%.

Conclusion and Future Work

The research project MEDICO aims to improve pixel-based medical image and volume data set segmentation algorithms by fusing existing techniques with formal knowledge about anatomy from ontologies. Based on the limited number of spatial relation instances in the FMA we have argued that there is a need for techniques which acquire additional knowledge about the spatial distribution of anatomical entities in human bodies.

We presented our automatic inductive approach which infers a set of spatial relation instances from manually annotated volume data sets. Our evaluation results show that this method is able to provide reasonable numbers of additional spatial relation instances with error rates below 1 percent.

For the future we plan to enable our learning approach to make use of implicit transitive spatial relation instances. This would allow to formalize the learned model using far less spatial relation instances.

We also plan to extend our corpus of annotated volume data sets. With a spatial model which is justified by a larger base of expert annotations we plan to investigate its suitability for high-level reasoning about potential diseases. Given a stable model of spatial relations already exists it could be used to detect differences to the spatial relation instances of a particular patient.

These differences could then be used to automatically produce hints about enlargements of certain anatomical entities which could be pathological.

Acknowledgements

This research has been supported in part by the research program THESEUS in the MEDICO project which is funded by the German Federal Ministry of Economics and Technology under the grant number 01MQ07016. The responsibility for this publication lies with the authors.

References

1. Möller M, Sintek M, Buitelaar P, Mukherjee S, Zhou XS and Freund J. “Medical Image Understanding through the Integration of Cross-Modal Object Recognition with Formal Domain Knowledge”, Proc. of HEALTHINF, 2008, 1, 134-141.
2. Möller M, Regel S and Sintek M. “RadSem: Semantic Annotation and Retrieval for Medical Images”, Proc. of The 6th Annual European Semantic Web Conference (ESWC2009), 2009.
3. Rosse C and Mejino JLV. “The Foundational Model of Anatomy Ontology” in “Anatomy Ontologies for Bioinformatics: Principles and Practice”, Springer, 2007, 6, 59–117.
4. McGuinness DL and van Harmelen F. “OWL Web Ontology Language Overview”, World Wide Web Consortium, 2004.
5. Beckett D. *(ed.)*. “RDF/XML Syntax Specification (Revised)”.
6. Möller M and Sintek M. “A Generic Framework for Semantic Medical Image Retrieval”, in Proc. of the Knowledge Acquisition from Multimedia Content (KAMC) Workshop, 2nd International Conference on Semantics and Digital Media Technologies (SAMT), 2007.
7. Noy NF and Rubin DL. “Translating the Foundational Model of Anatomy into OWL”, Stanford Medical Informatics Technical Report, 2007.
8. Pham TV and Smeulders AW. “Learning spatial relations in object recognition” in Pattern Recognition Letters, 2006, 27, 1673–1684.
9. Seifert S, Barbu A, Zhou SK, Liu D, Feulner J, Huber M, Suehling M, Cavallaro A and Comaniciu D. “Hierarchical Parsing and Semantic Navigation of Full Body CT Data”, in Proc. of SPIE, 2009.
10. Mechouche A, Golbreich C, Gibaud B, Marchiori M, Pan J and de Sainte Marie C. “Towards an Hybrid System Using an Ontology Enriched by Rules for the Semantic Annotation of Brain MRI Images”, in Lecture Notes in Computer Science, 2007, 4524, 219–228.
11. Hudelot C, Atif J and Bloch I. “Fuzzy spatial relation ontology for image interpretation” in Fuzzy Sets Syst., Elsevier North-Holland, Inc., 2008, 159, 1929–1951.

An Exercise on Developing an Ontology-Epistemology about Schizophrenia and Neuroanatomy

Rodolpho Freire[1], Danilo Nunes[1], Marcus V.T. Santos[2], Paulo E. Santos[1]
[1]FEI - São Paulo, Brazil; [2]Ryerson University, CA, USA

Abstract

This paper describes preliminary ideas on formalizing some concepts of neuroanatomy in ontological and epistemological terms. We envisage the application of this ontology to the assimilation of facts about medical knowledge about neuroimages deriving from schizophrenic patients.

Introduction

This paper is part of a major effort towards the formalization of the knowledge contained in neuroimages of patients with schizophrenia. Our long term goal is to build an ontology that is a formal basis for the expectations generated from statistical data analysis.

There are a number of biomedical ontologies; perhaps central to this area are the Foundational Model of Anatomy (FMA)[6] and the Open Biomedical Ontologies (OBO)[9], amongst others as summarized in by Friedman, Chen and Fuller, *et al*[3]. FMA is a knowledge source of classes and relations about observable characteristics of the human body structure; thus, FMA is mainly concerned with representing anatomical information. In contrast, the OBO Foundry project is a collaborative development which includes a large amount of biological information. Some attempts have also been made to build ontologies of neuroanatomical structures.[4,5]

The goal of the present paper is to relate a spatial ontology about the ventricular brain system (VBS) with findings about changes in this structure that are picked out in neuroimages from schizophrenic patients.

The structure of the VBS can evidently be represented within an ontology, however changes in neuroimages refer to knowledge about a domain, and not to the domain itself; findings about schizophrenia falls within the epistemology umbrella. A complete solution of combining ontologies with epistemology is still an open issue. However, we make explicit which are the classes related to the domain of neuroanatomy and which are related to the knowledge about the domain (the epistemological classes). In the present paper we propose a region-based ontology using the Basic Inclusion Theory (BIT)[1], due to its clear definitions of spatial regions through part-whole, taxonomic and topological relations, with the explicit use of logical relations. Another characteristic of BIT is that its underlying language is the first-order logic, which allows the inclusion of axioms about complementary theories into a single formalism. Figure 1 presents BIT base relations.

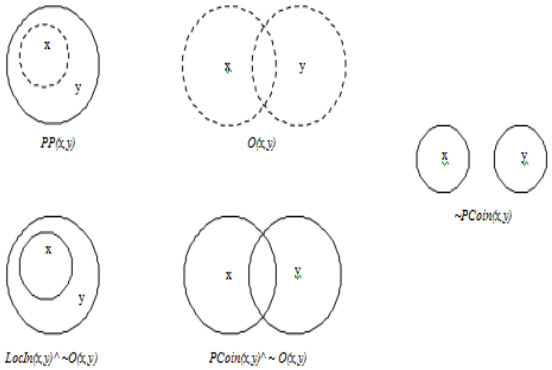


Figure 1. Basic Inclusion Theory relations.

Ventricular Brain System (VBS)

The ventricular brain system is a cavity disposed within the brain, which is composed of the third, fourth and lateral ventricles. The lateral ventricles are subdivided as body, frontal horn, occipital horn and temporal horn. The communication between the lateral ventricles and the third ventricle is done by the Monro foramina. The third ventricle is sub-divided into anterior commissure, optical recess, supraparienal recess and infundibulum and communicates with lateral ventricles by Monro foramina. The third ventricle also communicates with the fourth ventricle by the cerebral aqueduct and in the centre of the third ventricle is located the interthalamic connection. The fourth ventricle is composed by the lateral recess and the Luschka foramina linking up with the third ventricle through the brain aqueduct.

A Spatial Bio-Ontology for VBS

To represent the ventricular brain system and the medical knowledge about schizophrenia, we define Fiat Boundaries[11,10] and a notion of continuity[8,11] using BIT relations. The next section presents some ideas about how it is accomplished.

Fiat Boundaries and Continuity

Fiat boundaries are used for representing abstract limits, i.e., those limits that are commonly accepted, but which do not have a concrete existence[11,10]. In the biomedical area fiat boundaries can be used to delimit anatomical regions[2], such as the limits between ventricular regions. We define the relation LFiat, read as "x is a fiat boundary in y", and is axiomatised as follows:

$$LFiat(x, y) \quad (1)$$

$$LFiat(z, x) \rightarrow PCoin(z, x) \quad (2)$$

$$LFiat(x, y) \rightarrow \neg \exists z\, LFiat(z, x) \quad (3)$$

$$LFiat(x, y) \wedge LFiat(x, z) \rightarrow \neg Desc(y, z) \quad (4)$$

$$LFiat(x, y) \wedge LFiat(y, z) \rightarrow PCoin(x, y) \quad (5)$$

A discontinuity can be defined as a disjunction among two distinct spatial regions which became disconnected. Santos e Cabalar[8] proposed a theory based to represent discontinuity based in Varzi[11]. We use this notion to represent (for instance) the third ventricle, which has a material discontinuity called the interthalamic connection. We represent a discontinuity using the relation Disc(x,y) ("x is a discontinuity in y") and define the following axioms to constrain its meaning:

$$Disc(x, y) \quad (6)$$

$$Disc(x, y) \rightarrow LocIn(x, y) \quad (7)$$

$$Disc(x, y) \rightarrow \neg Disc(y, z) \quad (8)$$

Using Desc/2 we can define the notion of "continuous part": PCont(x, y) meaning that x is a continuous part of y, as shown in formula 9.

$$PCont(x, y) \equiv [PP(x, y) \vee P(x, y)] \wedge \forall z\, \neg Disc(z, x) \quad (9)$$

Then, we define a segment x of an object y (*Segm(x,y)*) as the "maximal continuous part" of y according to formula 10.

$$Segm(x, y) \equiv PCont(x, y) \wedge \neg \exists z[PP(y, z) \wedge PCont(z, y)] \quad (10)$$

Representing the VBS

There are 21 Fiat boundaries limiting all ventricular anatomical elements. The formulas 11 to 14 represent the fiat boundaries (represented by Z) that delimit the right and left lateral ventricles, third ventricle and fourth ventricle.

$$LFiat(Z_1;Left\_Lateral\_Ventricle) \quad (11)$$

$$LFiat(Z_2;Right\_Lateral\_Ventricle) \quad (12)$$

$$LFiat(Z_3;Third\_Ventricle) \quad (13)$$

$$LFiat(Z_4;Fourth\_Ventricle) \quad (14)$$

Given the definitions of Fiat boundaries and continuity, we have conditions to represent each ventricle individually without ambiguities. The foramina area is defined in similar terms. The formulas 15 define the right lateral ventricle, the volume of ventricle is given by variable φ. In the similar way the formulas 16 to 18 define the third left lateral ventricle, third ventricle and fourth ventricle, respectively:

$$Inst(x,Right\_Lateral\_Vent) \leftarrow (Vol(x) = \varphi) \wedge [\varphi > Vol\_TV] \wedge [\varphi > Vol\_FV] \wedge LFiat(Z_2,x) \wedge Segm(x,Ventricular\_Brain\_System) \quad (15)$$

$$Inst(x,Left\_Lateral\_Ventricle) \leftarrow (Vol(x) = \varphi) \wedge [\varphi > Vol\_TV] \wedge [\varphi > Vol\_FV] \wedge LFiat(Z_1, x) \wedge Segm(x, Ventricular\_Brain\_System\ ) \quad (16)$$

$$Inst(x,Third\_Ventricle) \leftarrow (Vol(x)=\varphi) \wedge [(\varphi<Vol\_LLV) \wedge (\varphi<Vol\_RLV)] \wedge (\varphi>Vol\_FV) \wedge LFiat(Z_3,x) \wedge Segm(x,Ventricular\_Brain\_System) \quad (17)$$

$$Inst(x,Fourth\_Ventricle) \leftarrow (Vol(x) = \varphi) \wedge [(\varphi<Vol\_LLV) \wedge (\varphi<Vol\_RLV)] \wedge (\varphi<Vol\_TV) \wedge LFiat(Z_4,x) \wedge Segm(x, Ventricular\_Brain\_System) \quad (18)$$

We include in Protégé all axioms that represent the Fiat boundaries and which anatomical structures they are related to. This definition is expressive enough to answer questions such as: "given one region x, that belongs to y, which is this region?", or "which ventricular region is the foramina x connected with the ventricles?".

Epistemological Classes

In order to define common characteristics among distinct groups, the medical specialist relies on the relative literature (using information from meta-analysis), image or statistical analysis. The information available in these sources is not part of the domain (so it cannot be captured by an ontology) but it is knowledge about it.

The knowledge about things are not the things itself, therefore, including it in the ontology would lead to Kantian confusion. In this work we avoid this confusion by assuming "epistemological classes", which are related to the ontological classes by a modified Is_a relation (Is_a2). Given an epistemological class E, an ontological class O and a binary primitive relation κ(x, y) (representing that x is the knowledge about a domain y), we define Is_a2 in BIT in the following way:

$$Is\_a2(E,O) \equiv \forall x(Inst(x,E) \rightarrow \neg Inst(x,O) \wedge \kappa(x,O)) \quad (19)$$

Informally, E is an epistemological class within the ontology O iff every instance of E is not an instance of O but is knowledge about O. In Figure 2 we can see a graphical schema that shows epistemological classes about the right lateral ventricle. The epistemological classes are described by the formulas 20 to 27.

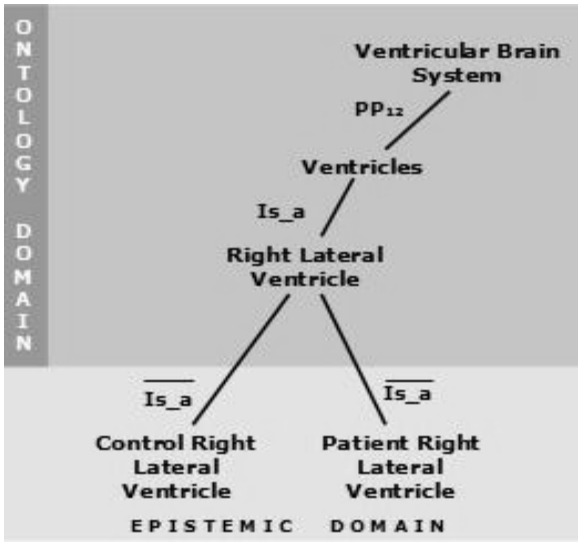


Figure 2.
Differences about Ontological and Epistemological domains.

Is_a2(CONTROL_RIGHT_LV,RIGHT_LV) (20)
Is_a2(PATIENT_RIGHT_LV,RIGHT_LV) (21)
Is_a2(CONTROL_LEFT_LV,LEFT_LV) (22)
Is_a2(PATIENT_LEFT_LV,LEFT_LV) (23)
Is_a2(CONTROL_TV,THIRD_VENTRICLE) (24)
Is_a2(PATIENT_TV,THIRD_VENTRICLE) (25)
Is_a2(CONTROL_FV,FOURTH_VENTRICLE) (26)
Is_a2(PATIENT_FV,FOURTH_VENTRICLE) (27)

Therefore, we can include both ontological and epistemological individuals in the same formalism. In this work, an epistemological individual is a piece of knowledge about anatomical changes in the VBM (related to schizophrenia) that comes from the medical literature (meta analysis for instance) or from image data analysis procedures. It is now possible to execute queries about, for instance, the composition of the ventricular brain system, or about specialist knowledge about the domain.

An example of an ontological query is: "Which structures compose the ventricular brain system?" This query in Protégé (using Manchester syntax) becomes "PP only Ventricular_Brain_System" and results in all classes that compose the ventricular brain system. Figure 3 shows us a part of some of these results.

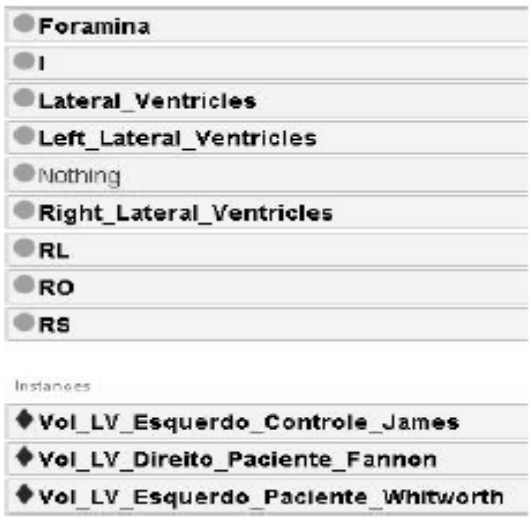


Figure 3. Ontological Query Result

Epistemological reasoning is possible in a similar way: the query "the volume 6.52 of the right or left lateral ventricles is classified as patient or control groups?". In Protégé this query becomes "Lateral_Ventricles and Vol value 6.52", and produces the result: "Vol_Right_LV_Control_Barr", which means that the classification of an individual whose lateral ventricle (LV) has a volume of 6.52 is "control" according to Barr [7].

Conclusion and Future Work

This paper briefly described a formalization for ontological and epistemological classes about the ventricular brain system defined using BIT, and realized computationally in Protégé. This allows us to include and consult the domain entities as well as the knowledge about the domain. Future work will consider the formalization of new evidences about schizophrenia to be included in this framework.

References

1. Bittner T and Donnelly M. A formal theory of qualitative size and distance relations between regions. In Proceedings of the 21st International Workshop on Qualitative Reasoning, 2007.
2. Fielding JM and Marwed D. The image as spatial region: Location and adjacency within the radiological image. In Formal Ontology in Information Systems, pages 444–448, 2006.
3. Friedman C, Chen H, Fuller SS and Hersh W (eds.). Medical Informatics, volume 8 of Integrated Series in Information Systems, chapter Biomedical Ontologies, pages 211–236. Springer US, 2006.

4. Bota M and Swanson LW. Bams neuroanatomical ontology: Design and implementation. Front. Neuroinform, 2(2), 2008.
5. Martin RF, Mejino JLV, Bowden DM, Brinkley JF and Rosse C. Foundational model of neuroanatomy: Implications for the human brain project. In Proc. of the American Medical Informatics Association Fall Symposium, pages 438–442, 2001.
6. Rosse C and Mejino JLV. A reference ontology for biomedical informatics: The foundational model of anatomy. Journal of Biomedical Informatics, 36(6):478–500, 2003.
7. Shenton ME, Dickey CC, Frumin M and McCarley RW. A review of mri findings in schizophrenia. Elsevier, 49(1-52):141–144, 2001.
8. Santos P and Cabalar P. The space within fisherman's folly: Playing with a puzzle in mereotopology. Spatial Cognition and Computation, 2008.
9. Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL and Rosse C. Relations in biomedical ontologies. Genome Biol, 6(5), 2005.
10. Smith B and Varzi AC. Fiat and bona fide boundaries. Philosophy and Phenomenological Research, 6, 2000, 401–20.
11. Varzi AC. Boundaries, continuity and contact. Time, Space and Movement: Meaning and Knowledge in the Sensible World (Proceedings of the 5th International Workshop), 1995.

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

Towards an Ontology of Biomedical Educational Objectives

Martin Boeker, Holger Stenzhorn, Felix Balzer, Stefan Schulz
Freiburg University Medical Center, Freiburg i. Br., Germany

Abstract

The domain of medical education combines educational objectives and cognitive functions with the subject-matter of organism structures, functions and malfunctions as well as diagnostic and interventional techniques.

A re-structuring of learning objectives guided by principles of ontology seems promising as this gives the means to compare, classify, and validate learning objectives using formal methods.

Although large parts of the subject-matter of medical education are already covered by existing biomedical vocabularies, there still exist several challenges for designing an ontology of educational objectives. Emphasis is given to the representation of plans and cognitive entities on the one hand, and on prototypical "blueprint" entities on the other.

Introduction

A high-quality education of undergraduate and graduate students as future health professionals is a cornerstone for the sustained delivery of high quality and effective health services. Formalization and standardization of educational content can be an important means to reduce the complexity and inconsistency in this highly dynamic field with its rapidly emerging and evolving contents.

Educational objectives are the core of every proficient teaching and learning assessment process[1]. For over 50 years educational objectives have served as "explicit formulations of the ways in which students are expected to be changed by the educative process"[2]. Given this important function, in medical education, a large number of educational objectives have been collected in catalogues, either on a central basis for a group of medical schools[3] or on a local basis for single institutions.

Despite the laborious effort to compile large bodies of educational objectives, the targeted audience – teachers, learners and curriculum developers – has had limited benefit. It is difficult to access the tremendous amount of educational objectives in a certain subject-matter in any practical way: Subject-matters in a medical curriculum are as diverse as the expected knowledge and skills of the future professionals, covering the whole range of pre-clinical and clinical disciplines. Curriculum content also includes teaching-related factors such as contact hours or specific learning aids.

Up to now, learning objectives have been published in narrative form without any reference to standardized terminologies. As a consequence, their content is subject to different interpretations. Further, their arrangement in hierarchies is intuitive but informal, e.g., the Taxonomy of Educational Objectives[4]. As a result, in current practice, the sequence, consistency and coherence of the complex structure of a compilation of learning objectives within a curriculum can neither be exhaustively checked nor displayed in a principled fashion.

This is the reason why we propose to follow principles of formal ontology for the representation of learning objectives. This *Ontology of Biomedical Educational Objectives* (*OBEO*) is intended to support tools for annotation, consistency checking, and navigation within educational objective catalogues. We propose to align this ontology with an upper-level ontology, like DOLCE[5], and also link it to established biomedical terminologies and ontologies. Here, SNOMED CT[6] would be the first choice due to its high coverage of the clinical domain and being an accepted standard resource. For the basic biomedical sciences, several OBO Foundry ontologies[7] would constitute a useful completion as the most prominent ontological source in this domain.

The Function and Structure of Learning Objectives

The formulation of educational objectives plays a central role in the development of a medical curriculum by addressing the needs of the learners[8]. The importance of educational objectives becomes obvious in the scope of their different roles in the educational process[1].

Educational objectives provide

- focus for instruction,
- guidelines for learning,
- targets for formative and summative assessment,
- instructional intent to others, and
- possibilities for instruction evaluation.

Today, educational objectives are usually placed as intended learning outcomes at the final stage of an instruction process which allows assessing the

students' performance after the learning process. Defined in such a way, educational objectives represent clearly what is expected from the students *after* the training, e.g., through demonstrating knowledge, performance in psycho-motor or communication skills, or even in the complex behavior associated with certain attitudes.

Educational objectives in a medical curriculum are typically formulated as follows:

- *The physician is able to assess a patient presenting this problem* [from a list of given medical problems] *in a well-structured way, and to establish a differential diagnosis.*
- *She/he is able to propose appropriate diagnostic, therapeutic, social, preventive and other measures, and to provide urgent intervention in case of life-threatening problems.*[3]

This learning objective states *who* will do *how much* (or *how well*) of *what*. Thus a learning objective statement comprises an *agent* (usually the learner) who performs a certain *action* which indicates a defined *performance level* to proof his or her acquired *knowledge*, *skills*, or *attitude* towards some given *subject*.

Using the BioTop domain upper-level ontology together with the DOLCE upper-level ontology[5] and the OBO Relation Ontology[9], we can express a learning objective as a *goal* represented by a *biotop:immaterial-nonphysical-entity*[a] and which is part of some *biotop:plan*. Since many different things can constitute an *obeo:learning-objective* in one context but not in another one, we express this by introducing the *obeo:learning-objective-role*. The learner is an instance of *biotop:human* who is *ro:agentIn* in a *biotop:action*, as defined in the *obeo:learning-objective* and in whom it is *internally represented*. For the existence of an *obeo:learning-object*, the *biotop:action* does not need to be instantiated but it can be *only obeo:realized-by* the specified *biotop:action* (cf. Fig. 1).

Complex relations exist between the type of action to be performed by the learner and the type of subject. E.g., there are subtypes of *biotop:action* for the cognitive domain, *such as obeo:remembering*, *obeo:understanding*, *obeo:applying*, *obeo:analyzing*, *obeo:evaluating*, or *obeo:creating*. This can be represented as a hierarchy of cognitive actions[2, 4].

To broaden the view on educational objectives as intended learning outcomes, a list of possible learning outcomes is given according to[10]:

- reactions to learning,
- modification of attitudes and perceptions,
- acquisition of knowledge and skills,
- behavioral changes,
- changes in organizational practice, and
- benefits to patients.

In this context, educational objectives can be defined more generally in terms of professional conduct or competencies a health care professional should exhibit, as well as more specific tasks a physician is expected to perform in respect to the medical domain.

Reference to the Subject-Matter in the Definition of Educational Objectives

The large body of educational objectives in medicine is related to well-defined medical topics as those covered by a broad range of biomedical vocabularies.

Therefore, an educational objectives ontology will have to include or to refer to existing terminological and ontological sources in order to cover the subject-matter of the domain knowledge and skills to be demonstrated or performed. Important features for the definition of specific medical educational objectives are anatomical and biomolecular structures, etiology, epidemiology, clinic, and diagnostic features of diseases, clinical pathways, diagnostic and interventional techniques, etc.

Although most of this content is already covered by current biomedical ontologies, terminologies, and classification systems, e.g., SNOMED CT, ICD, OBO, their inclusion into the definition of learning objectives leads to new ontological challenges, as there are major differences between the standard usage of domain ontologies and their usage in the context of describing learning objectives:

- In the standard approach, ontology classes are instantiated by particular objects or processes that have a concrete spatiotemporal existence. For example, the rationale of having a class *snomed:influenza* in a clinical ontology is to describe what is *universally true* for all instances of this class and what, consequently, can be asserted for each individual influenza that instantiates this class.

[a] *biotop* identifies classes and relations from BioTop, *obeo* classes and relations from the Ontology of Educational Objectives (OBEO), *dol* classes and relations from DOLCE, *snomed* concepts from SNOMED CT and *ro* relations from the OBO Relation Ontology.

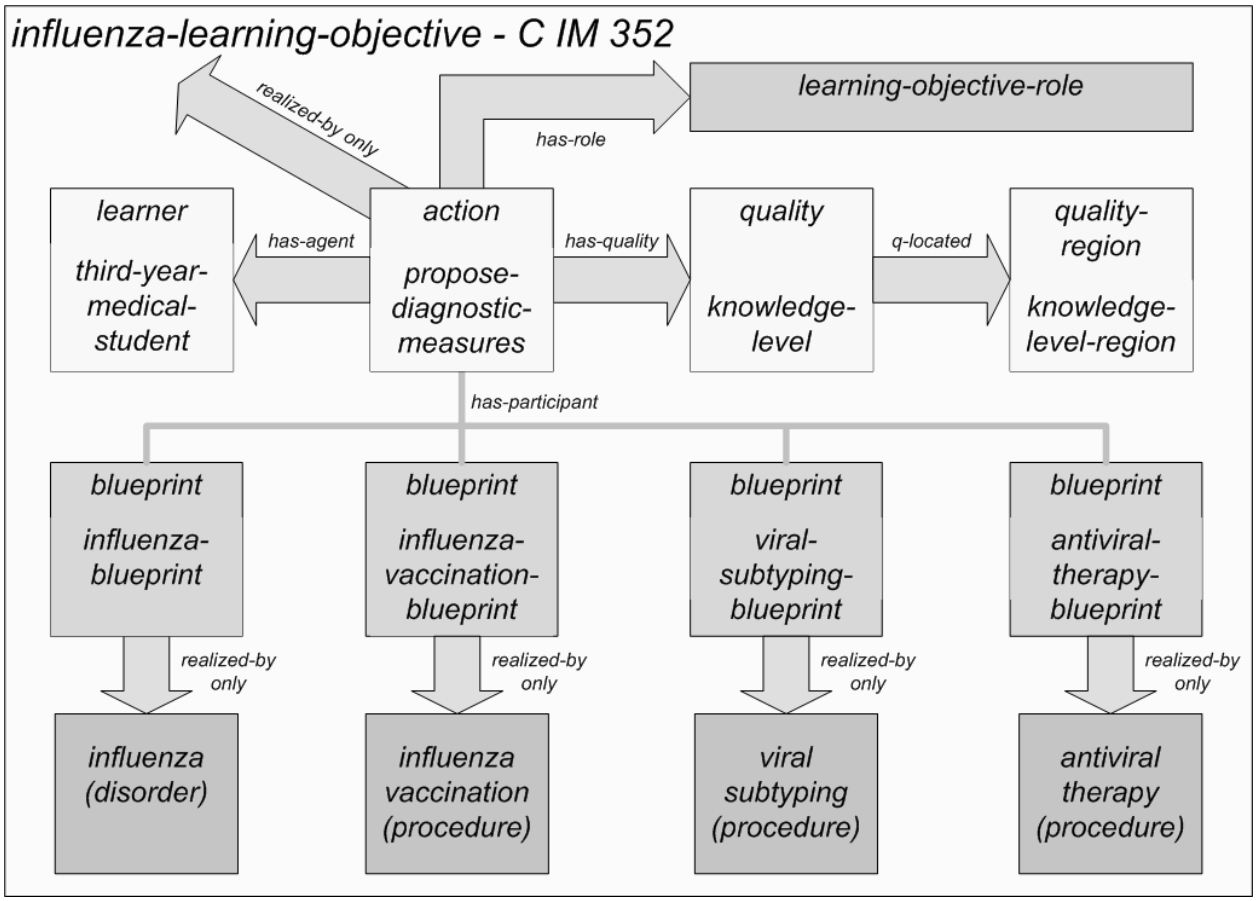


Figure 1: Displayed is a part of a complex *obeo:influenza-learning-object* which can be instantiated *only* by a *biotop:action* with several *biotop:qualities* and the *biotop:role obeo:learning-objective-role*. Biomedical concepts are linked via *oboe:blueprint* classes which are abstractions of concrete medical entities referred by standard terminologies/ ontologies. Blueprints can be tailored to the specific needs in the learning-objective, e.g., by the extension of a subset of the subclasses.

- But if *influenza* is referred to in an educational scenario, the focus is, on the contrary, to convey what is *typically true* for an influenza. So, instead of referring to concrete instances of *influenza* of concrete patients, the reference to *snomed:influenza* in an educational scenario targets a kind of "blueprint" of this disease, but not a class of really existing influenza instances.
- Such representational objects (blueprints of anatomical structures but also of disease courses, clinical protocols, etc.) can be tailored on a broad scale to the learning objectives and specific needs of the learners, thus existing on several levels of abstraction. For instance, a "heart-blueprint" for a gross anatomy course in a medical school will include the ramifications of the heart's electrical conduct system. But this is probably not the case in an introductory course on human biology at high school level. This, again, might be very different from the heart as a topic of pathological anatomy where it is likely to include the different congenital defects, as taught, e.g., in a course for pediatric nurses.

The general problem is to ascertain the ontological nature of blueprint objects and to formally relate them to classes in biomedical ontologies. In BioTop, blueprint objects are best classified under the type *biotop:immaterial-nonphysical-entity* which in turn is a subclass of DOLCE's *dol:non-physical-object* (formerly known as *dol:description*). Although not explicitly provided for neither by BioTop nor by DOLCE, we propose to relate blueprints to the related classes by the relation *obeo:realized-by* For example, we state that a *obeo:influenza-blueprint* is a *obeo:blueprint* that can only be realized by some instance of *snomed:influenza*. In OWL Manchester Syntax notation[11]:

obeo:influenza-blueprint equivalent-to
  obeo:blueprint and
    oboe:realized-by only *snomed:influenza*

Although not very relevant in an educational context, it should be noted that a blueprint object can exist

without ever being realized. This may be the case of a new therapeutic protocol, or the structure formula of a drug that has not yet been synthesized.

Classes like *obeo:influenza-blueprint* can be further specialized in terms of different ways of description, e.g., graphic, textual, etc., or different granularities as deemed adequate for each group of learners.

In any case we have to reject naïve models such as

oboe:learning-objective-I implies
includes some *snomed:influenza* (…)

This states that for each *oboe:learning-objective-I* at least one instance of *snomed:influenza* exists. But the educational objective certainly has not any specific relation to any particular disease instance of a particular person.

Conclusion

Summing up, the reference to biomedical ontologies in the context of creating an ontology for learning objectives requires new modeling patterns, because it is always the (prototypical) description of some type of domain entities which has to be represented in a biomedical ontology. In contrast to clinical ontologies, where process, procedure, or disease types are instantiated by particular patients with their particular diseases, operations, signs, symptoms and diagnostic parameters, these (particular) things are not of interest in an ontology of educational objectives. This motivates the introduction of a new type of entity, a kind of prototypical description we have termed "blueprint". Blueprint entities have not been subject to current biomedical ontologies, although some of them – above all the Foundational Model of Anatomy – show a clear (however not explicit) tendency toward this kind of representation.

As a second but equally important conclusion of this paper we emphasize the need to ontologically redesign the existing informal catalogs and taxonomies which are currently being used in medical education.

The OWL implementation of the ontology can be retrieved at http://purl.org/imbi/obeo.owl.

References

1. Gronlund NE and Brookhart SM. *Gronlund's Writing Instructional Objectives;* Pearson Education, Inc.: Upper Saddle River, New Jersey, 2009.
2. Bloom BS, Engelhart MD, Furst EJ, Hill WH and Krathwohl DR (eds.): *Taxonomy of educational objectives: handbook I: Cognitive domain.;* David McKay: New York, 1956.
3. Bürgi H, Rindlisbacher B, Bader C, Bloch R, *et al*. (eds.): *Swiss Catalogue of Learning Objectives for Undergraduate Medical Training: Under a mandate of the Joint Comission of the Swiss Medical Schools*, 2nd ed.; University of Bern: Bern, 2008.
4. Anderson LW, Krathwohl DR, Airasian PW, Cruikshank KA, *et al*. (eds.): *A Taxonomy for Learning, Teaching, and Assessing.: A Revision of Bloom's Taxonomy of Educational Objectives.;* Longman: New York, 2001.
5. Masolo C, Borgo S, Gangemi A, Guarino N and Oltramari A. *Report Title: WonderWeb Deliverable D18. Ontology Library (final)*, 2003.
6. Standardised Nomenclature of Medicine - Clinical Terms (SNOMED CT). IHTSDO - International Healthcare Terminology Standards Development Organisation. http://www.ihtsdo.de. Last accessed June 10, 2009.
7. Smith B, Ashburner M, Rosse C, Bard C, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, The OBI Consortium, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone S-A, Scheuermann RH, Shah N, Whetzel PL and Lewis S. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration, Nature Biotechnology (2007). 25, 1251–1255.
8. Kern DE, Thomas PA, Howard DM and Bass EB. *Curriculum Development for Medical Education: A Six-Step Approach;* The John Hopkins University Press: Baltimore, Maryland, 1998.
9. Beisswanger E, Stenzhorn H, Schulz S and Hahn U. BioTop: An upper domain ontology for the life sciences. A description of its current structure, contents, and interfaces to OBO ontologies; Applied Ontology (2008) 3(4), 205–212
10. Della Freeth, Hammick M, Koppel I, Reeves S and Barr H. *A Critical Review of Evaluations of Interprofessional Education*, Higher Education Academy H S a P N, Ed.: King's College London. London, 2002. http://www.health.heacademy.ac.uk/publications/occasionalpaper/occasionalpaper02.pdf. Last accessed June 10, 2009.
11. Horridge M and Patel-Schneider PF. OWL 2 Web Ontology Language Manchester Syntax http://www.w3.org/2007/OWL/wiki/Manchester Syntax. Last accessed June 10, 2009.

An Ontology-Based Framework for Clinical Research Databases

Megan Kong[1], Carl Dahlke[2], Diane Xiang[1], David Karp[4], Richard H. Scheuermann[1,3]
[1]Department of Pathology, [3]Division of Biomedical Informatics, [4]Division of Rheumatology, U.T. Southwestern Medical Center, Dallas, TX, USA; [2]Health Information Systems, Northrop Grumman, Inc., Rockville, MD, USA

Abstract

The Ontology-Based eXtensible data model (OBX) was developed to serve as a framework for the development of a clinical research database in the Immunology Database and Analysis Portal (ImmPort) system. OBX was designed around the logical structure provided by the Basic Formal Ontology (BFO) and the Ontology for Biomedical Investigations (OBI). By using the logical structure provided by these two well-formulated ontologies, we have found that a relatively simple, extensible data model could be developed to represent the relatively complex domain of clinical research. In addition, the common framework provided by the BFO should make it straightforward to utilize OBX database data dictionaries based on reference and application ontologies from the OBO Foundry.

Overview

The U.S. National Institutes of Health are interested in maximizing the return on the public investment in biomedical research. This had led many institutes to develop policies that encourage sharing of data generated from research supported by this public funding. In this regard, the National Institute of Allergy and Infectious Disease (NIAID) has supported a number of bioinformatics initiatives to provide the infrastructure to capture and manage research data for re-use and re-analysis. The Bioinformatics Integration Support Contract (BISC) was awarded to develop a long-term sustainable archive of data generated by the ~1500 investigators supported by the Division of Allergy, Immunology and Transplantation (DAIT). DAIT investigators conduct a wide range of research program types, including basic scientific research of immune system function, translational research to determine the underlying mechanisms of immune system disease and response to infection, and clinical trials to evaluate the safety, toxicity, efficacy and mechanisms of immune disease therapies and vaccination strategies. We have developed the Immunology Database and Analysis Portal (ImmPort) to support not only the archiving of these valuable data sets, but also to support their integration with the biological knowledge contained in other public data repositories (e.g. GenBank, UniProt, the Immune Epitope Database, the Protein Data Bank, etc.) and their analysis using state-of-the-art data mining analytical tools (www.immport.org).

One of the biggest challenges in ImmPort design is how best to manage the data derived from the wide range of different experiment methodologies being used by DAIT-funded investigators, which includes everything from gene expression and SNP genotyping microarrays up through clinical trials, and methodologies that are somewhat unique to the immunology research domain (e.g. flow cytometry and ELISPOT). And so we have adopted a general strategy for database development in which our database structure is designed around the general features of any biomedical investigation, rather than based on experimental details that might be methodology specific.

In addition to this design constraint, ImmPort would also like to ensure that our data and analytical infrastructure is maximally interoperable with other external databases and bioinformatics resources. Thus ImmPort has been an active participant and early adopter of many data standards development initiatives, including the development of minimum data standards like MIFlowCyt[1] through the MIBBI consortium[2] and ontology standards like the Ontology for Biomedical Investigations (OBI) through the Open Biomedical Ontology (OBO) Foundry consortium[3].

Through this work, we have considered how minimum data standards and ontology structures might be utilized to help inform the design of databases. However, it is also important to be clear about the distinction between ontologies and data/information models. Well-formulated ontologies are designed to describe classes of entities in reality and how these classes invariably relate to each other. The structure of ontologies should not be context dependent. In contrast, data models are focused on supporting instance level data in which specific representatives of entity classes are described together with their characteristics that distinguish individuals from each other within the class. Thus, data models need to be able to capture and integrate instance-level characteristics and context dependencies.

In the study reported here, we have attempted to investigate whether it would be possible to integrate these two components of knowledge representation in such a way as to leverage the class-level structural characteristics provided by a set of well-formulated reference ontologies as an underlying common database framework that could then be extended in a consistent fashion to incorporate the instance-specific details. We have specifically applied this strategy to the representation of clinical research data, including the study design components found in clinical protocols, clinical assessment results captured in case report forms and laboratory results obtained from the evaluation of derived human specimens. The end result is the Ontology-Based eXtensible (OBX) data model.

Methods and Results

Two reference ontologies were chosen as the foundation for OBX design – Basic Formal Ontology (BFO) and Ontology for Biomedical Investigation (OBI). BFO (http://www.ifomis.org/bfo) was originally conceived of by Smith and Grenon as an upper level ontology that could serve as a framework to support the development of domain-specific ontologies for scientific research[4]. The BFO structure is based on the central dichotomy between objects (continuants) and processes (occurrents), reflecting their distinct relationships with time. Continuants endure through time and retain some notion of their identity even while undergoing various kinds of changes. Occurrents unfold in time and can be defined to include temporal starts and ends. Continuants can be further sub-divided into those physical objects that exist independent from other entities – independent continuants (e.g. organs, tissues, cells, molecules, etc.) – and things that depend on physical objects for their existence – dependent continuants (e.g. the color red, the investigator role, the ribonuclease molecular function). OBI (http://purl.obofoundry.org/obo/obi/) builds upon BFO, extending the core structure by describing those entities that are specific to the biomedical research domain. For example, occurrent is extended to include subtypes of various planned process like biomaterial transformation, assay and data transformation; independent continuant is extended to include biomaterial and instrument; dependent continuant is extended to include investigator role, analyte role, evaluant role. Both BFO and OBI have been built using a strict is_a hierarchy of type/subtype relations and are compliance with the principles for ontology development best practices as promulgated by the OBO Foundry (http://www.obofoundry.org/crit.shtml).

In order to determine if the structure of these ontologies could be used to build a database that could support the management of a wide range of data derived from clinical and translational research studies, we extracted the core structure of the OBI extension of the BFO and developed a conceptualization of the core components as a starting point for data modeling (Figure 1A). The central component of the core conceptual model is the *Event* table, which includes descriptions of the actual events that happened it the study. These actual events may or may not be planned. A planned event is a realization of *Procedure Specification*; this separation allows for situations in which the actual event deviates from what was planned. Each event may also include one or more objects that play defined input and output roles. Each event also occurs in a specified time context. And finally, each event occurs in the context of a specific study that describes the actual realization of a study design.

Next we took this OBX Core Conceptual Model and used it as a framework to describe specific entities that need to be described in the clinical research database component of ImmPort and how they relate to each other. Again, we relied on OBI/BFO ontology design principles to capture the specific distinctions of the specific entities. For example, events are further specified to include *Biomaterial Transformation* defined as events with one or more biomaterials as inputs and outputs, which can be further specified into *Merging* and *Biosampling* subtypes (Figure 1B). One example of an important merging type of event is the substance intervention, whose description includes details about the type of compound included, and the formulation, dose and route of delivery used. In the case of biosampling, the input subject and the output biosample are specified in the data model. In this way, a wide variety of different events can be defined by describing the event type, the input and output continuants and the roles that they play in the process. In addition, the OBX Core is also compatible with the modeling of unplanned processes, including protocol deviations and adverse events of critical importance to the clinical research domain.

The following class and class subtypes have been modeled in this way:

- Object – population, population arm, human subject, animal subject, biological sample, compound, complex compound, software, instrument, site;
- Biomaterial transformation – substance merging, device intervention, surgery intervention,

biosampling process, environment exposure process;

- Assay – subject assessment, lab test, questionnaire, medical history taking, ECG;
- Data transformation – diagnostic process, research data analysis, outcome measure process, baseline characteristic process, protocol deviation determination.

In each case, the subtype tables contain attributes that are specific to the given subtype. In some cases we have made practical decisions to directly link tables even though they could be indirectly linked through table joining procedures in order to optimize database performance. A complete representation of the resulting OBX Conceptual Model can be found at http://pathcuric1.swmed.edu/Research/scheuermann/OBX.html.

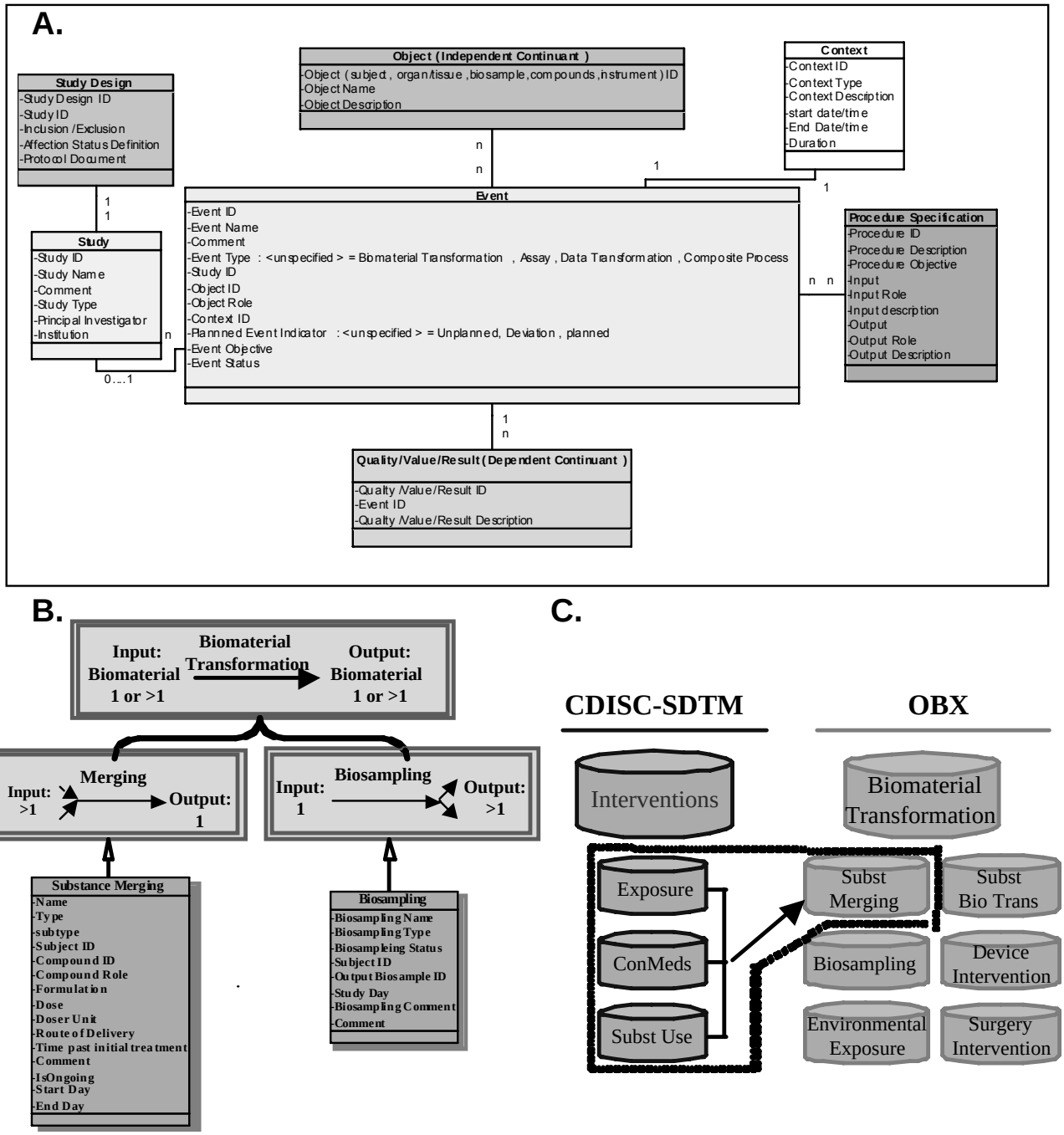


Figure 1. Schematics representations showing the OBX Core Conceptual Model (A), the specification of subtypes of biomaterial transformation events (B), and a comparison of interventions types from the CDISC-STDM with the relevant component of OBX (C).

We compared the OBX approach to other relevant data representation approaches in the clinical research domain. The Clinical Data Interchange Standards Consortium (CDISC, www.cdisc.org), a global multi-disciplinary organization, has developed a set of clinical data standards to facilitate global clinical data interoperability and exchange. The Study Data Tabulation Model (STDM) is one of the data standards developed by CDISC, which has been adopted by the U.S. Food and Drug Administration (FDA) to be the standard format for clinical trial data to be submitted to the FDA. In STDM, observations collected during the study are divided into three classes: Interventions, Events, or Findings. Interventions class captures investigational treatments and is further divided into three domains (Figure 1C): Concomitant Medications (ConMeds), Exposure, and Substance Use (Subst Use).

The OBX model places CDISC-STDM Interventions class under Biomaterial Transformation given that both the input and output in the Interventions class are biomaterials. The inputs for Concomitant Medications, Exposure or Substance Use are subjects of the study and concomitant medications, investigational drugs or self-administered substances, respectively. The outputs of these interventional processes are also subjects of the study. However, instead of using three different domains to represent essentially the same process, OBX recognizes the difference between Concomitant Medications, Exposure, and Substance Use is the role that the substance plays in this Substance Merging process (see Figure 1B). By adding a Compound Role attribute, the Substance Merging class encompasses the information that is captured in all three STDM interventional domains.

Discussion

The Ontology-Based eXtensible data model was developed to support the implementation of the clinical research database component of the ImmPort system. We are currently in the process of mapping components of a variety of clinical studies from the Atopic Dermatitis Vaccinia Network and the Immune Tolerance Network into this model representation. Based on this exercise, we are continuing to refine the conceptual model to ensure that we can not only describe the basic entities in a clinical study, e.g. human subject, biosamples, assays, assessments and assessment results but also the more complex components of a clinical study, e.g. protocol deviations, adverse events, study arm specifications and composite events like the clinical visit.

During the refinement process, several advantages of the OBX approach have been noted. The relatively simple structure of OBX has made it relatively easy to add new class tables to the schema without disrupting the existing structure. The logical framework used provides a consistent mechanism for linking component entities together. It is relatively easy to re-use entity tables as needed in generating primary key-foreign key relationships. The fact the OBX is based on the logical framework of BFO/OBI allows for its obvious integration with ontology term use as values for specific data elements in the database record instances.

We have recently completed a physical database schema based on this model, which is made freely available at www.immport.org. The OBX schema is being used to support the capture, managements and query of clinical research data in the ImmPort system for the National Institute of Allergy and Infectious Disease. Adoption of OBX by other organization interested in managing clinical research data would support data sharing, system interoperability and semantic query of this value data content.

Acknowledgements

We would like to thank the OBI consortium for helpful discussion about biomedical investigations that forms the basis for the described work. Supported by NIH N01AI40076 and U54RR023468.

References

1. Lee JA, *et al*. (2008) “MIFlowCyt: The Minimum Information about a Flow Cytometry Experiment” Cytometry: Part A 73(10): 926–30.
2. Taylor CF, *et al*. (2008) “Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project” Nat Biotechnol. 26(8):889–96.
3. Smith B, *et al*. (2007) “The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration” Nat Biotechnol. 25(11): 1251–5.
4. Grenon P, Smith B and Goldberg L. (2004) “Biodynamic Ontology: Applying BFO in the Biomedical Domain” from DM Pisanelli (ed.), Ontologies in Medicine, Amsterdam: IOS Press, 2004, 20–38

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

An Advanced Clinical Ontology

Riichiro Mizoguchi[1], Hiroko Kou[1], Jun Zhou[1], Kouji Kozaki[1], Takeshi Imai[2], Kazuhiko Ohe[2]
[1]The Institute of Scientific and Industrial Research (ISIR), Osaka University, Japan
[2]Department of Medical Informatics, Graduate School of Medicine, The University of Tokyo, Japan

Abstract

This article discusses a fundamental issues of medical ontology based on ontological theory. We focus on "anatomical structure of organs" and "abnormal states in the human body". On the basis of the investigation, we distinguish organ-specific types from those independent of any organ to maximize the explicitness of ontology. The next feature of our ontology is to allow on-demand reorganization of is-a hierarchy of diseases instead of one fixed hierarchy to cope with various viewpoints which physician might have. We also take care of the notorious issue related to conflict of is-a and part-of relations.

Introduction

Ontology is one of the most promising techniques for enabling semantic interoperability of medical information among various data across domains/tasks. This is why there have been developed some ontologies such as SNOMED-CT[1], ICD-10[2], Galen[3], etc. In this situation, there has been strong need of a sophisticated medical ontology in Japanese which is highly expected to compatible with those existing ontologies. The authors believe that the ontology cannot be a simple translation of the existing ontologies because that would hide some possible concepts specific to Japanese clinical practice. We should first establish our own ontology to reflect Japanese clinical practice and then investigate alignment between the Japanese one and existing ones to make them interoperable with each other. Another reason for this policy is that those existing ontologies suffer from so-called "legacy problem", that is, some of them are incomplete in terms of ontological theories since they had started their project when ontological engineering was not matured enough. As a late comer, we aim at building a medical ontology which is ontologically sound.

In this background, the Japanese ministry of health, labour and welfare has launched a three-year project on Foundation of Database for Clinical Knowledge in 2008. The expecting deliverable is a clinical ontology composed of roughly 30,000 concepts or more covering a couple of thousands of diseases in typical clinical and anatomical domains. This paper is an intermediate report on the ontology development conducted in the project and is structured as follows. The next section discusses the underlying policy in the ontology development. Human body structure with the focus on organs is discussed in Section 3. Diseases are discussed in Section 4. Section 5 presents related work to locate our project in the right context followed by concluding remarks.

Underlying Policy

Our ontology is being developed having the following issues in our mind. These issues lead us to introduce several new theories and ideas as explained below.

a) Commonality vs. specificity: In order to make it more articulate, common characteristics and target-thing-specificities should be clearly captured and differentiated. We introduced *generic structural /disorder* components each of which represents common characteristics of structural and disorder components as much as possible.

b) *is-a vs. part-of* issue 1: For example, the two relations <disease of a pulmonary valve *is-a* disease of heart> and < pulmonary valve *part-of* heart> cause a problem, since both "disease of a pulmonary valve" and "disease of a heart" have a slot of site of the disease and the filler of the former must be a subclass of the latter from the theory of inheritance, in reality, however, the former must be a part of the latter. To solve this problem, on the basis of our latest theory of roles[4], we introduced *"p-"* operator in our ontology building tool Hozo[4,5] which automatically generates a generic concept representing all the parts of the thing the operator is attached.

c) *is-a vs. part-of* issue 2: The atrium is composed of left and right atriums. At the same time, however, both left and right atriums are subclasses of the atrium. Fortunately, the *"p-"* operator can solve this issue at the same time.

d) No single hierarchy of diseases does not work well for all the stake holders such as pathologists, clinicians and surgeons, etc. To cope with these various viewpoints, we introduced an innovative technique to realize **on-demand** reorganization of *is-a* hierarch according to the specified viewpoints.

Human Body Structure

Upper-Level Types

Figure 1 shows the top-level types of the structural part of human body. *Organ in general* consists of *organ* and *organ system*. *Organ* consists of *internal organ, body part, portion of tissue & cells* and *generic structure* component. *Internal organ* represents ordinary organs such as heart, *portion of tissue & cells* includes finer-grained organs such as *gastric gland* as well as tissue and cells, *body part* includes structural parts such as face, arms, legs, et al. The design rationale of the top-level structure is that to represent it in a compact recursive structure reflecting the essential properties of several important types of organs. In fact, the nested structure of *cells, tissues, minimal-organs, organs* and *organ system* are nicely represented in the recursive structure. The details of *generic structure* are explained below.

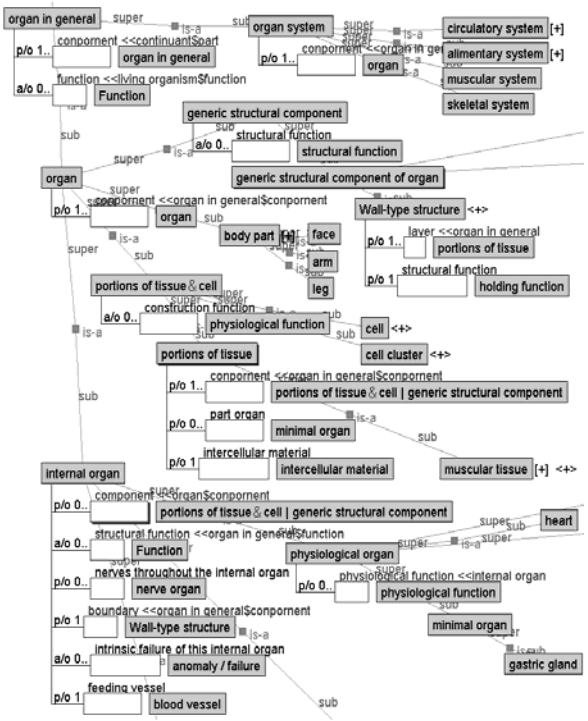


Figure 1. Top-level structure of human body.

Generic Structural Components

Many of the organs which consist of several sub-organs share common structural components. For example, we can identify *hollow structure* component in a stomach and atrium. *Tubular structure* components are found in *blood vessel* and *esophagus*. Although the materials are different, *gastric wall* and *cardiac wall* share three-layer structure of tissue which we call *wall-type structure* component. Those common structural components enable us compact representation of those organs because common properties can be defined once at those components. Typical examples of the utility of such representation include that *tabular structure* has a potential malfunction of *arctation* caused by narrowed cross section area. When it occurs at *blood vessel*, it is called *angiostenosis* which would occur at *cerebral vasculature, coronary artery, etc.* as well. All share similar properties. *Arctation* occurs at esophagus and it blocks the flow of pieces of foods down to the stomach, which is analog to the *angiostenosis* case. All of them share widening operations as a possible treatment for its remedy, though how to implement it would be different from each other. Properties specific to each organ can be defined additionally or by specializing the properties inherited from the common structural component. For example, although vein and esophagus share the *tubular structure* component, vein has a valve of vein but esophagus does not. While both vein and esophagus are composed of a three-layer wall; that of esophagus has two-layer *muscle fiber* structure to perform *peristaltic action*.

In order to represent such specificity, we introduce the concept of Roles supported by Hozo which is a tool for building ontology developed by us[5]. Figure 2 shows the legend of type definition in Hozo as well as role definition in which "bike" is defined by specifying its part. "p/o" stands for "*part-of*" link. At the same time, a role named "front wheel role" is defined by referring to "wheel" defined elsewhere. In Hozo, an entity playing a role is called "role-holder". In the case of Figure 2, a wheel which is incorporated as a part of a bike and playing the role of "front wheel role" is thereby called a "front wheel". Hozo, thus, realizes representation the mutual dependency between the whole and its parts[4].

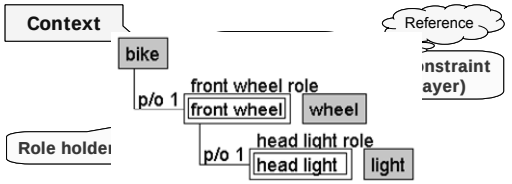


Figure 2. Legend of concept and role definition in Hozo.

Advanced Issue of *is-a* and *part-of* Relations

Contrary to the superficially simple characteristics of *is-a* and *part-of* relations, there have been two serious issues to solve such as b) and c) shown in section 2. Figure 3 shows the first difficulty b). Following the property inheritance, pulmonary valve must be a specialization of heart, but it is not. To represent the *is-a* relation between disease of heart and disease of pulmonary valve, we need to invent to inherit parts of heart rather than its subclass from heart.

The second difficulty seems to be more serious than b), since it is related to fundamental conceptualization of what is a whole of collectives.

This difficulty has been discussed by Udo Hahn and his group[6]. In order to solve this difficulty, they introduced SEP-triple which consists of three concepts: the original entity called E-node as a whole together with two concepts derived from the original one: One called S-node and the other called P-node. S-node is a super class of both E-node and P-node. The key idea is the introduction of a generic concept representing all the parts of the original entity under consideration.

We first tackled the issue b) and came up with a new operator named "*p-*" operator explained above. The operator enables parts to be inherited by ordinary property inheritance mechanism. In the case of Figure 3, for example, we write "*p-heart*" instead of "heart", then the slot of its subclass inherits not subclass of "heart" but its parts. Although this method would suggest we need complicated hidden processes in Hozo, it is not the case. When *p-X* is used, Hozo automatically generates a generic concept representing all defined parts of X including all parts which have X as their ancestor. This is valid because each part *is-a* subclass of "X's parts class" which coincides with *p-X*. According to mereology, the theory of parts, *p-X* includes itself which is not the very X as an entity but X as its part. This is why "*p-*" operator can solve the issue c).

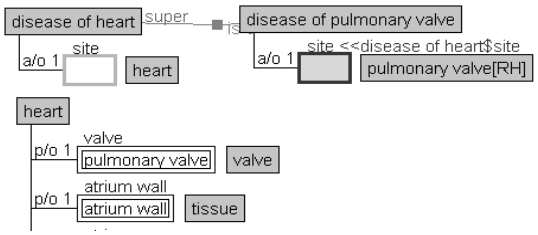


Figure 3. A problematic situation related to the property inheritance from super to sub concepts.

Definition of Disease

Disease as an Abnormal State

It is apparent that capturing diseases is a tough problem from the beginning. In fact, there are many ways of categorization of diseases. Patients use common sense names of diseases. What doctors of primary care deal with and what the government deals with to calculate statistics of the cause of death are very different. In addition, pathologists, clinicians and surgeons see the same disease different points of view. Those differences result in multiple taxonomies of diseases. When ontology developers build an ontology, they tend to present one *is-a* hierarchy believing it is the essential structure of the world under consideration. Although it is often true, it is not the case in medical ontology in which disease classification is essentially perspective-dependent as we see above. To cope with this well-known difficulty, we adopted the strategy as follows: (1) building the most fundamental *is-a* hierarchy of diseases based on "state" and (2) on-demand generation of *is-a* hierarchy according to the viewpoint specified. Ontologically, pathological state, disease, symptom, syndrome, disorder, dysfunction, failure, cause, etc. are kinds of disorder of human body and can be represented as "states". On the basis of this fact, the top-level categories of disease are developed as shown in Figure 4.

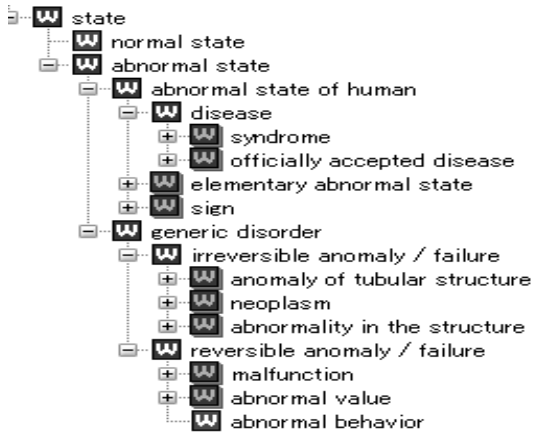


Figure 4. Top-level categories of state.

Upper Level Categories of Disease Quality

State, the top-level category, has *normal state* and *abnormal state*. The latter has two subclasses such as *abnormal state of human* and *generic disorder* which is the type corresponding to the *generic structure* component. *Officially accepted disease* is the central type of disease and is defined by referring to other states. *Elementary abnormal state* is disorder which cannot be disease by itself and is mainly used for characterizing *officially accepted disease*. The main component of disease is *pathological state* which is a role played by *abnormal state* in the context of diseases which clinical experts accept as disease. Basically, each state is defined in terms of <Entity, Attribute, Value>, EAV-triple. We have investigated the survey of the current clinical practice in Japan and found EAV-triple works quite successfully. We also analyzed quality descriptions in ICD-10 and Galen and found that most of them are covered by the ontology of quality and quantity defined in YATO[7] and we can convert them into the form of EAV-triple.

Officially Accepted Disease

Figure 5 shows the framework of *officially accepted disease*. It is defined by specifying typical *disorder roles* played by *abnormal state* which is based on EAV-triple. Depending on expert's decision, some abnormal states become

pathological state, some become *symptom* derived by the *pathological state*. This is our way of defining diseases based on ***states*** with the help of role-defining function of Hozo. Exploiting the fundamental characteristics of states and the expressive power of role representation of Hozo, definition of disease can be easily adjusted to the current understanding of the disease under consideration, which makes the ontology both solid and flexible at the same time. In fact, the same abnormal state can be *pathological state* or *symptom* according to the context of disease of interest. Furthermore, the fact that boundaries between those states are intrinsically vague prevents us from defining them as established types which are hard to change. Roles which intrinsically change according to the context best fits to definition of those states.

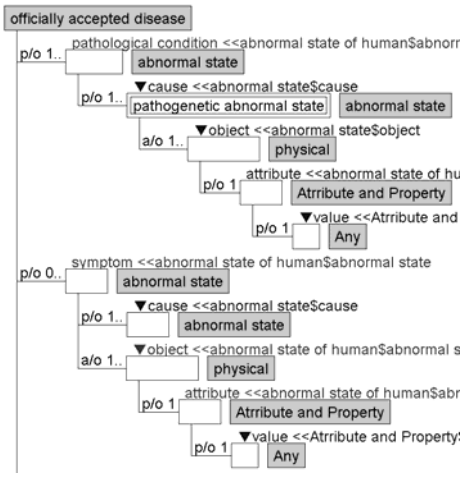


Figure 5. The framework of *officially accepted disease.*

On-Demand Reproduction of is-a Hierarchy

As discussed above, on-demand function is critically important to make clinically acceptable for experts in many divisions in medicine. The state-based disease definition with Hozo tool allows us to achieve this demanding goal. Diseases are defined as subclasses of *officially accepted disease* shown in Figure 5 and they have several slots with classes for constraining slot values. In Figure 5, examples are *abnormal state* and *physical*. If users want to see *is-a* hierarchy of diseases in terms of the partonomy of organs, then they just specify *physical* which is where diseases locate. Of course, the partonomy is converted into *is-a* hierarchy by Hozo using *"p-"* operator technology. If they want to see the hierarchy in terms of *pathological state*, then they specify it. We have already built a prototype system for this functionality and confirmed its feasibility. We have also manually mapped the ICD-10 concepts to ours in the prototype system of ontology navigation to enable users who want to browse our ontology from the ICD-10 contents. The demonstration of the functionality is available at: http://www.ei.sanken.osaka- u.ac.jp/MedOnto/.

Concluding Remarks

We did preliminary comparison between our ontology with existing ontologies such as SNOMED-CT, FMA, CARO[8], ICD-10 and GALEN and confirmed the ontological soundness of our ontology which is compliant with YATO which is comparable to BFO[9] and DOLCE[10]. In addition to this, it has major advantages over them with respect to the following three perspectives: 1) explicit representation of commonality and specificity of structure and diseases, 2) resolution of the notorious problem of inter-dependence between *is-a* and *part-of* relations and 3) on-demand reorganization of *is-a* hierarchy of diseases. In the preliminary comparison, we investigated FMA in terms of the difficulty 2) and found that FMA tries to solve it by introducing a lot of redundant virtual classes and ends up with partial solution of the problem in the sense that it fails to solve the issue of c) in Section 2. We are currently in the phase of increasing diseases of several clinical divisions by tight collaboration with clinical doctors using a description support system we developed to help them input data.

Acknowledgement

This research is supported by the Ministry of Health, Labour and Welfare, Japan.

References

1. SNOMED-CT, http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html
2. ICD10, http://www.who.int/classifications/icd/en/
3. OpenGALEN, http://www.opengalen.org/
4. Mizoguchi R, *et.al*. A Model of Roles within an Ontology Development Tool: Hozo, Applied Ontology, 2007:2:159-179
5. HOZO, http://www.hozo.jp/
6. Hahn U, *et al*. Turning Lead into Gold? EKAW2002, LNCS; vol. 2473:pp182-196.2002
7. Mizoguchi R. Yet Another Top-level Ontology: YATO, Proc. of InterOntology 2009: 91-101
8. CARO, http://www.bioontology.org/wiki/index.php/CARO:Main_Page
9. BFO, http://www.ifomis.org/bfo
10. DOLCE, http://www.loa-cnr.it/DOLCE.html

Concepts, Confusion and Modeling

Harold R. Solbrig, Christopher G. Chute
Mayo Clinic, Rochester, MN, USA

Abstract

The term "concept" continues to be used in models to reference categories, classes, universals, individuals and other less well defined artifacts. The fact that this obfuscates the purpose and usefulness of the model itself has already been well documented. Here, we show how the use of the term "concept" as a class name in a model can introduce serious confusion and propose a simple way that such confusion can be avoided.

Introduction

There are many modeling efforts underway that attempt to address the relationship between knowledge about the external world and information recorded in databases, forms and messages. Despite admonitions to the contrary[1,2], most of these models still use the term "concept" to designate a variety of different entities (and non-entities), including individuals, universals, categories, words, imaginings, etc.

With a couple of notable exceptions[3,4], the term "concept" is also used as a label for a class within the models themselves. In the sections that follow, we begin by discussing the role of class labels in modeling meta-levels and then show how using "concept" as a class name obfuscates the intent of the model and introduces crippling confusion. We then discuss a simple solution that, while not addressing the more fundamental issues introduced by of the use of the notion of "concept" itself, at least helps to clarify the purpose of such a class in this sort of modeling effort.

M0-M3 Modeling Levels

Modelers typically assign labels to classes, attributes and associations in a model that match the name of external entity being modeled. An exception to this rule, however, is when the models are used to describe aspects of the modeling effort itself. In this situation, modelers have learned the importance of unambiguously differentiating the names of the modeling artifacts from the names of the entities being modeled. Let's start with a simple example, "person".

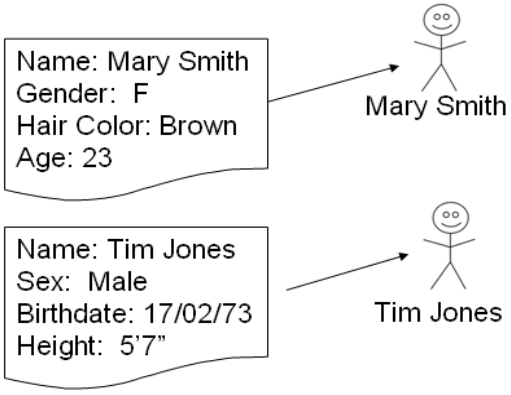


Figure 1. M0 Modeling Level

In figure 1, we've recorded some information about two "real world" (well, stick figures for this document, but they represent real world) people. This information is recorded at the M0 metalevel[5].

In order to catalog and share information about people in general, we need to arrive at:

1. An understanding and definition of the abstraction (universal) that we wish to represent. In this case, we might decide that the abstraction is "Human".
2. A list of the salient characteristics that are shared by representative members of this abstraction. (Figure 2)

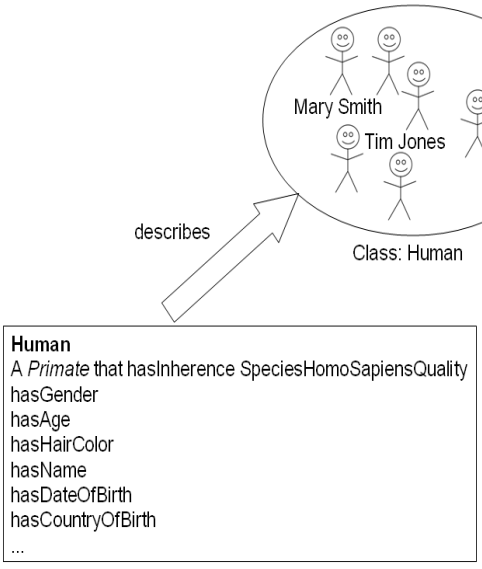


Figure 2. Abstraction of Characteristics

We then need to use the modeling context to determine which of these characteristics are:

(a) Are assumed to have fixed values
(b) Are considered irrelevant
(c) Carry real information

We then add any additional characteristics that are specific to the model itself, such as a unique identifier for the instance record. Finally, we assign labels to characteristics in category (c) above and produce a model that becomes the M1 model.

In our example, we may be constructing a database of people who were born in Iceland, meaning that the country of birth attribute is fixed. We may determine that the hair color and age are irrelevant, and that the information that we want to carry is the person's name, their gender and how old they were at some known point in time (Figure 3).

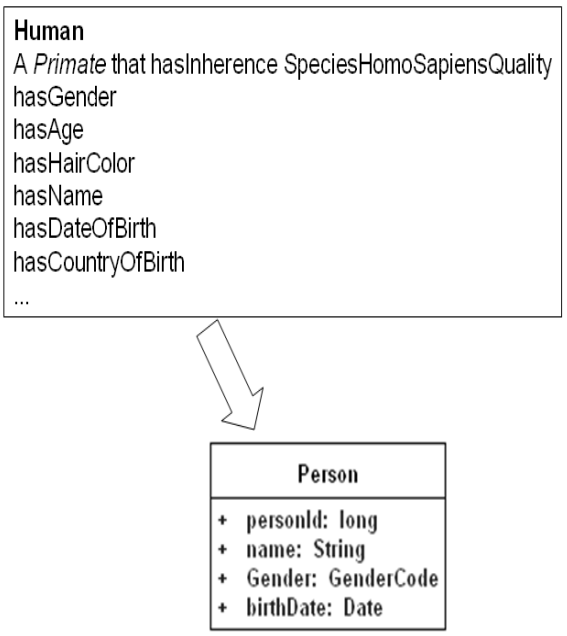


Figure 3. M1 Modeling Level

This model is of little value, however, unless we have a shared understanding of the semantics of the model itself. We need to understand what "class", "attribute", "data type", "visibility", etc. represent and how they are rendered in a class diagram. To do this, we repeat the process described above but, instead of describing people, age, hair color, etc. We now describe *class*, *attribute*, *visibility*, etc. The result of this process is a "meta-model" – a model of a modeling language.

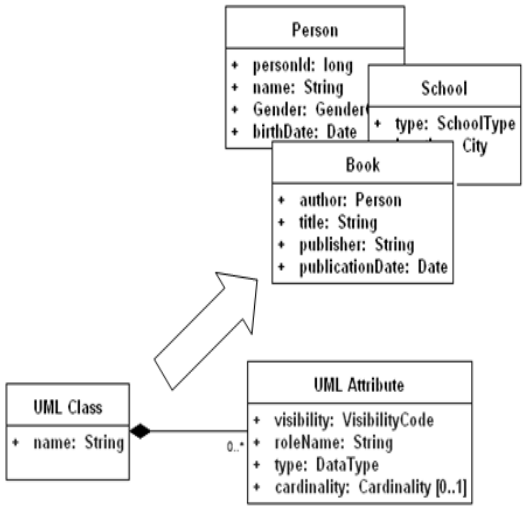


Figure 4. M2 Modeling Level

We now have a model of the semantics of what is, in our case, a UML model (Figure 4). We still need one more step, before we can close the loop – we need a model of the language of modeling itself (Figure 5):

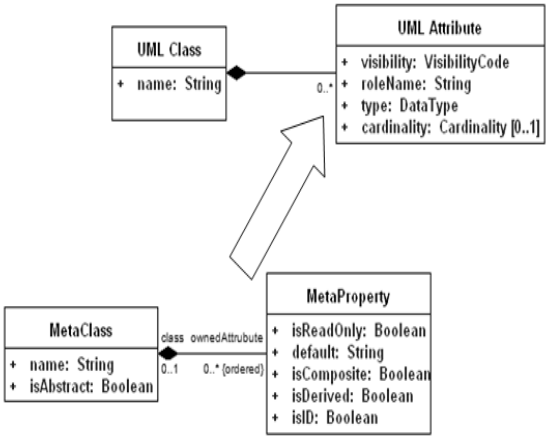


Figure 5. M3 Modeling Level

The M3 level gives us closure, as M3 level models are either axiomatic or self describing. Note how different labels have been used for the M2 and M3 level entities. Even though they appear very similar, they are *referencing different things*. The M2 level describes a model and the M3 level describes the language of modeling itself.

"Concept" in the Modeling Levels

We now repeat this description, but this time we will replace models involving people with models of the *descriptions* of entities, which, unfortunately, are typically called "concepts".

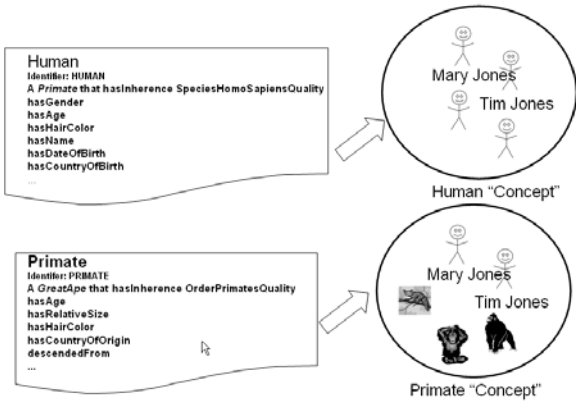


Figure 6. M0 Modeling Level for "Concept"

In the figure above, the concept of "human" and the concept of "primate" become the real world entities about which we record information.

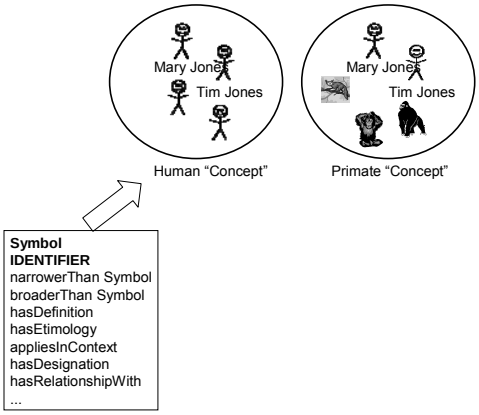


Figure 7. Abstraction of Characteristics

We then proceed to abstract the salient characteristics of "concept" and determine the fixed, irrelevant and information carrying characteristics in our particular context (Figure 7). As with the previous example, we may also add modeling specific information to arrive at an M1 level model (Figure 8).

This is where the confusion begins. In the previous example, instances of the class "Person" are obviously data records which, in turn, are *about* real world people. In this example, however, this distinction isn't clear. Following our previous model, instances of the class "Concept" should be data records which, in turn, are *about* real world concepts. The notion of "concept" is so ephemeral, however, that it is tempting to try to assert that, instead of being *about* a concept, the data record *is* the concept. If you define "concept" as "A unit of knowledge created by a unique combination of characteristics," you need to know when you have got one of those "units", and a database record is an obvious candidate. Defining "concept" as "anything that can be conceived or perceived" does little to help, as the data record may well be more conceivable or perceivable than the "thing" that it is intended to represent.

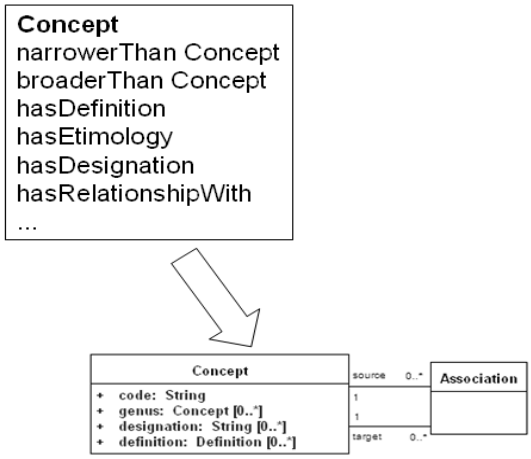


Figure 8. M1 Modeling Level

Once we've allowed the M0 level data record to *become* the target of the M1 class model, we find ourselves in a quandary. How do we differentiate the M0 instance "human" which describes the external universal, "human" and the M1 instance "person" in the previous example which describes an M0 description of a particular person? Even more insidious, however, is the discovery that the M1 class "Concept" now *describes* concepts and there is strong temptation to say that "Concept" is an instance of itself. This, in turn, leads to countless hours of useless discussion about "Isn't *Person* a concept?", "Why haven't we made *Ordered List* a subtype of concept?", "Do we have the concept "red" stored in our database?" and "Do all concepts have codes?" One is reminded of a scene from the movie, *Being John Malkovich*, where John physically ventures inside his own head – recursively perceiving how he perceives the world. The result is mass confusion – everything that John Malkovich perceives *is* John Malkovich. All the characters he interacts with are himself and the only words that are spoken are "Malkovich, Malkovich, …".

How Can We Fix This?

We need to focus on the misunderstanding that got us into this mess to begin with – the confusion of a data record that *describes* a concept with the concept itself.

The obvious solution to this problem would be to stop using the term "concept" altogether. As argued by Smith and many others, our business is *not*

creating models of anything that can be perceived or conceived – our business is creating models of reality – models that can be communicated, verified and used constructively in the discipline of science. The notion of "concepts" obfuscates the model and makes errors inevitable. Unfortunately, to date, this is still a losing battle. The best can do is to mitigate the damage.

The key fact that we need to recognize that (a) we are working with a *model* of universals and (b) like it or not, people are going to continue to refer to these as "concepts" and because of this we are unable to readily different the model and the thing being modeled. This is the same problem faced by the M2 and M3 models described earlier – when you are using a class in a model to describe an *instance* of a class in another model, it is important to give the class and instances different names. In the previous example, we named the instance "UML Class" and the class "MetaClass".

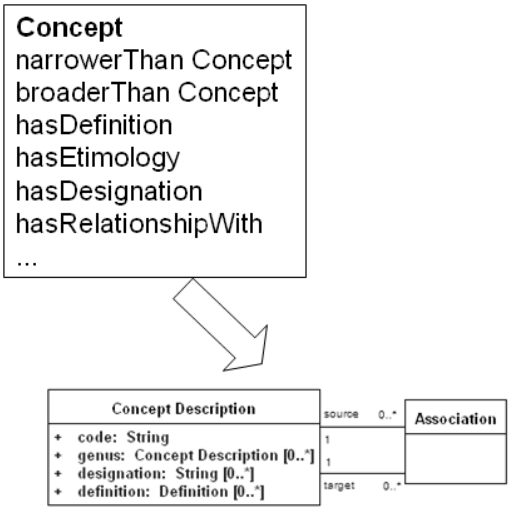


Figure 9. Revised M1 Level Model

Just like the M2 / M3 levels above, *we need to assign the M1 model element a different name!* Were we to replace the label "Concept" with "Concept Description" at the M1 level, we have effectively broken the circuit. Now, the M0 "Primate" is clearly an instance of a Concept Description, which, in turn describes the external entity, "Primate". Where it previously seemed to make sense to ask "Isn't *Concept* a concept?" the question "Isn't *Concept Description* a concept" is somewhat more sensible.

Some of the questions we posed previously now have more obvious answers. "Is a person a concept description?" "Why isn't an ordered list a concept description?" "Do we have the concept description "*red*", in our database?" Concept descriptions may have codes – but concepts do *not* have codes, at least in any normal sense of the word.

Additional benefits include the recognition that concept (description) systems contain sets of concept descriptions, which, in turn, reference concepts. Questions about can the same concept *be* in more than one concept system become irrelevant, as the concepts aren't in a concept system. The question, instead, is can the same *concept description* belong to more than one system, which can be answered by a sensible discussion about what constitutes identity for a concept description rather than debating about what constitutes the identity for a concept.

Summary

The ideal solution to the "concept" problem would be to cease using such an ambiguous and confusing label. Short of that, however, much of the current modeling confusion could be eliminated through the simple act of renaming the class "concept" to more accurately reflect what it actually represents.

References

1. Smith B, Ceusters W and Temmerman R. Wüsteria. Proceedings Medical Informatics Europe 2005. 116: 647–652.
2. Smith B. Beyond Concepts: Ontology as Reality Representation. Proceedings of FOIS 2004. 73–84.
3. Motik B, Patel-Schneider P and Horrocks I. OWL 1.1 Web Ontology Language Structural Specification and Functional Style Syntax. 2007. http://www.webont.org/owl/1.1/owl_specification.html
4. HL7 Common Terminology Services v. 1.2. http://informatics.mayo.edu/LexGrid/downloads/CTS/specification/ctsspec/cts.htm
5. Frankel D. Model Driven Architecture. John Wiley and Sons, 2003. 105–109.

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

A Quality Evaluation Framework for Bio-Ontologies

Jesualdo Tomás Fernández-Breis[1], Mikel Egaña Aranguren[2], Robert Stevens[2]
[1]University of Murcia, Murcia, Spain; [2]University of Manchester, Manchester, UK

Abstract

Over the past few years the number of bio-ontologies has rapidly increased. The evaluation of ontologies has long been a problematic issue. The growing number of ontologies makes the need for a strategy for evaluating quality more urgent. We propose a framework for evaluating the quality of bio-ontologies. This framework is inspired by a well-known software quality standard, which has been adapted to the needs of ontology evaluation. An example of how to use the framework, comparing two versions of the Open biomedical Ontologies' Cell Type Ontology, is included as an illustration.

Introduction

Bio-ontologies have increased in number and importance since the development of the Gene Ontology. Many research groups are collaborating in the development of an orthogonal collection of bio-ontologies, the Open Biomedical Ontologies (OBO) Foundry (http://www.obofoundry.org). In addition, there also exist independent efforts for developing other bio-ontologies. The development of application ontologies, for example, usually requires the reuse of different ontologies, so bits from different ontologies have to be combined. For this purpose, developers have to decide which ontology to use, but they lack support for making an informed decision. Hence, there is a clear need for methods for evaluating the quality of bio-ontologies. Ontology quality evaluation has usually been the concern of the Ontology Engineering community, and has been addressed from different perspectives and hence related work in ontology evaluation can be classified according to the particular evaluation aim: ranking, correctness, or quality.

Ontology Engineering has historically adapted methods from the Software Engineering field since they have many stages in common. Recent examples are ontology development methodologies[1] or Ontology Design Patterns[2]. There has not, however, been any attempt to adapt Software Engineering approaches for evaluating ontology quality. In this work, we propose an evaluation framework for bio-ontologies that is inspired by the ISO 9126 (http://en.wikipedia.org/wiki/ISO_9126) standard for software quality, which has been applied in other fields, for different purposes, such as the evaluation of e-learning systems[3] or software design documents[4]. Its application is recommended because: (1) it provides a comprehensive specification and evaluation model for software product quality; (2) it addresses user needs of a product by allowing for a common language for specifying user requirements that is understandable by users, developers and evaluators; (3) it objectively evaluates quality of software products based on observation; and (4) it makes quality evaluation reproducible. All these properties are desirable for an ontology quality evaluation approach, and hence they represent a potentially useful tool e such a framework.

Furthermore, this standard does not attempt to provide mechanisms for accumulating the metrics into an overall numeric evaluation. Given the different possible uses of ontologies, there is no need for such mechanisms, but rather there is a need for mechanisms capable of indicating which ontologies are more appropriate for particular situations. Also, this standard incorporates elements from the state of the art on ontology evaluation frameworks. An example of the usage of the framework is provided by evaluating two versions of the Cell Type Ontology[5]: the OBO version and a version that was re-engineered using a technique called Normalization[6].

Framework for Bio-Ontologies Quality Evaluation

In Software Engineering, software quality measures the quality of software design, and to which extent the software conforms to that design. The ISO 9126 standard for software quality evaluation provides a model based on internal, external and in-use quality metrics: functionality, reliability, portability, usability, maintainability, efficiency, effectiveness, productivity, physical security and user satisfaction. An internal metric can be used for measuring an attribute of a software product, derived from the product itself, either directly or indirectly (it is not derived from measures of the behavior of the system).Internal metrics are applicable to a non executable software product during designing and coding in early stages of the development process. An external metric can be used for measuring an attribute of a software product, derived from the behavior of the system of which it is a part. External metrics are applicable to an executable software product during testing or operating in later stages of development and after entering to an operational process. Quality in use metrics are those applicable to the final product in real conditions.

Using such a standard as a reference for defining an ontology evaluation framework is reasonable due to the intrinsic benefits provided by the use of standards, and the context that it would provide for a systematic evaluation of ontology quality. Therefore we propose a framework for evaluating ontology quality based on such a standard. The framework comprises seven quality dimensions, and these categories have these evaluation metrics associated (Figure 1):

Structural: This category is the only one in this framework that is not specified as such in the ISO 9126, but it is important when evaluating ontologies, since it accounts for software quality factors such as consistency, formalization, redundancy or tangledness.

Functionality: How the ontology performs in its intended roles.

Reliability: Capability of an ontology to maintain its level of performance under stated conditions for a given period of time.

Usability: Readability and ease of reuse.

Efficiency: Relationship between the level of performance of the software and the amount of resources used, under stated conditions, taking into account elements such as the time response, or memory consumption. Unfortunately, the field of OE has not developed good mechanisms to evaluate efficiency appropriately.

Maintainability: The effort needed to make specified modifications, how changes affect the rest of the ontology, etc.

Quality in Use: Quality in a particular context of use, provided by the users.

Next, we describe the interpretation of some of the metrics, when applied to ontologies, as follows:

Structural – Formalization: An efficient ontology has to be built on top of a semantically strict model to support reasoning. In the case of bio-ontology languages, the Web Ontology Language (OWL) has a strict semantics, the Open Biomedical Ontologies language (OBO) does not have such semantic definition, but has been defined in relation to OWL[7].

Structural – Formal Relations Support: Most ontologies only have formal support for taxonomy. This would indicate if any other formal theories are supporting the relations. The evaluation of this criterion for a bio-ontology depends on the number of formally supported relations included in it, for instance, through the use of the Relations Ontology (RO)[8].

Functionality – Competence Adequacy- Consistent Search and Query: The formal model of the ontology allows for better querying and searching methods.

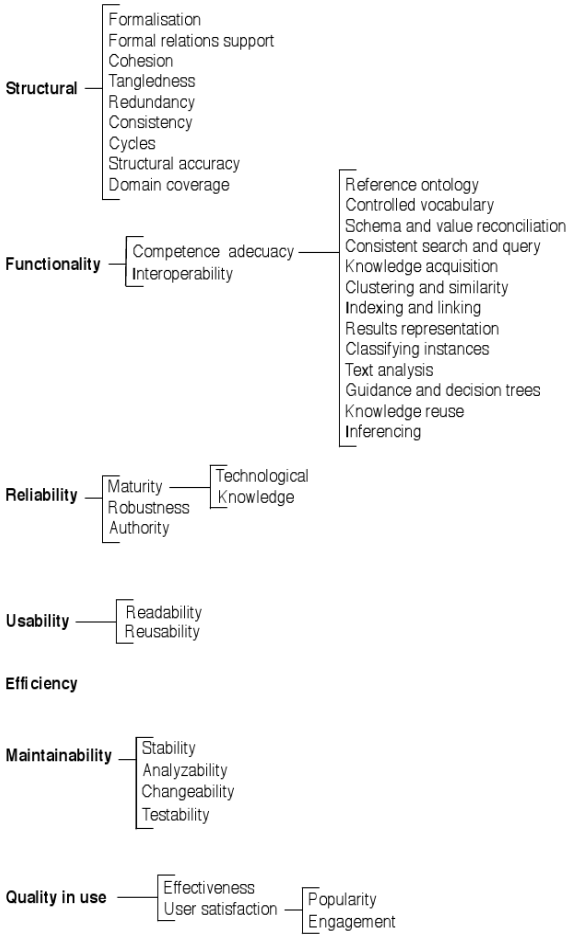


Figure 1. Evaluation framework.

Results

The Cell Type Ontology (CTO) was designed as a structured controlled vocabulary for cell types. CTO was constructed to integrate the model organism databases and other bioinformatics databases. In order to test the evaluation framework two versions of CTO were evaluated. The original version of CTO, oCTO, was the conversion of the OBO file to OWL. The normalized CTO, nCTO, was created by collaboratively dissecting the original CTO and then recreating the structure using reasoning (see: http://www.gong.manchester.ac.uk/odp/html/Normalisation.html).

The evaluation of the quality of these ontologies was performed by eight MSc students of the Semantic Web course at the University of Murcia. Before doing this work, the students were trained in this course for 20 hours in the design of ontologies, they analyzed

some of the most prominent ontologies (including biomedical ones) and they were also trained in the application of this evaluation framework. Then, they were given two weeks to evaluate both ontologies.

Each student had to fill in a form for each ontology, providing a quantitative evaluation for each quality metric included in the framework. The value ranged between 1(worst) and 5 (best). They were optionally allowed to provide comments on their evaluations. The usage of a quality evaluation framework does not require providing a numerical score for the evaluated items. In this case, we have averaged the results for each quality criterion for descriptive purpose, and all the quality criteria have been equally weighted. The results of this experiment are shown in Figure 2. A radar graph has been used for such purpose, since it allows an easy comparison of the quality of the two ontologies. The evaluators have given to nCTO a higher score in terms of structural, functional, usability, reliability and maintainability quality, whereas no big differences are found in terms of efficiency and quality in use.

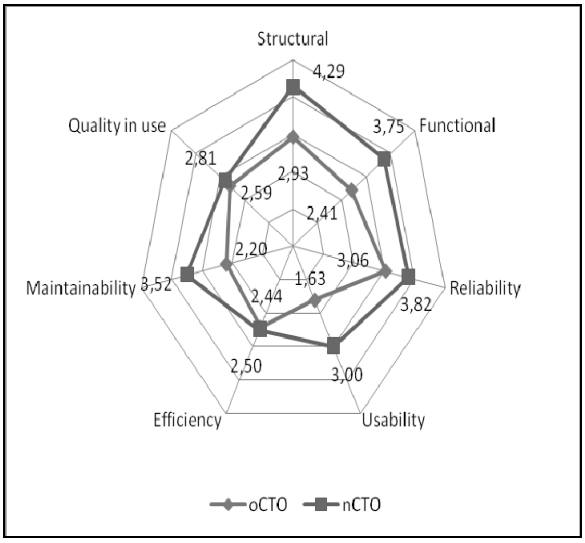


Figure 2. Results of the experiment

As has been mentioned, eight people have participated in this evaluation experiment, so the analysis of the degree of agreement between them is an interesting issue. All the evaluators gave a higher score to nCTO for the structural dimension; seven did so for functionality and usability; six did so for reliability and maintainability. It might be said that there is a consensus across these categories. In terms of efficiency, four evaluators gave a higher score to nCTO and three to oCTO. Four evaluators gave a higher value to nCTO and four to oCTO in the quality of use criterion. The evaluation of quality in use is the average of effectiveness and user satisfaction, which is split into popularity and engagement. In this sense, oCTO gets a higher score for user satisfaction, and a lower for effectiveness, due to its better structure. Hence, due to the effects of the numeric average, nCTO gets a slightly higher value for this quality dimension. So, in terms of efficiency and quality of use, there is no consensus. Both ontologies and the complete results of this experiment can be found at http://dis.um.es/~jfernand/icbo.

Discussion and Conclusions

The evaluation of ontology quality is a critical issue that remains unsolved. Different approaches accounting for different perspectives and aspects of ontology evaluation have been proposed in recent years, although none has become standard. In general, usability, reliability, and functionality criteria are identified in such approaches for evaluating quality, whereas those focused on ranking and correctness mainly consider structural properties.

In our opinion, the quality of an ontology is related to the degree of excellence. International quality organizations do not assign a numerical quality value to all kinds of processes and products, but they give them a quality stamp. This also occurs with software development processes. Such stamps certify their degree of excellence, which is checked against a series of criteria. The ISO 9126 has been criticized for being too general and abstract, and for not providing a concrete framework to be applied, obtaining a numerical evaluation as a result. The approach presented in this paper is based on the ISO 9216 and the framework includes most of the quality categories identified in the standard and incorporates the structural one to account for issues of particular importance for ontologies and it has been applied to two different ontologies, oCTO and nCTO. Both ontologies were built by applying a different methodology; oCTO was built in OBO and then transformed directly into OWL, and nCTO was built from scratch by applying the Normalization technique. This evaluation experiment has shown the usefulness of our approach, since we have obtained a vision of the quality of the ontologies, their strengths and their weaknesses, so that users have extensive information about the properties of both ontologies that can be used for making their decisions. In fact, quality evaluation approaches do not have to make decisions for the users, but provide enough information for them to make such decisions. As mentioned, the students were trained in the evaluation framework. This training consisted on explaining the meaning of the different quality dimensions used in the framework. Examples with ontologies were provided, using good practices in ontology

construction as the evaluation criteria. Obviously, the ontologies used in the training were not the ones to evaluate. Consequently, we think the scores were not biased by the training received by the students.

We were also concerned by how difficult the application of the framework could be and if this would require much technical knowledge. The students did not report problems in understanding how to apply it. This makes us think that any person with knowledge in ontology construction can do it as well without much effort. Another issue would be who should apply it and evaluate the quality of bio-ontologies[9,10], but this discussion is out of the scope of this work.

It should be said that this is early work, and that some improvements are needed. This experiment is as much an evaluation of the framework as it is of the ontologies themselves. In addition, the low number of relatively inexperienced ontologists makes any profound conclusions on the nature of the two ontologies suspect. We aim to design an objective quality evaluation framework, and this has been partially achieved in this work. First, the quality dimensions and criteria are the ones defined in the ISO standard, which provides an objective definition of quality evaluation. We have added the structural dimension and defined the concrete competences of an ontology. For this, we have used standard criteria for the structural dimension, drawn from the best practices and which are generally used for evaluation purposes in literature. Concerning competences, we are using the ones considered by the community. From this perspective, the framework is objective and not biased by our interests or preferences. What is not completely objective is the measurement of the values given by the experts. We will do further research in this area to gain objectivity in this part of the process. Finally, we plan to enrich the framework including metrics related to the ontology inference power based on the theory of justification[11].

Acknowledgements

This research was funded by the Spanish Ministry of Science and Innovation through the José Castillejo Fellowship JC2008-00120, EPSRC and the University of Manchester. Work on nCTO was funded by EPSRC-funded Ontogenesis Network (EP/E021352/1).

References

1. Lopez MF, Gomez-Perez A, Sierra JP and Sierra AP. Building a chemical ontology using methontology and the ontology design environment. IEEE Intelligent Systems 1999, 14(1):37–46.
2. Aranguren ME, Antezana E, Kuiper M and Stevens R. Ontology design patterns for bio-ontologies: A case study on the cell cycle ontology. BMC bioinformatics 2008, 9(Suppl 5):S1.
3. Chua BB and Dyson LE. Applying the ISO 9126 model to the evaluation of an e-learning system. In Proceedings of the 21st ASCILITE Conference, December, 2004, Perth, Australia.
4. Al-Kilidar H, Cox K and Kitchenham B. The use and usefulness of the ISO/IEC 9126 quality standard. In Proceedings of the International Symposium on Empirical Software Engineering, November, 2005, Noosa Heads, Australia.
5. Bard J, Rhee SY and Ashburner M. An ontology for cell types. Genome Biology 2005, 6(2):R21.
6. Rector A. Modularisation of domain ontologies implemented in description logics and related formalisms including OWL. Proceedings of the 2nd International Conference on Knowledge Capture, October, 2003, Sanibel Island, USA.
7. Horrocks I. OBO flat file format syntax and semantics and mapping to OWL. Available from: http://www.cs.man.ac.uk/~horrocks/obo/.
8. Smith B, Ceusters W, Klagges B, *et al*. Relations in biomedical ontologies. Genome Biology 2005, 6(5):R46.
9. Kalfoglou Y and Hu B. Issues with evaluating and using publicly available ontologies. Proceedings of the 4th International EON Workshop, Evaluating Ontologies for the Web, May 2006, Edinburgh, Scotland.
10. Obrst L, Hughes T and Ray S. Prospects and possibilities for ontology evaluation: The view from NCOR. Proceedings of the 4th International EON Workshop, Evaluating Ontologies for the Web, May 2006, Edinburgh, Scotland

    Horridge M, Parsia B and Sattle U. Laconic and precise justifications in OWL. In Proceedings of the 7th International Semantic Web Conference, October, 2008, Karlsruhe, Germany.

LexOWL: A Bridge from LexGrid to OWL

Cui Tao, Jyotishman Pathak, Harold R. Solbrig, Christopher G. Chute
Mayo Clinic College of Medicine, Rochester, MN, USA

Abstract

The Lexical Grid project is an on-going community driven initiative that provides a common terminology model to represent multiple vocabulary and ontology sources as well as a scalable and robust API for accessing such information. In order to add more powerful functionalities to the existing infrastructure and align LexGrid more closely with various Semantic Web technologies, we introduce the LexOWL project for representing the ontologies modeled within the LexGrid environment in OWL (Web Ontology Language). The crux of this effort is to create a "bridge" that functionally connects the LexBIG (a LexGrid API) and the OWL API (an interface that implements OWL) seamlessly. In this paper, we discuss the key aspects of designing and implementing the LexOWL bridge. We compared LexOWL with other OWL converting tools and conclude that LexOWL provides an OWL mapping and converting tool with well-defined interoperability for information in the biomedical domain.

Introduction

The Lexical Grid project (LexGrid)[2,12] coordinated by the Mayo Clinic Division of Biomedical Statistics and Informatics provides support for a distributed network of lexical resources such as terminologies and ontologies via standards-based tools, storage formats, and access mechanisms. The LexGrid system supports formats such as HL7 RIM, OBO, OWL/Protégé frame, UMLS RRF, and LexGrid XML. It models ontology information including versioning, provenances, entities, associations, and instances. LexGrid loads ontologies and terminologies from different sources, maps the information into the LexGrid model, and stores them in a backend database. Information modeled by LexGrid can be accessed through LexBIG, an interface that implements the LexGrid model, on top of which standard tools and services can be built.

A valuable augmentation to LexGrid is the adoption of Semantic Web technologies. The recent emergence of the Semantic Web and the Web Ontology Language (OWL)[4] is fostering a new level of interoperability. The biomedical informatics community greatly benefit by applying OWL's combination of formal semantics, rich expressiveness and shared software base to biomedical and clinical terminologies. The LexOWL project provides a round trip between LexGrid and OWL. In this paper, we focus on the direction from LexBIG to OWL. Through LexOWL, information modeled in LexGrid can be represented in OWL. Hence, tools and services that have been developed by the Semantic Web community can be directly applied to the biomedical and clinical domain. To name a few, we can use Protégé, which is a widely-used OWL ontology authoring tool, to browse and edit the information modeled in LexGrid. We can apply different reasoning tools to medical and clinical terminologies, to check consistency or to infer new knowledge. We can use OWL ontology modularity tools to integrate or extract ontology modules as well as use OWL ontology mapping tools to map ontologies. The biomedical terminology community has been actively seeking connections to OWL. OBO2OWL[1], OBOInOWL[9], Protégé OBO to OWL Tab[10], and Protégé 4 OBO loader provide mappings and conversions from OBO to OWL. The conversion from UMLS Semantic Network to OWL has been studied[6,8]. The NCI Thesaurus to OWL DL conversion is discussed in Noy, el al[11]. The International Healthcare Terminology Standards Development Organization also released a Perl converter for converting from SNOMED CT to OWL in recent SNOMED CT releases. LexOWL augments all these efforts by providing LexGrid a converter to OWL. Compared to the other tools, LexOWL has an inherit advantage in that, it can convert all the ontologies and terminologies from different sources modeled by LexGrid without individual mappers and converters. As an immediate benefit, LexOWL provides a well-defined interoperability across these sources since all the different resources are modeled by LexGrid.

We make the following contributions in this paper:

- LexOWL functionally converts LexGrid to OWL through an API bridge and represents the information modeled in LexGrid in the OWL API representation. By doing so, we can leverage the services and tools developed for OWL and the Semantic Web directly.
- LexOWL provides an OWL converter with relatively well-defined interoperability for different biomedical terminologies and ontologies.
- LexOWL provides a dynamic interface between LexGrid and Protégé so that Protégé can use LexGrid as its backend database, which could be a valuable addition to Protégé 4.

The rest of the paper is structured as follows. We begin with an overview of the LexOWL system in Section 2. In Section 3, we discuss how LexOWL maps LexGrid components to OWL. In Section 4, we compare the OWL ontologies exported by LexOWL to those converted by the existing tools. Finally, in Section 5 we summarize and consider future work.

LexOWL System Overview

Figure 1 shows the LexOWL system overview. The core component of LexOWL is the LexOWLManager. It manages both the LexBIG service through which we can access the LexBIG API, and the OWL Ontology manager through which we can access the OWL API. On the left hand side of the system overview, the LexGrid system loads ontologies in different formats from different sources, translates them to LexGrid representation as well as saves the knowledge to a relational database. Through the LexBIG API, LexOWL can access the ontologies loaded in the database. On the right hand side of the system overview, through the OWL API, LexOWL re-represents the information in the LexGrid database virtually to the OWL API Ontology representation, which can the be used directly by Protégé 4 and other Semantic Web tools.

Thus, in essence, LexOWL maps LexGrid to OWL on the API level. It is not just a tool that maps and converts from one format to another. In addition to that, it generates a "bridge" between the two APIs. The "bridge" accesses information from the LexBIG API and translates it to the OWL API's representations. The benefit of an API "bridge" is that even if the backend representations for ontologies change, the "bridge" still performs the same way and an update is not necessary.

We also defined the LexGrid to OWL mapping and a lexgrid2owl meta-ontology[3], based on which LexOWL can re-represent a selected LexGrid ontology to the OWL API representation. In the next section, we discuss how LexOWL maps LexGrid to OWL.

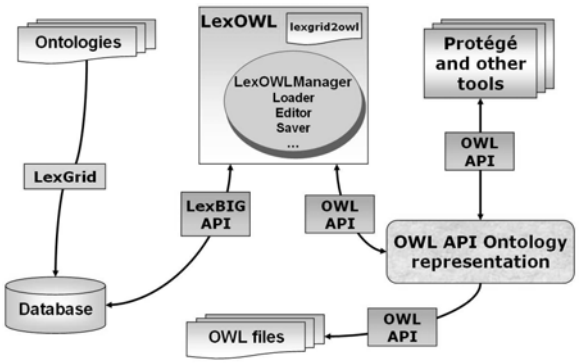


Figure 1. LexOWL System Overview.

LexGrid to OWL Mapping

LexOWL first maps the general ontology information. This includes information about the ontology itself such as name, version, and copyright. For some information, we can find equivalent representations in OWL (e.g., codingScheme to *owl:ontology*, localName to *rdfs:label*, and representsVersion to *owl:versionInfo*). For some information, we can find equivalent representations in standard name spaces such as dublin core (e.g., formalName to *dc:title* and copyright to *dc:rights*). We used the lexgrid2owl meta-ontology to represent the rest information (e.g, we define ApproxNumConcepts and isNative as two annotation properties in the meta- ontology).

LexOWL maps each LexGrid concept[†] to an OWL class. A concept in the LexGrid model can have properties such as a concept code, descriptions, presentations, definitions, and sources. LexOWL uses the concept code as the OWL class name and assign concept descriptions to a set of *rdfs:label*. In the lexgrid2owl meta-ontology, we define three OWL classes, Presentation, Definition, and Source, to represent the presentations, definitions, and sources in the LexGrid concept properties. We also defined annotation properties: *hasPresentation*, *hasDefinition*, and *hasSource* in the meta-ontology, to represent the relationships between concepts and such properties. Figure 2(a) shows a sample OBO term and Figure 2(b) shows its LexGrid representation. Figure 2(c) shows how LexOWL represents this concept and its properties in OWL. LexOWL creates an OWL class for the Concept Code "TAIR:0000055" and assign the Entity Description "pollen development" as a *rdfs:label*. The class has three annotation properties, one *hasDefinition* and two *hasPresentation*s, which link to "definition21" (an instance of the Definition Class), "presentation37", and "presentation38" (two instances of the Presentation Class) respectively. In addition, "definition21" has an annotation property *hasSource*, which links to "source21". Each of these instances also has annotation properties that represent contents such as synonyms and definitions from the source document.

LexGrid also has a special kind of concepts – anonymous concepts – which it uses to represent the anonymous classes in OWL. LexOWL parses each anonymous class and translates it back to OWL based on concept properties. Figure 3 shows an example. The upper part shows the LexGrid representation. The concept "A38" is the anonymous concept which is

[†] A "concept" represents a "kind" or "universal" entity in the LexGrid 2008 model. Here we still use "concept" to be compatible with LexGrid 2008. We are upgrading both LexGrid and LexOWL to avoid using this confusing label.

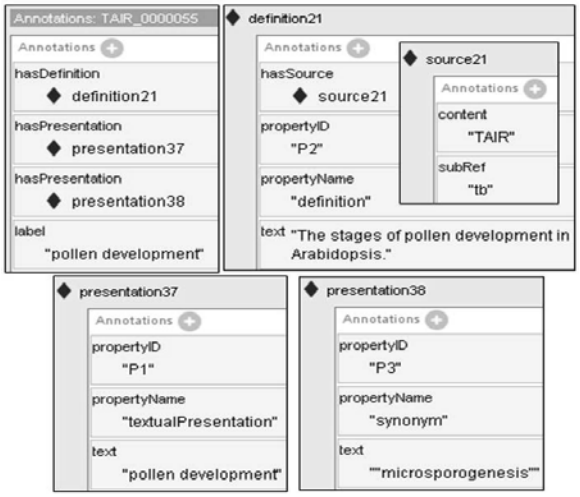
equivalent to the concept "Father". LexOWL can translate it back to OWL as the lower part of Figure 3 shows, which is identical to the original OWL representation.

```
[Term]
id: TAIR:0000055
name: pollen development
def: "The stages of pollen development in Arabidopsis." [TAIR:tb]
synonym: "microsporogenesis" []
is_a: TAIR:0000022 ! body part developmental stages
```

(a) A Sample OBO Term

Coding Scheme: arabidopsis_development.ontology - urn:lsid:bioontology.org:arabidopsis_development.ontology
Concept Code: TAIR:0000055
Entity Description: pollen development
Is Active: true
Presentation: pollen development
  Property Name: textualPresentation
  Property Id: P1
  Is Preferred: true
Presentation: "microsporogenesis"
  Property Name: synonym
  Property Id: P3
  Is Preferred: false
Definition: The stages of pollen development in Arabidopsis.
  Property Name: definition
  Property Id: P2
  Is Preferred: false
  Source: TAIR , **Role:** null, **SubRef:** tb

(b) LexGrid Representation for the Sample Term



(c) LexOWL Representation for the Sample Term

Figure 2: An Example for Entity Mapping

An association in the LexGrid model establishes a relation between two LexGrid entities. LexOWL classifies the LexGrid associations into two types: pre-defined associations and other associations. A pre-defined association can be directly mapped to an OWL element. For example, the associations "subClassOf" (OWL), "CHD" (ICD 10), and "is a" (OBO) are all mapped to *owl*:*subClassOf*. The association "hasSubtype" (UMLS) is mapped as an inverse of OWL element subClassOf. The associations "equivalentClass" (OWL) and "same as" (UMLS) are mapped to *owl*:*equivalentClass*. For detailed information about the pre-defined-association mapping, see https://cabig-kc.nci.nih.gov/Vocab /KC/index.php/ LexGrid_ Documentation[3].

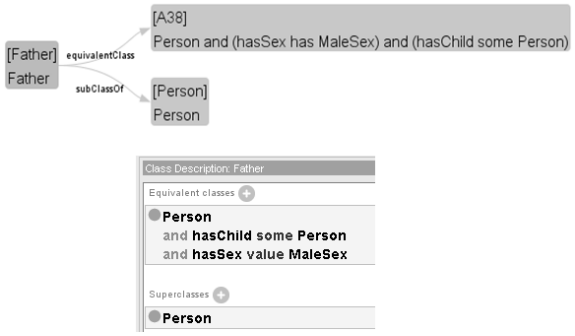


Figure 3: An Example of Anonymous Concept

Evaluation and Discussion

We tested LexOWL using different ontologies from various sources: OWL, OBO, UMLS Semantic Network, and WHO ICD10. We used Protégé Prompt[5] to compare the OWL ontologies generated by LexOWL and by other tools. We also sampled concepts and associations in each test ontology and compared them with the original source and checked whether all the related information are represent properly. The details of the results are listed below.

We tested on 5 OWL files. We chose these 5 ontologies carefully so that they cover most of the OWL Lite syntax introduced in OWL Web Ontology Language Reference[4]. We compared the OWL ontologies generated by LexOWL with the original ontologies. Each pair of ontologies is semantically equivalent to each other.

We also tested on 10 OBO files. For each OBO file, we compared the OWL ontology translated by LexOWL with those converted by OBO2OWL[1], Protégé 3.3.1 OBO to OWL Tab[10], and Protégé 4.0 OBO loader. All the four tools mapped OBO terms to OWL classes, OBO "isa" to OWL *subClassOf*, and used OWL *someValuesFrom* to represent relationships two classes. Semantically, the corresponding ontologies from all the 4 converters are identical. However, each converter defined its own annotation properties and used different annotation properties to represent the same OBO information. OBO2OWL and Protégé 4.0 OBO loader have relatively simple and straightforward conversions where they used the OBO labels directly as the OWL annotation property names. Protégé OBO to OWL Tab and LexOWL processed information in a lower granularity (e.g., the "def" in Figure 2(a) is parsed and the source information is annotated separately.)

We used LexOWL to export UMLS Semantic Network loaded in LexGrid to an OWL file and compared it

with the one converted by Jimenez-Ruiz[6]. LexOWL uses the UIs as the OWL class names versus Jimenez-Ruiz uses the actual names. Hierarchically, these two ontologies are identical. Jimenez-Ruiz introduced some annotation properties that are specific for the UMLS Semantic Network where LexOWL used lexgrid2owl meta-ontology to represent all the information. For example, Jimenez-Ruiz mapped SRDEF to *rdfs:comment*, whereas LexOWL mapped it to *lexgrid2owl:Definition*, which can bring better interoperability since definitions of terms from other sources are also mapped to *lexgrid2owl:Definition*. Jimenez-Ruiz used *owl:allValuesFrom* to represent relationships between two classes and LexOWL used *owl:someValuesFrom* since this is the default restriction LexOWL uses for representing relationships between classes[‡].

We also used LexOWL to export ICD10 WHO second edition loaded in LexGrid to an OWL file and compared it with the OWL file converted by Cardillo, *et al.*[7]. Hierarchically, these two ontologies are identical. The ontology converted by Cardillo, *et al.*[7] only covered hierarchical information, however. Information such as exclusions and inclusions are ignored whereas LexOWL considered them as OWL *ObjectProperties*, thereby preserving the semantics.

In summary, the test results show that LexOWL can convert information modeled in LexGrid to OWL successfully. LexOWL uses a single meta-ontology for all different sources where other tools use different meta-ontologies even for the same format. Hence, the ontologies converted by LexOWL has better Interoperability that will bring benefits in ontology mapping, integration and reasoning in the future.

Concluding Remarks and Future Work

We introduced LexOWL, a system that functionally connects LexGrid to OWL through a bridge over the LexBIG and the OWL APIs. LexOWL can represent information modeled in LexGrid in the OWL API representation, so that tools and services that are developed for OWL can be applied to the biomedical terminologies and ontologies. LexOWL also provides a LexGrid-to-OWL converter with a well-defined interoperability for information from different sources and in different formats.

As for the future work, several directions remain to be pursued. First, we would like to investigate performance of LexOWL with large-sized ontologies such as SNOMED CT, the Gene Ontology, and ICD10. We would like to add the editing and saving function as Figure 1 shows, so that we not only can browse, but also edit information represent in LexGrid using Protégé. Finally, LexOWL serves as a foundational pillar for ontology reasoning and inference. Our next step is to explore toward that direction on biomedical and clinical information.

[‡] How to represent the semantic relationships between classes in a more precise way is a problem we are investigating when mapping information to LexGrid and is out of the scope of this paper.

References

1. OBO2OWL: Lossless transformation between OBO and OWL. http://www.cs.utexas.edu/~hamid/ research/obo2owl.cgi, 2008.
2. LexGrid: The Lexical Grid. https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/LexGrid, 2009.
3. LexGrid to OWL Mapping documentations. https:// cabig-kc.nci.nih.gov/Vocab/KC/index.php/LexGrid_ Documentation, 2009.
4. OWL Web Ontology Language Reference. http://www.w3.org/TR/owl-ref/, 2009.
5. The Protégé Prompt Tab. http://protege.stanford.edu/plugins/prompt/prompt.html, 2009.
6. The UMLS Semantic Network in OWL. http://krono.act.uji.es/people/Ernesto/UMLS_SN_OWL, 2009.
7. Cardillo E, Eccher C, Serafini L and Tamilin A. Logical analysis of mappings between medical classification systems. In *Proceedings of the 13th International Conference on Artificial Intelligence*, pp 311–321, Sep 2008.
8. Fensel D, Sycara K and Mylopoulos J. Representing the UMLS semantic network using OWL. In *Proceedings of the Second International Semantic Web Conference*, pp 1–16, Sanibel Island, Florida, Oct 2003.
9. Golbreic C, Horridge M, Horrocks I, Motik B and Shearer R. OBO and OWL: Leveraging semantic web technologies for the life sciences. In *Proceedings of the 6th International Semantic Web Conference*, pp 169–182, Busan, Korea, Nov 2007.
10. Moreira D and Musen MA. OBO to OWL: A Protégé OWL tab to read/save OBO ontologies. *Bioinformatics*, 23(14):1868–1870, 2007.
11. Noy N, de Coronado S, Solbrig H, Fragoso G, Hartel F and Musen M. Representing the NCI Thesaurus in OWL DL: Modeling tools help modeling languages. *Journal of Applied Ontology*, 3(3):173–190, 2008.
12. Pathak J, Solbrig H, Buntrock J, Johnson T and Chute C. LexGrid: A framework for representing, storing, and querying biomedical terminologies from simple to sublime. *Journal of the American Medical Informatics Association*, 16(9), 2009.

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

Using the Gene Ontology to Annotate Biomedical Journal Articles

Michael Bada, Lawrence Hunter
University of Colorado Denver, Aurora, CO, USA

Abstract

We are creating a gold-standard corpus of manually annotated full-text biomedical journal articles toward natural-language-processing applications. Central to this is our use of entire ontologies of the Open Biomedical Ontologies initiative as well as other terminologies as term sources, in contrast to most other such annotation projects, which have used small, ad hoc schemas. In addition to the standard difficulties in such annotation projects, each of the terminologies we have used has idiosyncrasies and ambiguities that present further challenges to consistent, high-quality annotation of these articles. In this paper we present and discuss the most salient of these with regard to the Gene Ontology that we have encountered and addressed in our annotation guidelines and training. The utility of these guidelines can be seen in the high and still-increasing interannotator-agreement statistics that we continually monitor.

Introduction

Gold-standard annotated biomedical corpora are essential for the training and evaluation of advanced biomedical natural-language-processing (NLP) systems, as evidenced by the significantly improved performance statistics of such systems trained on relevant corpora[1,2]. We have therefore embarked on an ambitious project to create a manually annotated corpus of full-text biomedical journal articles as a gold-standard community resource, which we are calling the CRAFT (Colorado Richly Annotated Full-Text) Corpus. In addition to manually annotating full-text papers, we are using all terms of select ontologies of the Open Biomedical Ontologies (OBO) initiative[3] (as well as other terminologies) as the annotating term sources, the first such effort of which we are aware. One part of this effort is annotation of these articles using the entire Gene Ontology (GO), which is comprised of terms representing biological processes (BP), molecular functions (MF), and cellular components (CC)[4].

In addition to the standard difficulties in this type of annotation project, our effort is particularly challenging due to our annotating full-text articles and using the full term sets of OBOs. Each of the terminologies we have used has its own idiosyncrasies and ambiguities that present further challenges to consistent, high-quality annotation and thus to text mining of these articles. In this paper we present and discuss the most salient of these issues that we have encountered in using the GO and addressed in our annotation guidelines and training. The utility of these guidelines can be seen in the high interannotator-agreement (IAA) statistics that we continually monitor.

Methods

Manual annotation is performed in an annotation tool developed within our lab called Knowtator[5], which is implemented as a plugin to Protege-Frames[6]. In an effort to reduce the annotators' workloads, the articles were preprocessed by automatically tagging them with terms from the relevant ontology; the annotators then check these annotations, making any needed deletions or corrections, and mark up anything else that was missed by the preprocessing. All of the annotators' work is further reviewed by the project lead (MB), with subsequent corrections being made. All tagging is saved as standoff annotation. IAA was calculated by comparing one annotator's set of annotations with the set of annotations created as a result of the project lead's review of the set, and Knowtator was used to calculate IAA statistics. Though the GO is continually evolving and growing, we are using a single static version (dated November 20, 2007, when this project was initiated) that contains 14,306 biological processes, 7,984 molecular functions, and 2,047 cellular components.

Guidelines for Applying the GO toward NLP Annotation

Given that we are using an ontology of more than 24,000 terms to annotate biomedical journal articles and that we strive for high IAA statistics, clear, well-conceived annotation guidelines are critical. Here we present the higher-level issues we have encountered and how we have addressed them in our guidelines. (We also have lower-level, linguistics-based guidelines, but this is outside the scope of this paper.)

1. Rules for General Biological Processes

We have found that general words indicating biological processes are particularly difficult to consistently annotate. For example:

(1) Bicarbonate formation is important for aqueous humor secretion from the ciliary processes and

carbonic anhydrase (CA) facilitates this secretion. [PMID:11532192]

(2) Cells lacking BRCA1/2 fail to form damage-induced subnuclear RAD51 foci with normal efficiency, suggesting that these proteins are required for the formation of recombinase complexes at the sites of DNA damage. [PMID:11597317]

In (1), "formation" is closest to the biological-process term `biosynthetic process` (which encompasses the building up of more complex molecules from simpler ones), while in (2), "formation" is closest to the term `cellular component assembly`, which includes assembly of protein complexes, as is the case here. Rules have helped greatly in the annotation of such general process words. For example, pertaining to the above examples, if lexical variants of "form", "create", "assemble", etc. are applied to molecules, it is likely a biosynthetic process; if it is applied to cellular components, it is likely a cellular-component assembly. Annotating words denoting formation of cells and higher-level anatomical structures consistently is still difficult, as the GO cell and anatomical-structure development terms are arranged in a complicated (and some would say unintuitive) way. We plan on submitting our suggestions to the GO curation team toward making this part of the ontology clearer.

2. Molecular Functions vs. Biological Processes

The GO MF subontology is in a somewhat ambiguous state. The overwhelming majority of the MF terms are defined as molecular-level processes. (For example, the definition of `binding` is the "selective, often stoichiometric, interaction of a molecule with one or more specific sites on another molecule".) Using the terminology of the Basic Formal Ontology (BFO), an upper-level ontology that is being used by members of the OBO Consortium toward evolution of the OBOs[7], a molecular function in this view is an occurrent, *i.e.*, a process, at the molecular level. However, there are those within the OBO Consortium who assert that a GO molecular-function term denotes not a process but a proper function, which is a dependent continuant, *i.e.*, an abstract entity that depends on the existence of another entity – essentially, the potential or functionality inherent in an entity to have a process realized[3]. It appears that the MF subontology will eventually evolve to this latter conceptualization.

This is relevant to our annotation work because this dichotomy between a process and the functionality to effect a process will often be expressed differently in natural-language text and thus should be annotated differently. (However, passages are often ambiguous with regard to process or function.) We have taken the view that molecular functions are processes, partly because this is how they are mostly currently defined and partly because mentions of processes are more common and more straightforwardly indicated in natural language than are functions to effect processes and thus more easily annotated.

A conflated issue is the fact that there are corresponding terms in the BP and MF subontologies that are extremely difficult to differentiate given a textual mention, even using their defintiions, *e.g.*, BP `signal transduction` and MF `signal transducer activity`, BP `regulation of transcription` and MF `transcription regulator activity`, BP `caspase activation` and MF `caspase activator activity`, as well as many corresponding BP transport and MF transporter activity terms. We have mostly dealt with this thus far by using most of the MF terms only to annotate text matching the term itself or an exact synonym (*e.g.*, kinase activity), an acceptably close synonym (*e.g.*, kinase functionality), or the corresponding continuant (*e.g.*, kinase, as discussed later in the paper). This is a suboptimal solution, but using the MF terms in this restricted way has allowed us to maintain our high levels of interannotator agreement.

A merging of the current MF subontology into the BP subontology to create one ontology of occurrents, from molecular-level to organism-level, would go a long way in ameliorating these two interrelated issues. We realize that this may seem radical, but it has been considered before, *e.g.*, at the 2008 NCBO Relation Ontology Expert Meeting[8]. We assert that this should involve merging the corresponding BP & MF term pairs such as those aforementioned. While the BP ontology would be an ontology of biological occurrents, the MF ontology could be redefined as an ontology of biological functions. An ontology of functions could likely be managed semiautomatically, as it would mostly mirror the relevant portions of the process ontology.

3. Noncanonical, Pathological, and *Ex Vivo* Entities and Processes

The GO is charged with representing canonical biological processes, molecular functions, and cellular components, and so we attempted at first to limit annotation to such canonical entities and processes. However, this turns out to be a deceptively difficult task. Sometimes the noncanonicality or pathology is explicit, as in:

(3) There were enlarged extracellular spaces between cells in the equatorial/bow region in 5 wk old alphaA/BKO lenses. [PMID:12546709]

In (3), it is obvious that the extracellular spaces are noncanonical in that they are larger than normal. But many times the noncanonicality or pathology can only be inferred from a very careful reading and comprehension of the article, and many other times it is not at all clear. This is especially due to the fact that most biological articles involve experiments with organisms or components of organisms in which they are subjected to all sorts of procedures, substances, and environments they would not normally encounter. Our solution is to annotate all mentions of GO entities and processes, even those that are explicitly noncanonical or pathological (so long as all of the other rules are followed, of course). Thus, in (3), "extracellular spaces" is annotated with the GO CC term `extracellular space` even though they are noncanonical in terms of their sizes.

A related issue is, given that the GO is in the domain of naturally occurring *in vivo* processes nd cellular entities, whether or not their *ex vivo* counterparts should be annotated. Analogously, to maximize our IAA, we decided to annotate all such *ex vivo* entities and processes; thus, a binding is a binding whether it takes place in an organism or in a beaker.

Smith *et al.* have written of noncanonical anatomical parts, which, in their representation with the Ontology of Biomedical Reality, are siblings of canonical anatomical parts; both of these are subsumed by a superclass of anatomical structures[9]. We analogously are viewing the entities and processes of the GO as these more general concepts that encompass both canonical and noncanonical instances.

4. Verb Nominalizations as Occurrents

Verb nominalizations can refer to either occurrents or continuants; Simon and Smith have written of such duality of certain biomedical terms (*e.g.*, dilation, dislocation) and how they address it in their LinKBase system[10]. As an example from our corpus:

(4) The vesicle formation goes along with several other changes in the red blood cell like cytoskeleton rearrangements and changes in the phospholipid orientation in the cellular membrane. [PMID:12925238]

In (4), "formation" clearly refers to a process, while "orientation" is a dependent continuant in that it is an attribute of the cellular membrane. We instruct our annotators to not annotate relevant mentions of such words if they clearly denote continuants since GO biological processes are occurrents. However, it is sometimes ambiguous whether such a relevant mention refers to either the occurrent or to the dependent continuant, *e.g.*, "distribution" in:

(5) The differentiation and distribution of specific mature neurons was examined in our previous study at adult stages with the expression of striatal markers such as preproenkephalin and Gad65/67. [PMID:15882092]

In such a case where one of the possible readings denotes an occurrent, we instruct the annotator to mark up the mention. If we had an ontology of the corresponding dependent continuants, we would also mark up such a mention with the dependent continuant term, as we encourage multiple annotation to capture the ambiguity of the expression.

Results

At this stage of our annotation of our 97-article corpus, we have created 8,279 annotations of cellular components (which is completed) and 18,996 annotations of molecular functions and biological processes in 44 articles (which is ongoing). (One annotator marked up articles with GO CC terms and another is annotating with GO BP and MF terms.)

To demonstrate the utility of our guidelines, we present the IAAs (calculated approximately weekly) for our two GO annotation passes in Figure 1. The annotation of the cellular components quickly rose to approximately 90% or higher, while the annotation of the biological processes and molecular functions started very low (9.7%) but has significantly risen since, with the last few data points at approximately 80%. There are large oscillations in the graph that are partly due to the fact that this annotator is typically able to annotate only one or two articles per period. A given article often has many mentions of a relatively small set of GO terms, and the IAA statistics are subject to such variation if the two annotators consistently annotate these numerous mentions of this relatively small set of GO terms differently.

5. Continuants with Molecular Functions

Mentions of continuants that have functions that can be realized in processes are often more frequent than the processes themselves; this is especially true for the molecular-function terms. For example, mentions of recombinases are more frequent than mentions of the MF term `recombinase activity`. We wished to capture these lexically analogous mentions, so each such mention is annotated with the corresponding term and also with the class `continuant`, the general term denoting an entity in the BFO. Thus, each such mention is doubly annotated as a process and as an

entity. This is actually not semantically correct, as something cannot be both a continuant and an occurrent according to the BFO since these are disjoint classes in the BFO. It would be better to annotate each such mention once, as a continuant that has the corresponding function, *i.e.*, by annotating as a continuant and then adding a restriction to it, but Knowtator is not currently capable of this type of representation. Nevertheless, our methodology enables us to capture all information that can then be easily transformed into a more semantically correct representation in our annotation repository.

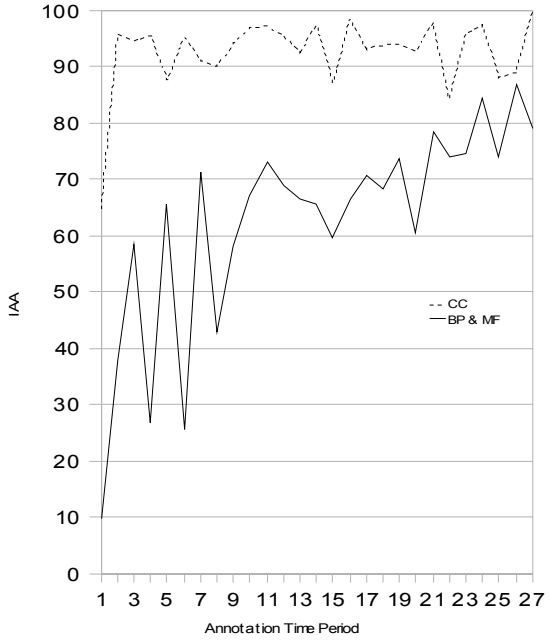


Figure 1. IAA statistics for the GO CC and BP/MF annotation projects over time. Each time period is approximately one week.

Conclusions

We have encountered issues in the course of our project to annotate biomedical journal articles with the whole of the GO, and we have attempted to address them with the guidelines presented in this article. These guidelines have in part allowed us to achieve high IAA statistics using the ontology as a very large annotation schema. We assert that these issues will also be troublesome for attempts at programmatic annotation and text mining of biomedical journal articles, which many see as necessary in the future. We have presented several suggestions to the GO that we believe could ameliorate these issues.

Acknowledgements

This work is supported by NIH 5 T15 LM009451-02 and JDF 110200801921.

References

1. Tsuruoka Y, Tateishi JD, Ohta T, McNaught J, Ananiadou S and Tsujii J. Developing a robust part-of-speech tagger for biomedical text. Proc 10th Panhellenic Conf on Informatics. 2005; 382–392.
2. Lease M and Charniak E. Parsing Biomedical Literature. Natural Language Processing, Springer Berlin/Heidelberg. 2005; 58–69.
3. Smith B, Ashburner M, Rosse C, Bard C, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, The OBI Consortium, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone S-A, Scheuermann RH, Shah N, Whetzel PL and Lewis S. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. Nat Biotech. 2007;25:1251–1255.
4. 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 and Sherlock G. Gene Ontology: Tool for the unification of biology. Nat Genet. 2000;25:25–29.
5. Ogren PV. Knowtator: A plug-in for creating training and evaluation data sets for Biomedical Natural Language systems. Proc 9th Internat Protege Conf. 2006.
6. Gennari JH, Musen MA, Fergerson RW, Grosso WE, Crubezy M, Eriksson H, Noy NF and Tu SW. The Evolution of Protege: An Environment for Knowledge-Based Systems Development. Internat J of Human-Comp Studies. 2003; 58(1):89–123.
7. Grenon P, Smith B and Goldberg L. Biodynamic Ontology: Applying BFO in the Biomedical Domain. Ontologies in Medicine. IOS Press, Amsterdam. 2004;20–38.
8. OntologyRelations, http://bioontology.org/wiki/index.php/
9. Smith B, Kumar A, Ceusters W and Rosse C. On carcinomas and other pathological entities. Comp and Func Genom. 2005;6:379–387.
10. Simon J and Smith B. Using Philosophy to Improve the Coherence and Interoperability of Applications Ontologies: A Field Report on the Collaboration of IFOMIS and L&C. Proc 1st Workshop on Phil and Inform. 2004.

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

A Unified Ontological-Semantic Substrate for Physiological Simulation and Cognitive Modeling

Sergei Nirenburg, Marjorie McShane, Stephen Beale
University of Maryland Baltimore County, Baltimore, MD, USA

Abstract

This paper briefly describes a system that provides a constructive proof of the versatility of ontologies and ontology-based knowledge resources. The Maryland Virtual Patient (MVP) environment models a team of automatic and human agents diagnosing and treating a virtual patient. It uses a uniform ontological substrate to support both the simulation of the world outside the cognitive agents – specifically, physiological processes in the "body" of the virtual patient – and perception, reasoning and action in the "minds" of the virtual patient and other intelligent agents.

Introduction

The term ontology has come to be used so broadly that calling a resource an ontology carries little information without further specification. Our ontology, work on which was started in the early 1980s, was initially intended as a substrate for natural language processing[1]. It has been recently extended to model a society of simulated embodied artificial intelligent agents in the Maryland Virtual Patient (MVP) environment[2,3,4,5] (inter alia).[§] The human agents play the roles of an attending physician and, optionally, a human mentor. The artificial agents play the roles of virtual patients (VPs) and additional members of a medical team, such as lab technicians and medical specialists. The environment also features an automatic tutor agent. The VP agent is at present the most fully developed and the most complex of the artificial agents. A core application of the MVP environment is to help teach medical students cognitive decision-making skills in diagnosing and treating patients. To make this process as close as possible to the experience of treating humans, we have simulated both the patient's "body" and its "mind."

The processes involved in the two kinds of simulation cover the VPs' physiological processes; perception (interoceptive perception and language understanding); reasoning (including decision-making and memory-related operations); and action (physical, verbal and mental). All these processes are supported in the MVP environment by a single set of knowledge resources based on an ontology. In developing the MVP environment, our initial hypothesis was that the above processes could be modeled within a unified, knowledge-based paradigm. Our vested interest in seeing this hypothesis validated was the substantial economies we expected in the knowledge acquisition task. As it happens, the hypothesis was indeed constructively validated in the proof-of-concept MVP system, in which all the above processes have been implemented on the basis of a minimal extension of the OntoSem ontology, which was originally developed to support language understanding.

The Ontology. The OntoSem ontology currently contains about 9,500 concepts – described using, on average, 16 properties each – which are divided among objects, events and properties. Most of the concepts are general-purpose, with the exception of several hundred from the medical domain that were added to support the MVP project. The ontology shares its metalanguage with two other knowledge bases: a lexicon and a language-independent fact repository. There is a many-to-one linking from the lexicon to the ontology, as descriptive specifications of lexical meaning are permitted.[6] OntoSem's metalanguage is unambiguous, which permits reasoning about language and the world to be carried out without the interference of lexical and morphosyntactic ambiguities.

The Fact Repository. The distinction between descriptions and assertions, standard in AI and cognitive modeling, is the criterion for recording a knowledge element as an ontological concept (a description) or an ontological instance (an assertion, stored in the fact repository). This distinction proves useful in modeling all the processes necessary for supporting the MVP environment. For example, the preferred mode of modeling language understanding in our approach is to use the OntoSem analyzer to generate disambiguated text-meaning representations (TMRs) from input texts, store the TMRs in the fact repository, then use the fact repository as a source of heuristics for all further processing, including subsequent language understanding itself. In other words, the fact repository both helps the processing of new texts and is augmented by semantic information obtained from those texts.

A non-linguistic example of the use of the ontology vs. fact repository distinction is authoring libraries of specific MVPs on the basis of a "prototype" disease stored in the ontology.

[§] Patent pending.

Artificial Agent Capabilities in MVP

Several types of processes in the MVP environment are supported by ontological representations of typical sequences of causally and temporally connected events, often referred to as causal chains, scripts or plans.[7] These processes include physiological simulation, the reasoning used for language analysis, and the reasoning used for decision-making. In this section we present brief examples in each of these domains to show how the ontology seamlessly ties together a complex multi-agent system.

Physiological Simulation. Physiological simulation in the MVP environment is implemented using ontological representations of complex events. As an example, consider the representation of the complex event SWALLOW. The SWALLOW script includes many subevents (muscles contracting, nerves firing), conditionals (food cannot pass if there is an obstruction), loops (peristalsis throughout the segments of the esophagus), and so on, this script is conceptually straightforward in that every event must occur in the order specified, given that its preconditions are met, with no optional or variously ordered events. (For lack of space, we show only a few of the dozens of subevents, omit variable bindings, local properties and property facet markers.)

```
(swallow
  (has-event-as-part
    oropharyngeal-phase-of-swallowing
    esophageal-phase-of-swallowing))
(oropharyngeal-phase-of-swallowing
  (has-event-as-part
    motion-event:mouth_to_pharynx
    contract-muscle:contract_pharynx
    motion-event:pharynx_to_larynx
    relax-muscle:crico_relaxes
    relax-muscle:LES_relaxes))
(esophageal-phase-of-swallowing
  (has-event-as-part
    peristalsis:from_larynx
    contract-muscle:crico
    peristalsis:R  ; Regular peristalsis in the esophagus
    peristalsis:to_stomach))
(motion-event:mouth_to_pharynx
  (agent              human-a)
  (theme              bolus-a)
  (instrument         human-a.tongue)
  (source             human-a.mouth)
  (destination        human-a.pharynx)
    (duration         (value 0.08)(default-measure second))
    (effect
    (location (domain    bolus-a)
              (range     human-a.pharynx))))
(contract-muscle:contract_pharynx
 (agent  human-a))
 (theme (set (element human-a.pharynx-constrictor-muscle)
             (cardinality >1))))
 (effect  (openness
             (domain  human-a.pharynx. epiglottis)
             (range    0)))
…)
```

A more complex type of physiological simulation-supporting script is a disease script, which not only has more parameterizable features but can also be modified midstream by external factors, like medical interventions or changes in the person's lifestyle.[2,3]

Cognitive Capabilities. Viewed in a simplified manner, the cognitive capabilities of a VP are implemented as an infinite perception– decision-making–action loop. In the MVP environment, the world that is perceived by the VP is constrained to its own body (interoception) and to its language-based interactions with the agents playing the roles of medical personnel. The VP's reasoning covers not only goal-oriented decision making, it is also central to language analysis and generation. The VP's actions include dialog-related verbal actions, manipulating the agenda of goals and plans, remembering events and facts, and a few physical events – such as presenting to the MD – that are not simulated in great detail at the moment.

Modeling Perception I: Interoception. Interoception connects the "body" and the "mind" of the VP by signaling the agent's becoming aware of a symptom (e.g., pain), understood as a side effect of its physiological state. Procedurally, the moment the VP perceives a symptom, the latter is added to its short-term memory. This triggers the addition of an instance of the goal *be-healthy* onto the agenda, with the symptom as a parameter. What is important for this paper is that the ontologically grounded format in which symptoms are formulated is identical to that of text meaning representations (produced by the language analysis system) and elements of the fact repository.

Modeling Perception II: Language. Many aspects of language processing – from disambiguation[1] to paraphrase detection[5] to reference resolution[8] – can be supported by knowledge like that provided by the OntoSem ontology. For lack of space, we briefly discuss just one aspect of the OntoSem ontology – multivalued selectional restrictions – and two of the many types of language processing it permits.

(i) *Resolution of type incongruity.* Type incongruities are situations in which typical semantic constraints are not met: dogs can eat newspapers, even though the THEME case role of the ontological concept INGEST should be constrained to food or drink; parrots can speak, even though humans are the only full-fledged agents of speaking; babies and dogs can earn money (e.g., as clothing models or in pet food ads), even though they are hardly typical agentive workers.

In cases where extensions of meaning can be foreseen, they can be encoded using multivalued selectional restrictions. In OntoSem, these are implemented using *facets* of property values, which reflect the different confidence levels in semantic decisions. Thus, the INSTRUMENT case role of the event PAY can be constrained to MONEY on the DEFAULT facet, but also license GOODS on the SEM facet and SERVICE on the RELAXABLE-TO facet. Of course, many such extended meanings cannot be anticipated and must be processed using runtime reasoning, which is an ongoing line of work in OntoSem.

(ii) *Reference resolution.* Among the most difficult aspects of reference resolution is selecting the most appropriate antecedent from among a list of candidates when the standard non-semantic heuristics fail to come up with a strong preference. Multivalued selectional restrictions can sometimes cast a deciding vote. For example, whereas the typical agent of a surgical procedure is a surgeon, any doctor – or, in a pinch, any person – can perform some types of surgery. If a text included a sentence like *He botched the surgery*, and if there were several potential antecedents for *he* that had similar non-semantic scores, any candidate known to be a surgeon should be preferred; barring that, anybody known to be a physician, though not known to be a surgeon, should be preferred; and barring this, any human, not known to be either a physician or a surgeon, should be preferred (this example is simplified to save space; the actual combination of heuristic evidence is much more complex). The relevant ontological concept (once again, simplified) is:

PERFORM-SURGERY
AGENT DEFAULT SURGEON
SEM PHYSICIAN
RELAXABLE-TO HUMAN

Whereas the facets SEM, DEFAULT and RELAXABLE-TO are used in the ontology, the fact repository uses the facet VALUE whose semantics is that of actuality, not typicality.

Modeling the VP's Decision Making. When making decisions, the VP uses both knowledge it is aware of and knowledge that it might not be expressly aware of. The kinds of conscious knowledge that the VP uses for making decisions are: (a) an inventory of ontologically grounded goals and an inventory of plans that the VP knows are instrumental in attaining a particular goal; (b) information about the VP's physiological state, particularly the intensity and frequency of symptoms, as perceived via interoception and remembered in its memory; (c) information available to the VP about certain properties of tests and treatments for its condition: pain, unpleasantness, risk and effectiveness; if this information is not available to the VP, the VP has the option of activating a plan of determining the values for these parameters; in the current implementation, this involves asking questions of the agent playing the role of attending physician; and (d) two time-related parameters: the follow-up-date, i.e., the time the doctor told the patient to come for a follow-up, and the current-time of the given interaction. The largely subconscious traits the VP uses in decision-making are: (a) character traits like trust, suggestibility and courage; and (b) certain physiological traits, like physiological-resistance (e.g., how well the MVP tolerates chemotherapy), pain-threshold (how much pain the MVP can tolerate) and the ability-to-tolerate-symptoms (how intense or frequent symptoms have to be before the MVP feels the need to do something about them).[4]

Discussion

Much recent work on ontology has been devoted to compiling "ontologies" – under any definition of the word – as quickly and inexpensively as possible. Most such efforts exploit machine learning techniques and, as would be expected, produce noisy results that are useful for some applications but will certainly not support simulation or high-level reasoning by an advanced, conversational intelligent agent. In this paper we have attempted to show that keeping human acquirers (largely) out of the loop is not the only way to keep ontology development from being prohibitively expensive. Another way is to manually or semi-automatically create resources but reuse them across modules of an environment. In the case of MVP, the physiological, general cognitive and language processing capabilities of all the agents rely on the same ontological substrate, the same organization of the fact repository (agent memory) and the same format of knowledge representation. This uniformity not only provides significant savings in development, testing and debugging time, it also facilitates interoperability. The MVP system provides a constructive proof of the versatility of ontologies and ontology-based knowledge resources.

Naturally, when starting to develop our first medical application we sought domain-specific ontologies that might be incorporated into our general purpose ontology. Two large and well-known resources are MeSH, the National Library of Medicine's (NLM's) tree of medical subject headings, arranged hierarchically, and Metathesaurus, NLM's ontology of hundreds of thousands of medical terms along with their synonyms and morphological variants (hereafter referred to together as M/M). These resources overlap in part (MeSH being much smaller) and use the same concept identifiers (CUIs).[9] After experimentation with these resources – which reflects

the best understanding of them we could acquire in a limited time – we concluded that importation would not benefit our system for the following reasons: (a) M/M is geared toward the needs of library science, not having the semantic precision to support high-level reasoning by artificial agents; (b) the content is English terms, including all synonyms, which introduces the language issues that our language-independent ontology avoids; (c) there are very few properties, and 61% of properties (at the time of our experiment, in 2005) had no properties at all; (d) there is no division of concepts and instances; (e) the is-a relation is not interpreted as strictly in M/M as in OntoSem: e.g., in the following are all siblings: Gait; Lower extremity pain walking; Lower limb length difference; Barefoot walking; and Extensor thrust[10]; (f) many concepts in M/M contain a very large set of parents – i.e., 651,000 have one or two parents but another 30,000 have 3 or more parents, with the following reckoning (number of concepts: number of parents): 17075:**3**, 6787:**4**, 3434:**5**, 1907:**6**, 1203:**7**, 715:**8**, 432:**9**, 1,000:**>=10**. As mentioned earlier, many of these "parents" are not parents in the narrow sense of the term used in OntoSem but, instead, concepts related in some unspecified way: e.g., of the 38 root nodes of the hierarchy, a number are sources of information, like SNOMED Intl. 1998 and Medical Entities Dictionary; (g) the physicians collaborating in the work found the content too noisy to be helpful; and, as would be expected of any large resource, (g) there are many errors that would need to be cleaned manually to keep our ontology to its current standard: e.g., over 14,000 concepts are parents of themselves. In terms of utility to our ontology, the UMLS resources have a similar status as WordNet[11] has for building our lexicon: acquirers can use them to provide ideas for resource development, but no automatic, full-blown incorporation can be usefully carried out.

There is, however, one resource that has been very useful in our work: the Foundational Model of Anatomy (FMA)[12]. FMA provides both inheritance (is-a) and meronymic (part-of) trees for elements of human anatomy, as well as a number of other properties like distal to/proximal to, has mass, and so on. The names of concepts are English terms. Synonyms and some foreign language equivalents are included, but they are linked to the "preferred term," making this truly an ontology rather than a word net. In supplementing the OntoSem ontology for use in the medical domain, we are consulting the FMA model because, first, it represents a fine organization of anatomical concepts and, second, we aim to keep our knowledge resources compatible with what we believe will become the accepted standard. However, it would be incorrect to assume that FMA answers all our needs in the medical domain: it treats only anatomical objects, whereas we need a full treatment of relevant events and their relationship to objects, both anatomical and extra-anatomical. In addition, as might be expected, our collaborating doctors do not agree with all of the decisions of the FMA developers with respect to the specific needs of our environment.

References

1. Nirenburg S and Raskin V. Ontological Semantics. MIT Press, 2004.
2. McShane M, Fantry G, Beale S, Nirenburg S and Jarrell B. Disease interaction in cognitive simulations for medical training. Proceedings of MODSIM World Conference, Medical Track, 2007, Virginia Beach, Sept. 11-13 2007.
3. McShane M, Nirenburg S, Beale S, Jarrell B and Fantry G. 2007. Knowledge-based modeling and simulation of diseases with highly differentiated clinical manifestations. 11th Conference on Artificial Intelligence in Medicine (AIME '07), Amsterdam, The Netherlands, July 7-11, 2007.
4. Nirenburg S, McShane M and Beale S. A simulated physiological/cognitive "double agent." AAAI 2008 Fall Symposium on Biologically Inspired Cognitive Architectures, 2008.
5. McShane M, Nirenburg S and Beale S. Two kinds of paraphrase in modeling embodied cognitive agents. AAAI 2008 Fall Symposium on Naturally Inspired AI, 2008.
6. McShane M, Nirenburg S and Beale S. An NLP Lexicon as a Largely Language Independent Resource, Machine Translation, 19(2): 139–173, 2005.
7. Schank R and Abelson R. Scripts, Plans, Goals and Understanding, Erlbaum Assoc., Hillsdale, N.J., 1977.
8. McShane M. 2005. A Theory of Ellipsis. Oxford University Press.
9. Bodenreider O. The Unified Medical Language System (UMLS): Integrating biomedical terminology. Nucleic Acids Research 32: 267–270, 2004.
10. Woods W. What's in a link: Foundations for semantic networks. In Bobrow DG and Collins A (eds.), Representation and Understanding, Academic Press, 1975.
11. Fellbaum C. A semantic network of English: The mother of all wordnets. Computers and the Humanities 32: 209–220, 1999.
12. Rosse C and Mejino JLV. A reference ontology for bioinformatics: The Foundational Model of Anatomy. Journal of Biomedical Informatics, 2004.

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

Using Ontology Fingerprints to Evaluate Genome-Wide Association Study Results

Lam C. Tsoi[1], Michael Boehnke[2], Richard L. Klein[3,4], W. Jim Zheng[1]

[1]Department of Biostatistics, Bioinformatics & Epidemiology, Medical University of South Carolina, Charleston, SC, USA; [2]Department of Biostatistics and Center for Statistical Genetics, School of Public Health, University of Michigan, Ann Arbor, MI, USA; [3]Division of Endocrinology, Metabolism, and Medical Genetics, Department of Medicine, Medical University of South Carolina, Charleston, SC, USA; [4]Research Service, Ralph H. Johnson Department of Veterans Affairs Medical Center, Charleston, SC, USA

Abstract

We describe an approach to characterize genes or phenotypes via ontology fingerprints which are composed of Gene Ontology (GO) terms overrepresented among those PubMed abstracts linked to the genes or phenotypes. We then quantify the biological relevance between genes and phenotypes by comparing their ontology fingerprints to calculate a similarity score. We validated this approach by correctly identifying genes belong to their biological pathways with high accuracy, and applied this approach to evaluate GWA study by ranking genes associated with the lipid concentrations in plasma as well as to prioritize genes within linkage disequilibrium (LD) block. We found that the genes with highest scores were: ABCA1, LPL, and CETP for HDL; LDLR, APOE and APOB for LDL; and LPL, APOA1 and APOB for triglyceride. In addition, we identified some top ranked genes linking to lipid metabolism from the literature even in cases where such knowledge was not reflected in current annotation of these genes. These results demonstrate that ontology fingerprints can be used effectively to prioritize genes from GWA studies for experimental validation.

Introduction

Genome-wide association (GWA) studies have become a feasible and important method to identify loci that are associated with a particular phenotype[1]. Assessing quantitatively the likely importance of genes identified as significant to disease risk based on biological facts is essential to proceed efficiently toward experimental validation processes and, ultimately, to define the causal relationships between genes and phenotypes.

Various text-mining methods have been developed to extract information from the biomedical literature for gene annotation[2-5]. In addition, GO provides a standardized characterization of gene functions[6]. Despite the fact that biomedical literatures were written without GO in mind, it has been shown that GO terms that can be identified in PubMed abstracts tend to occur frequently in the literature[7]. Therefore, GO as a standardized terminology provides a semantic grounding to mine the PubMed literature.

Here we describe a comprehensive analysis combining text mining of PubMed abstracts and GO with quantitative measure to assemble ontology fingerprints for genes and phenotypes, and a method to calculate a similarity score between two ontology fingerprints. We further describe how comparing the ontology fingerprints of a phenotype with that of genes identified in a GWA study can be used to prioritize genes for follow-up investigation, including fine mapping and functional studies.

Methods

Data

We used the June 13, 2007 version of GO and 2007 version of PubMed abstracts for this study. The PubMed abstracts and the genes annotated were obtained from the NCBI "pubmed2gene" file. Abstracts that contained GO terms were also annotated by mapping each term to the abstracts using exact string match. Since GO is a Directed Acyclic Graph (DAG)[6], abstracts containing a GO term were also labeled with all the parents of that GO term in the GO hierarchy as well. In addition, each abstract was labeled with a GO term only once regardless of how many times the term occurred. Because we were attempting to decipher human gene-phenotype relationships, the ontology fingerprints were derived from abstracts linked to human genes. In total, we retrieved 178,687 abstracts, and we constructed ontology fingerprints for all 25,357 human genes. There were 5,001 ontology terms that mapped to PubMed abstracts linked to human genes.

Enrichment Test

To test whether a GO term appeared more often in PubMed abstracts linked to a gene than in the rest of the PubMed abstracts linked to other human genes, we performed a hypergeometric test, resulting in a list of GO terms with p-values for each gene. Due to

the discreteness of the hypergeometric distribution, the mid p-value was used in the calculation[8]:

$$p - value = \frac{1}{2}P(A_{obs} = e) + P(A_{obs} > e) \qquad \text{Equation 1}$$

For each gene and ontology pair, A_T is the total number of abstracts considered, while A_O and A_G denotes the number of abstracts linked to the ontology term and gene respectively; number of abstracts that linked to both the ontology term and the gene is labeled as e. A_{obs} is the random variable of observing the number of abstracts linked to both the ontology term and the gene. The p-value was then adjusted to remove insignificant GO terms (See Supplementary information for details).

We also performed the same test on each phenotype-ontology pair. While each gene or phenotype has a list of ontology terms serving as ontology fingerprints defined as ontology terms with p-value <1, collectively the terms and the quantification reflect the characteristics of the gene or phenotype.

Similarity Score Calculation

The ontology fingerprint characterizes the cellular component, molecular function, or biological process of a gene or a phenotype with a quantitative measure. By comparing how similar the ontology fingerprints between a gene and a phenotype are, we can infer to what extent a gene may be related to the phenotype. We calculate a similarity score using a modified version of the inner product:

$$S_j = \frac{\sum_{i=1}^{O} log(q_i) log(r_{ij})}{max\left\{1, \sum_{i=1}^{O} [I(q_i < 1) I(r_{ij} = 1)]\right\}} \qquad \text{Equation2}$$

$i = 1, 2, ..., O$ represents the ontology terms, and the r_{ij} and q_i represent the adjusted p-values of the i^{th} ontology term of the gene j and the phenotype term, respectively. We took the logarithm of the probabilities to prevent underflow during computation. In the numerator, ontology terms that have adjusted p-values equal 1.0 for either the gene or phenotype (i.e. not in either of the gene's or phenotype's fingerprint) will have a score of zero for that ontology term i, and thus make no contribution. Each similarity score is then normalized by $\sum_{i=1}^{O} I(q_i < 1) I(r_{ij} = 1)$, which is the number of ontology terms in the fingerprint of the phenotype but not in that of gene j. The normalization intends to give more weight on a gene's ontology fingerprint that has a higher degree of overlapping terms with the phenotype's ontology fingerprint. If all of the ontology terms of a phenotype overlap with those of a gene, 1 is used in the denominator. Note from Equation 2 that an ontology term with low adjusted p-values for both the phenotype and the gene would contribute significantly to the similarity score. Therefore, the equation considers both the number of GO terms in the ontology fingerprints and the significance level indicated by the p-value. A p-value threshold (λ) was selected and applied to calculate similarity score between genes and phenotypes (See Supplementary information for detail).

Significant Genes Identified from GWA Study

We applied our approach to a GWA study that investigated the influences of loci on lipid concentrations, HDL, LDL, and triglyceride[9]. Genes within or overlap with the top linkage disequilibrium (LD) blocks of best SNPs for each trait were obtained as significantly associated with the corresponding trait (top 199, 201 and 200 LD blocks for LDL, HDL and TG respectively). Independent loci were defined as having low correlation ($r^2 < 0.2$) with any other higher ranking SNP. The p-value of the most significant SNP within each block was used.

Results

Ontology Fingerprints

We computed the association of genes or phenotypes with GO terms by using the hypergeometric enrichment test. The p-values from the test (raw p-values) were then adjusted, taking into consideration the number of ontology terms associated with the genes or phenotypes. The purpose of the adjustment was to reduce the impact of insignificant ontology terms on the ontology fingerprints of genes or phenotypes that have been extensively studied. The resulting ontology terms with adjusted p-values collectively served as the ontology fingerprint for the gene or phenotype, with the p-value for each term reflecting the significance of the term's enrichment among the abstracts associated with the gene or phenotype. Only terms with adjusted p-values < 1.0 were used to define the ontology fingerprints for the gene or phenotype. Table 1 illustrates a small portion of the ontology fingerprint for the gene *VEGFA*, which encodes vascular endothelial growth factor A. This ontology fingerprint serves as a comprehensive, quantitative characterization of the gene using well-defined ontology terms.

GO id	GO term	Adjusted p-value
GO#GO_0008083	Growth Factor	1.00×10^{-323}
GO#GO_0001525	Angiogenesis	1.00×10^{-323}
…	…	…
GO#GO_0008283	Cell Proliferation	1.52×10^{-6}
GO#GO_0006928	Cell Motility	1.71×10^{-6}
…	…	…
GO#GO_0004714	Transmembrane Receptor Protein Tyrosine Kinase	2.60×10^{-1}
GO#GO_0002253	Activation of Immune Response	2.64×10^{-1}
…	…	…
GO#GO_0042098	T Cell Proliferation	9.35×10^{-1}
GO#GO_0003773	Heat Shock Protein	9.58×10^{-1}
…	…	…

Table 1. Eight out of the 279 GO terms in the ontology fingerprint for *VEGFA*. Full list is shown in Supplementary Table 1.

Similarity Scores between Genes and Phenotypes

By comparing the genes' and phenotypes' ontology fingerprints, we calculated similarity scores to quantify the relevance of particular genes to phenotypes. We tested our approach by using 10 randomly selected KEGG pathways as phenotype domains for evaluation. The AUCs for the 10 pathways are shown in Table 2 (column "Ontology Fingerprint AUC"). We compared our approach to a similar text-mining approach which uses "concept profiles" to evaluate the association between different biological concepts[10]. Table 2 shows how well the ontology fingerprint approach and this Anni 2.0 system correctly associated genes with their corresponding KEGG pathways. Specifically, our ontology fingerprint-based method has higher AUC for associating genes with their corresponding pathways than Anni 2.0. 1. We attribute such significant improvement to the employment of Gene Ontology, a well-developed controlled vocabulary to characterize the biological features of genes and phenotypes, the hypergeometric test, which highly increases the sensitivity for detecting the associated ontology terms, and our scoring method, which emphasizes on the number of ontology terms characterizing both the gene and the phenotype.

Using Ontology Fingerprints to Prioritize Genes from GWA Studies

We applied our method to evaluate the results from a GWA analysis[9] studying the genetic variants influencing plasma lipid concentrations, including High-density lipoprotein (HDL), Low-density lipoprotein (LDL), and Triglyceride (TG). Among genes strong associations with lipid concentration, many are not clearly identified in their annotation as being relevant to lipid metabolism. Within the top-ranked genes are quite a few well-known cholesterol related genes, including cholesterol ester transfer protein, plasma (*CETP*), low density lipoprotein receptor (*LDLR*), lipoprotein lipase (*LPL*). Simply based on the gene annotations alone, there are 10, 8, and 12 genes related to the lipid mechanism among the top 20 genes with highest similarity scores. For the remaining genes that do not have Entrez Gene annotation to be associated with the lipid metabolism, we found that there are additional 3, 9 and 7 genes that could potentially influence the HDL, LDL and TG concentrations respectively by tracing back to the GO terms and the literatures that contributed to the similarity scores. One example is transferrin (*TF*), which is ranked by the similarity score among the top 20 genes for HDL. While current annotation of *TF* does not show any relevance to lipid or lipid metabolism, we found that Cubilin (*CUBN*), an endocytic receptor, can act as a receptor for both transferrin and apolipoprotein A1[11]. Another example is thyroid hormone receptor beta (*THRB*). *THRB* was found to negatively regulate the lipoprotein lipase inhibitor[12], and the agonist of *THRB* is associated with a decrease of triglyceride concentration in rats[13,14]. Neither the relationship of *THRB* to nor its influence on the concentration of triglycerides in humans is established, so the annotation for this gene shows no direct link to lipid metabolism. Our results indicate that the ontology fingerprint method can identify genes relevant to the phenotypes revealed through GWA study (The top 20 ranked genes are listed in supplementary Table 2).

Pathway	Ontology Fingerprint AUC	Anni 2.0 AUC	p-value from Wilcoxon Test
Apoptosis	0.96	0.85*	5.56×10^{-19}
Biosynthesis of steroids	0.75	0.73	0.66
Fatty acid metabolism	0.88	0.86	0.14
Focal Adhesion	0.94	0.87*	4.06×10^{-11}
Galactose metabolism	0.90	0.78*	7.64×10^{-9}
Glycolysis	0.80	0.72*	1.86×10^{-6}
MAP kinase signaling	0.90	0.78*	2.21×10^{-14}
Prostate cancer	0.95	0.91*	3.80×10^{-8}
Renal cell carcinoma	0.93	0.81*	1.65×10^{-12}
Sphingolipid metabolism	0.89	0.72*	2.09×10^{-9}

Table 2. Ontology Fingerprints-derived similarity scores can correctly assign genes to their corresponding pathways. The area under ROC curves for each of 10 KEGG pathways are shown. The middle column shows the results from the Ontology fingerprint method, while the right column is the result from the Anni 2.0; * represents the difference between the two methods is significant at 0.0001 level by the Wilcoxon rank-sum test.

Conclusion

Even though several text mining approaches have been developed to identify relationships between genes and phenotypes, our approach is significantly different in several aspects: 1) a hypergeometric enrichment test was used to focus on identifying overrepresented ontology terms for genes and phenotypes in relevant PubMed abstracts; 2) ontology fingerprints with quantitative measures, rather than individual ontology term annotations, were used to capture comprehensive characteristics of genes and phenotypes; 3) a method to calculate similarity scores between ontology fingerprints evaluated the relevance between genes and phenotypes.

*The Supplementary information can be found at:
http://genomebioinfo.musc.edu/OntoFinger/

Acknowledgement

This work is supported by grants IRG 97-219-08 from the ACS, a pilot project of Grant 5 P20 RR017696-05 and PhRMA Foundation Research Starter Grant (WJZ), DK62370 and HG00376 (MB), and NLM training grant 5-T15-LM007438-02 (LCT).

References

1. Thomas DC, *et al*. Recent development in genomewide association scans: A workshop summary and review. *American Journal of Human Genetics* 2005, 77:337–345.
2. Freudenberg J and Propping P. A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics* 2002, 18 Suppl 2:S110–115.
3. Turner FS, Clutterbuck DR and Semple CA. POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol* 2003, 4(11):R75.
4. Baral C, Gonzalez G, Gitter A, Teegarden C, Zeigler A and Joshi-Tope G. CBioC: Beyond a prototype for collaborative annotation of molecular interactions from the literature. *Comput Syst Bioinformatics Conf* 2007, 6:381–384.
5. Cheng D, Knox C, Young N, Stothard P, Damaraju S and Wishart DS. PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res* 2008, 36(Web Server issue):W399–405.
6. The Gene Ontology Consortium: The Gene Ontology project in 2008. *Nucleic Acids Res* 2008, 36(Database issue):D440–444.
7. Verspoor CM, Joslyn C and Papcun GJ. The Gene Ontology as a Source of Lexical Semantic Knowledge for a Biological Natural Language Processing Application. In: *SIGIR'03 Workshop on Text Analysis and Search for Bioinformatics: August, 1st 2003; Toronto, CA*; 2003: 51–56.
8. Agresti A. Categorical Data Analysis, 2nd ed. Hoboken, New Jersey: John Wiley & Sons, Inc.; 2002.
9. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, Heath SC, Timpson NJ, Najjar SS, Stringham HM, *et al*. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet* 2008, 40(2):161–169.
10. Jelier R, Schuemie MJ, Veldhoven A, Dorssers LC, Jenster G and Kors JA. Anni 2.0: A multipurpose text-mining tool for the life sciences. *Genome Biol* 2008, 9(6):R96.
11. Kozyraki R, Fyfe J, Verroust PJ, Jacobsen C, Dautry-Varsat A, Gburek J, Willnow TE, Christensen EI and Moestrup SK. Megalin-dependent cubilin-mediated endocytosis is a major pathway for the apical uptake of transferrin in polarized epithelia. *Proc Natl Acad Sci U S A* 2001, 98(22):12491–12496.
12. Fugier C, Tousaint JJ, Prieur X, Plateroti M, Samarut J and Delerive P. The lipoprotein lipase inhibitor ANGPTL3 is negatively regulated by thyroid hormone. *J Biol Chem* 2006, 281(17):11553–11559.
13. Prieur X, Huby T, Coste H, Schaap FG, Chapman MJ and Rodriguez JC. Thyroid hormone regulates the hypotriglyceridemic gene APOA5. *J Biol Chem* 2005, 280(30):27533–27543.
14. Erion MD, Cable EE, Ito BR, Jiang H, Fujitaki JM, Finn PD, Zhang BH, Hou J, Boyer SH, van Poelje PD, *et al*. Targeting thyroid hormone receptor-beta agonists to the liver reduces cholesterol and triglycerides and improves the therapeutic index. *Proc Natl Acad Sci USA* 2007, 104(39):15490–15495.

Practical Experiences in Concurrent, Collaborative Ontology Building Using Collaborative Protégé

Daniel Schober[1,2], James Malone[2], Robert Stevens[3]
[1]**Institute for Medical Biometry and Medical Informatics, University Clinic, Freiburg, Germany**
[2]**European Bioinformatics Institute, Cambridge, UK**
[3]**Manchester School of Computer Sciences, Manchester, UK**

Abstract

Creation of an ontology according to some common plan is best accomplished collaboratively. This is sometimes contradicted by the distribution of the ontology's developers. An obvious solution therefore is to build collaboration into ontology development tools. Such support necessarily includes both the technical means to perform editing operations upon an ontology, but also support for the communication that makes collaboration such a vital part of much ontology development. To investigate the distributed, collaborative ontology engineering process and the corresponding capabilities of the Collaborative Protege 3 (CP) tool, members of the OntoGenesis network came together and enriched the Ontology of Biomedical Investigations (OBI) with new content. The communications and interactions of the participants with each other, directly or through the tool, were tracked and analyzed. Our initial analysis of the degree to which this new tool fulfills the practical requirements of collaborative ontology engineering suggests the approach is promising. We present some observations and recommendations for CP based upon this experience.

Introduction

Engineering ontologies that are representative of a community consensus is of great interest to those working in bioinformatics and often requires close collaboration. Yet, the process of developing such ontologies often requires collaboration by many people in distributed geographical regions. There are a number of important requirements for ontology development tools that cope with the contradiction of the need for close collaboration and the distribution of developers[1]. Firstly, concurrent ontology editing, the ability for multiple edits to be made to the ontology at a single time and from different computers. Secondly, tracking annotations (called 'notes' in CP) associated with corresponding representational units (RUs). Thirdly, tracking annotations associated with actions of ontology change, such as deletions, axiom edits and annotation edits. Fourthly, a manageable mechanism for discussion threads and instant messaging for online editors that satisfy the need for communication between ontology developers that makes collaborative ontology building so useful.

The new Collaborative Protégé (CP) plugin[2] for the widely used open-source ontology editing tool Protégé 3, claims to support the above features. CP enables concurrent editing of a single OWL file. The tool also features notes on RUs, a change tracking log for RUs (such as class edits), a discussion thread and an instant messaging client for real time chat. The tool captures changes, notes and discussions as instances of an integrated *Change and Annotation Ontology (ChAO)*, thereby providing an audit trail of edits and decision making. This tool, therefore appears an appropriate choice for an evaluation of collaborative ontology engineering and we present an initial investigation into its use.

Materials and Methods

Thirteen members of the OntoGenesis Network came together at the European Bioinformatics Institute (EBI) for the 7th OntoGenesis Meeting (website: http://ontogenesis.ontonet.org/moin/NetworkMeeting 7). The instrument branch of the ontology of biomedical investigation (OBI, http://obi.sourceforge.net), an OWL-DL ontology for the annotation of the biomedical laboratory workflow, was enriched with new classes and relations needed to describe instruments. The instrument branch was chosen because it represents an area of daily experience upon which a broad range of biologists, such as is present in the OntoGenesis Network, have something valid to contribute. The Obi.owl file was populated with new device classes and functions a) coming from the domains of the OntoGenesis members and b) as taken from a list provided by the Metabolomics Standard Initiative (http://msi-ontology.sourceforge.net/). The development followed the methodology adopted by the OBI developers http://obi.sourceforge.net/ontologyInformation/index.php#designPrinciples).

Our methodology involved the following set of tasks:

Familiarization: Users had an initial familiarization with Collaborative Protégé 3.4, its GUI and collaborative features.

Ad hoc additions: Development of attendee's own lists of devices and concomitant functions. This essentially required the addition of new classes as children of the OBI device class and the OBI function class. This also meant that there was a possibility of duplication, i.e. addition of the same device by 2 different editors, as the edits were made concurrently.

Controlled additions: Placement of selected device classes from the MSI term list into OBI. The appropriate metadata required by OBI were also added.

'Agent Provocateur': During a specified time period known only to organizer, an *Agent Provocateur* added conflicting and deliberately incorrect content to the ontology. This was used to assess the transparency of the changes occurring to the other online editors.

Controlled Communication: Communication was restricted to specified channels during each editing session in order to evaluate CPs ability to foster communication, i.e. via notes, discussion threads and chat.

Initially, development occurred in a single group but was then divided into two groups. *Ad hoc* additions were made, where editors were able to add and edit classes as they saw appropriate. Participants were then further divided in 4 pairs of 2, which then tackled different subsets of the MSI device term list. Each pair picked new terms from the list and added them to OBI with discussion. Then the results of the pairs were reviewed and commented by other pairs adding annotations. After more MSI terms were added by the whole group, first the chat was used to comment, annotate and discuss these additions. Then they were discussed by voice only and after that by chat and voice together. During the latter stages of this session, the *Agent Provocateur* user was deployed.

Results

Editing the Ontology

The complete editing metrics, together with tables, diagrams and deeper discussions, can be found in the supplementary material accessible from the 7th OntoGenesis Meeting website.

The OBI file grew 4.3% over the meeting course, whereby the increase in added defined classes (10.2%) was nearly double that of primitive classes (4.8%). Three new object properties were created during the meeting. These were used in a total of 68 new existential restrictions (9.7% increase). By inspection, increased chat indicates increased editing activity (see Table 2 of the supplementary information). The data also show different users working on different parts of the ontology and on different RU types. Apparent roles of users differed, e.g. 'moderators' creating tasks for others, which showed up in the metrics.

The lack of a RU and module locking mechanism meant there was no way to temporarily prevent others from altering classes that have a logical impact on the class under current definition. If a highly nested class description is created, it is difficult to get it right, unless others are prevented from changing something higher up in the hierarchy that will contradict the definition currently worked upon. Another method would be to just highlight areas that are currently worked on according to a colour scheme identifying the users, which then could resolve this by chat.

Checking for duplicate class and property labels and notification of the users would be useful. If two users added the same class concurrently, there was no notification after the duplication had occurred.

Priority has to be the undoing of deleted classes, because this can occur accidentally very easily in Protégé, e.g. by a single wrong click on the delete button or by accidentally moving classes. A roll back function would aid in conflict resolution and would lead to a safer editing. Undo/redo functionalities would be another feature to help users to prevent conflicts. Some non-deprecated properties were found to be sub-properties of deprecated properties, which seemed odd. Since currently there is also no global change list, it is impossible to see changes and annotations on deleted entities. If a parent class is deleted, all of its annotations disappear, including all children. The annotations will still be there, but since the association to the annotated RU is done via the ID only, without the label it is difficult to know what was annotated.

Subscription and Notification of changes was requested, where users subscribe to certain areas of interest within the ontology and are then notified of any changes that occur to those areas. Getting notified on changes chosen by a user, such as discussion threads or certain RUs, would help to stay up to date and proceed faster in conflict resolution. For example warnings and alerts could be passed to subsets of users to prevent duplicate or contradicting editings. A 'change view' on selected items that are on a watch list would help users to keep track on recent developments in their interest or responsibility-domain. A feed of all classes could be used to notify developers to subscribed classes. For

the annotation flag in the class hierarchy it would be practical to see when someone added some new annotation, e.g. the annotation flag should then get an exclamation mark, or blink, or should display an analog bar that indicates the amount of attached annotations (a measure of *topic-hotness*).

Versioning

A side benefit of using a real time collaborative approach is that complicated versioning strategies are not needed: SVN change track and diff functions are not feasible for OWL files. Using SVN the threshold to do minor changes can be increased on the user side, because a complicated merge back and conflict resolution needs to be carried out on the whole artifact level, even when logically non-conflicting changes were made. However, even when SVN is used, the change track captured in the ChAO knowledge base (KB), can be copied and distributed in some SVN log after updating owl files. One drawback here is that small changes result in a textual information overkill: For a human readable change history, the tool should just state ‘class x was moved from A to B’, instead of listing all involved quantum changes, e.g. ‘class x was deleted from A’, ‘class x was created under B’, …. Users would like the changes to be described in a high level abstraction, rather than overly granular.

Annotations on RUs with Entity Notes

Due to its abundant connotation with owl annotation properties, the term “annotations” as used in the CP GUI caused some initial confusion. Consequently, the “Annotations Tab” has now been re-labeled to “Entity Notes” which is clearer and more specific. “Discussion Threads” has been renamed to “Ontology notes” correspondingly. Unfortunately these name changes are now out of sync with the nomenclature used in the ChAO ontology.

Each annotation has a freetext subject field to fill in as well as its freetext value. For the majority of small annotations, it turned out that people did not use the subject heading, potentially because they felt to provide an annotation type, subject heading and value for small annotations is overkill. Seeing the annotations in a table view, e.g. sorted according to type, subject and value would make viewing easier. Axiom annotations, as being currently investigated for OWL2, were requested by some users as well.

The group observed that, to avoid information overload and to keep quality up, users should be allowed to remove their unintended annotations e.g. for the first 5 min of their creation.

Detailed statistics on numbers and kinds of annotations made during the sessions are available in a spreadsheet and diagrams in the supplementary material.

We positively note that in cases where the present (meta-) annotations are not sufficiently granular, the annotation types in CP’s underlying ChAO can be expanded with new annotation types that suit special projects needs and evaluation approaches.

Search and filtering of user annotations: It is possible to filter by author, annotation text, annotation type or by creation date, alone or in logical combinations. Own metadata schemes, e.g. certain obi annotation properties like `has_curation_status` or `definition_source`, can be queried for by the queries tab.

Communication

In the beginning, threads and notes were misused for chats and *vice versa*, the latter due to the chats' instant visibility and notification. Once a topic had started, it seemed to be difficult to find a cut off, when to move from a chat into an RU note or thread and *vice versa*. A good example of the consequence of not using the right modality for annotations was, when a participant warned the group about an obsolete property (`is_device_for`) in the threads and not in the more appropriate entity note for the object property RU. As a consequence it was found that nonetheless a warning had been issued, people used this obsolete property.

Chats were used for general acute issues and planning, e.g. “vote being held on @'http://purl.obo foundry.org/obo/Class_44”. Chats were requested to be linked with specific RUs and axioms to aid a more immediate and direct conflict resolution and not overload the (persistent) entity notes. A closed 'retreat room' was desired as well as a filter function on user names to enable to see only the chats of certain people or on particular ontology fragments.

The integration of emoticons in text fields could increase transmittance of pragmatic aspects of communications and would aid in the prevention of tensions on a sociologic level, i.e. allowing irony to be expressed.

Integrated voting on change issues, proved to be not fully implemented, but was needed by users. A mechanism that changes the ontology automatically

could increase KB development time and could be implemented using ChAO information and formalized voting outcomes.

Issue tracker functions were requested, i.e. a scratch pad or todo list that can be worked through and 'checked', e.g. indicating a proposed plan and what has been already realized at a certain time point. E.g. when people add new classes from a spreadsheet they should have a checklist that indicates which class has already been taken care of.

Performance

Overall, the performance of CP was very usable and much can be done with configuration to optimize it further. In large artifacts, expanding the full class hierarchy at once for the first time in one client can take its time (ca. 20 sec in our setup). Also opening a class with many direct subclasses for the first time will slow down and impair performance initially.

Discussion and annotation update throughout the clients was so slow, that it led people to use the chat functionality, which was updated and immediately visible. To see an Annotation update, people needed to change a frame and only then was the GUI updated. This bug has since been rectified by the protégé team.

Conclusion

In this investigation, a realistic collaborative ontology building session was created using CP and its features were thoroughly tested. Areas where user requirements were not fulfilled have been highlighted. Although some caveats persist and some requirements could not be fulfilled at this time, it became clear that the CP tool is now in an advanced stage and can be used in practice with sufficient stability. It copes with complicated setups and is flexible enough to allow for corresponding adjustments.

From an overall technical point of view, collaborative ontology building was relatively trouble free. The main area for improvement comes from the need for more sophisticated communication mechanisms. In editing, a mechanism for conflict resolution, e.g. 'undo/redo' is needed, as well as some transaction management. Although crucial to editing in a collaborative, concurrent, real-time fashion, this is not presently available in CP. Some enhancement of editing functionality and the addition of notifications on changes to notes and threads was deemed necessary. The addition of chats to specific RUs and for specific groups would enhance annotation traceability of the tool further. In all, CP as it stands now is already usable as a collaborative tool that we can recommend. Our analysis provoked much feedback to the CP developers, and will be valuable for the CP version of P4, which is currently in preparation. Further use of CP in controlled settings will enable us to acquire further insights into the process of tool-based collaborative ontology building and such findings will be fed back to tool development in the future.

Acknowledgements

Thanks go to Tania Tudorache, Timothy Redmond and Natasha Noy from the Stanford Protégé team and James Watson from the EBI teaching facilities. Further thanks go to all the ontogenesis network participants, and to Melanie Courtot, Alan Ruttenberg and the OBI Consortium for providing the merged OBI.owl file. DS is funded by the DebugIT project of the EU 7th FP (ICT-2007.5.2-217139). Support was received from the EBI's NET-project (www.ebi.ac.uk/net-projects) during some initial work leading to this paper. DS would like to thank his former and current employers Susanna Sansone and Stefan Schulz for making this work possible in-between affiliation change. JM is funded by the EMERALD EU project number LSHG-CT-2006-037686. The OntoGenesis Network is funded by EPSRC grant EP/E021352/1.

References

1. Noy N, Chugh A and Alani H. The CKC Challenge: Exploring Tools for Collaborative Knowledge Construction. BMIR-2007; p. 1260.
2. Tudorache T, Noy NF, Tu SW and Musen MA. Supporting collaborative ontology development in Protege, Seventh International Semantic Web Conference, 2008, Karlsruhe, Springer, Germany

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

Overcoming the Ontology Enrichment Bottleneck with Quick Term Templates

Philippe Rocca-Serra[1], Alan Ruttenberg[2], Jay Greenbaum[3], Melanie Courtot[4], Ryan R. Brinkman[4] Patricia L Wetzhel[5], Daniel Schober[6], Susanna Assunta Sansone[1], Richard Scheuermann[7], the OBI Consortium and Bjoern Peters[3]

[1]EMBL-EBI, Wellcome Trust Genome Campus, Cambridge, UK; [2]Science Commons, Cambridge, Massachusetts, USA; [3]La Jolla Institute for Allergy and Immunology, La Jolla, USA; [4]Terry Fox Laboratory, British Columbia Cancer Agency; [5]Stanford University, Stanford, CA, USA; [6]Institute for Medical Biometry and Medical Informatics, University Clinic, Freiburg, Germany; [7]Department of Pathology and Division of Biomedical Informatics, U.T. Southwestern Medical Center, Dallas, TX, USA

Abstract

The developers of the Ontology of Biomedical Investigations (OBI) primarily use Protégé[1] for editing. However, adding many classes with similar patterns of logical definition is time consuming, error prone, and requires the editor to have some expertise in OWL. Therefore, the process is poorly suited for a large number of domain experts who have limited experience Protégé and ontology development. We have developed a procedure to ease this task and allow such domain experts to add terms to the ontology in a way that both effectively includes complex logical definitions yet requires minimal manual intervention by OBI developers. The procedure is based on editing a Quick Term Template in a spreadsheet format which is subsequently converted into an OWL file. This procedure promises to be a robust and scalable approach for ontology enrichment.

Introduction

The Ontology of Biomedical Investigation[2] (OBI) project is developing an integrated ontology for the description of biomedical investigations based on the requirements of a diverse set of scientific communities. Briefly, OBI's development process is as follows: First, we selected the Basic Formal Ontology (BFO)[3] and the Relation Ontology (RO)[4] to provide basic organizational cues and a formal framework for representation. Collection of case studies and solicitation of term submissions are used to select candidates for inclusion in the ontology. A variety of techniques, including card sorting exercises, frequent teleconferences, and biannual meetings are used to develop entity definitions which are subsequently categorized along BFO's 3 main axes, namely processes, independent continuants and dependent continuants. This classification helps ensure that terms are placed in a correct *is_a* hierarchy. In order to represent more complex artifacts, OWL restrictions are expressed in term of RO relations between OBI defined entities and those defined in partner ontologies from the OBO foundry[5]. In combination with the creation of 'defined classes' which have logical necessary and sufficient definitions, a complex, expressive and logically rigorous class hierarchy is constructed that can be practically maintained and validated by reasoners such as Pellet[6] and FaCT[7].

A fundamental disadvantage of the process outlined above is that it does not scale well. Thousands of community term requests are already in the pipeline of the OBI project. Logical definitions are required by most of these terms in order to position them properly in the OBI hierarchy. Manually adding these classes and restrictions is time-consuming, error prone, and significantly limits the number of people who can contribute productively to enriching the ontology.

Our *Quick Terms* proposal is motivated by the observation that a significant number of such term requests can be expressed using a limited number of pre-defined design patterns. In order to engage domain experts without extensive expertise in ontology development, we formulate the required input for each such design pattern as a *Quick Term Template* (QTT), which can be edited in an Excel spreadsheet. In the following, we illustrate an example of a common term request, namely assays that measure the concentration of a specified molecular compound in a given material. Requests for terms to identify such assays come from diverse communities, including EBI's BioInvestigation index[8], the Immune Epitope Database[9] and Influenza Virus BioHealthbase[10]. This example illustrates the QTT process as a proof of principle. We claim that the approach can be easily extended to different design patterns.

Methodology and Results

The Quick Term Template submission process has four main steps: (1) Agreement by the OBI consortium on the logical definition of the parent class for submissions that match a certain pattern; (2) Identification of entities that can be varied with respect to the parent class (the differentia), for which a QTT spreadsheet is generated. This spreadsheet

contains one column for each such entity; (3) Processing the QTT submission through the use of a script that translates each row of a QTT template into a logically definition of a class; (4) Agreement by the OBI Consortium on how these classes will be integrated into OBI (*e.g.*, are they managed as a separate, imported, OWL file, or are they integrated into the core OBI itself). We elaborate on these steps below.

Step 1: Develop the Representation of the Parent Class

The example used throughout this section is a QTT for assays that measure the concentration of a specified molecular entity relative to a given material entity. These are called 'analyte assays' in OBI. Each definition links the material in which the concentration is measured (the evaluant), the molecular entity that is detected (the analyte), and the measurement being made (a scalar with a concentration value and unit).

Textual definition: An analyte assay is an assay with the objective to determine concentration of one substance (bearer of analyte role) that is present in (part of) another (bearer of the evaluant role). The output of the assay is information about concentration – a relational quality of the analyte towards the evaluant.

Logical definition:

'achieves planned objective' some 'analyte measurement objective'

and realizes some ('evaluant role' that 'inheres in' some 'material entity')

and realizes some ('analyte role' that 'inheres in' some 'scattered molecular aggregate')

and has_specified_output_information some ('scalar measurement datum' and ('is quality measurement of' some 'molecular concentration') and ('has measurement unit label' some concentration unit label'))

Figure 1: RO based restrictions of OBI Analyte Assay

Step 2: Derive Tabular Quick Term Template

We found that a large number of our current requests for terms are subclasses of analyte assay. Their differentiae are what the analyte is (*i.e.*, what the concentration is being detected of), and what the evaluant is (*i.e.*, what material the concentration is detected in). Accordingly, a Quick Term Template for an analyte assay needs columns for only those two entities. Table 1 depicts a QTT with several example terms, as they would be seen in a spreadsheet.

Analyte	**Evaluant**
Glucose CHEBI:17234	Urine FMA:12274
Interferon gamma PRO:000000017	Cell culture supernatant OBI:1000023
Glucose CHEBI:17234	<deliberately left blank>

Table 1: A basic QTT for submitting an analyte assay term request.

The template hides the complexity of modeling by only identifying the differentiating entities needed for the definition of the class while hiding the actual relations binding those entities together. The burden of building the logical definition is hidden from the user, and is instead accomplished during processing of the template. A template such as this one is accompanied by guidelines for users explaining what values are allowed the columns, and how they will be interpreted in building the assay. Documentation, and possibly software, will restrict the source ontologies from which terms may be selected for column, and may further restrict to use of certain subclasses in the specified ontologies.

Step 3: Submission processing

Following submission of a completed QTT, Perl scripts generate the logical definition of the class, expressed in OWL, as follows:

1. Parse the incoming tab delimited file for syntactic accuracy.
2. Identify each class to be included from an external ontology. As we are using OWL, all external resources need to be imported. In most cases, we rely on the MIREOT mechanism[11] to do so. In short, OBI relies on an `external.owl` file containing references to external ontology terms that retain their ID space. This file is then read to determine whether additions are required and if so they are added.
3. Create new OWL classes by populating the owl template with relevant values. As in the normal development of OBI, class identifiers are made unique to ensure that no conflicts arise when incorporating the newly generated class into OBI.
4. Add metadata. As a QTT submission creates fully logically defined terms, the creation of labels and textual definitions can be automated as well.

For the examples in Table 1, the class defined by

Row 1 is assigned the label ‘glucose concentration measurement in urine’, the class corresponding to Row 2 the label ‘interferon gamma concentration measurement in cell culture supernatant’, and the class corresponding to Row 3 the label ‘glucose concentration measurement’. Note that Row 3 in the table automatically gets classified as a superclass of Row 2. The empty cell is interpreted as any evaluant, and therefore every glucose concentration measurement in urine (Row 2) *is_a* glucose concentration measurement (Row 3).

Step 4: Integration within OBI

Check consistency. A consistency check and classification is carried out to verify integration and the integrity of the newly augmented OBI ontology.

Conclusion and Future Work

The QTT process outlined here provides a straightforward way to incorporate a large amount of key information submitted by domain experts in the communities OBI is designed to serve. Included in such are existing artifacts, such the IUPAC clinical chemistry resources[12] that contain hundreds of assays. The procedure has the potential to address the pressing need expressed by many communities wanting to submit a large number of terms for inclusion in OBI.

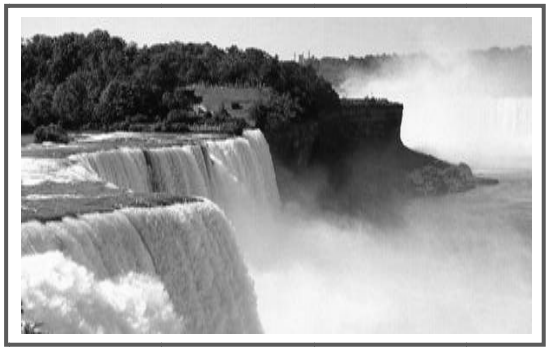
For future work, we plan to make the fully automated QTT submission process available for web-based submissions, with automated generation of OBI class identifiers. We would further like to explore the integration of OLS[13] or BioPortal[14] into the QTT process to streamline the process of finding existing terms, for instance by taking advantage of a full text search across appropriate external ontologies. Finally, the creation of a Protégé plug-in based on the QTT methodology would be a welcome addition to the functionality of the editor. OBI developers are keen to collaborate with other groups on this project in the hope others might benefit from this strategy for reducing bottlenecks associated with limited availability of knowledge engineers.

Acknowledgements

The OBI Consortium is (in alphabetical order): Ryan Brinkman, Bill Bug, Helen Causton, Kevin Clancy, Christian Cocos, Melanie Courtot, Dirk Derom, Eric Deutsch, Liju Fan, Dawn Field, Jennifer Fostel, Gilberto Fragoso, Frank Gibson, Tanya Gray, Jason Greenbaum, Pierre Grenon, Jeff Grethe, Yongqun He, Mervi Heiskanen, Tina Hernandez-Boussard, Philip Lord, Allyson Lister, James Malone, Elisabetta Manduchi, Luisa Montecchi, Norman Morrison, Chris Mungall, Helen Parkinson, Bjoern Peters, Matthew Pocock, Philippe Rocca-Serra, Daniel Rubin, Alan Ruttenberg, Susanna-Assunta Sansone, Richard Scheuermann, Daniel Schober, Barry Smith, Larisa Soldatova, Holger Stenzhorn, Chris Stoeckert, Chris Taylor, John Westbrook, Joe White, Trish Whetzel, Stefan Wiemann and Jie Zheng.

References

1. The Protégé Ontology Editor and Knowledge Acquisition System, http://protege.stanford.edu/.
2. OBI Ontology, http://purl.obofoundry.org/obo/obi.
3. BFO: http://www.ifomis.org/bfo
4. RO: Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL and Rosse C. Relations in biomedical ontologies. Genome Biol.;6(5):R46 (2005).
5. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, OBI Consortium, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL and Lewis S. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology, 25(11):1251–1255 (2007).
6. Sirin E, Parsia B, Grau BC, Kalyanpur A and Katz Y. 2007. Pellet: A practical OWL-DL reasoner. Web Semant. 5, 2 (Jun. 2007), 51–53.
7. Tsarkov D and Horrocks I. (2006) FaCT++ description logic reasoner: System description. In Proc. of the Int. Joint Conf. on Automated Reasoning (IJCAR 2006).
8. BioInvestigation Index: http://www.ebi.ac.uk/net-project/projects.html#bioinvindex
9. Peters B and Sette A. Integrating epitope data into the emerging web of biomedical knowledge resources. Nat Rev Immunol. Jun;7(6):485–90 (2007).
10. Squires B, Macken C, Garcia-Sastre A, Godbole S, Noronha J, Hunt V and Scheuermann RH. BioHealthBase: Informatics support in the elucidation of influenza virus host pathogen interactions and virulence. Nucleic Acids Res. 36(Database issue):D497–503 (2008).
11. MIREOT, http://obi-ontology.org/page/MIREOT
12. IUPAC clinical chemistry nomenclature
13. Cote RG, Jones P, Apweiler R and Hermjakob H. The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries. BMC Bioinformatics. 28;7(1):97 (2006).
14. NBCO Bioportal: http://bioportal.bioontology.org



ICBO

International Conference on Biomedical Ontology

July 24-26, 2009
Buffalo, New York, USA

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

The Cell Cycle Ontology: An Application Ontology Supporting the Study of Cell Cycle Control

Erick Antezana[1], Mikel Egaña[2], Ward Blondé[3], Robert Stevens[2], Bernard De Baets[3], Vladimir Mironov[4], Martin Kuiper[4]
[1]Department of Plant Systems Biology, VIB, Ghent, Belgium
[2]School of Computer Science, The University of Manchester, UK
[3]Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Belgium
[4]Department of Biology, Norwegian University of Science and Technology, Trondheim, Norway

Abstract

The Cell Cycle Ontology (CCO) is an application ontology that automatically captures and integrates detailed knowledge on the cell cycle process by combining, interlinking and enriching knowledge from various sources. CCO uses Semantic Web technologies, and it is accessible via the web for browsing, visualising, advanced querying, and computational reasoning. CCO facilitates a detailed analysis of cell cycle related molecular network components. Through querying and automated reasoning, it may provide new hypotheses to help steer a systems biology approach to biological network building. The ontology is available on http://www.cellcycleontology.org. Visual exploration can be done via the BioPortal, the Ontology Lookup Service, the Ontology Online service, or the DIAMONDS platform.

The Cell Cycle Ontology

The Cell Cycle Ontology captures detailed information (in terms and relationships) of the cell cycle process by combining representations from several, public sources.[1] CCO supports four model organisms (H. sapiens, A. thaliana, S. pombe and S. cerevisiae) with separate ontologies and one integrated ontology. It is an application ontology that is supplied as an integrated turnkey system for exploratory analysis, advanced querying, and automated reasoning.

CCO holds more than 13,000 concepts and 30 types of relationships. CCO comprises data from existing resources such as the Gene Ontology (GO), the Relations Ontology (RO), the IntAct database (MI), the NCBI taxonomy, the UniProt Knowledge Base as well as orthology data. An automatic pipeline builds CCO from scratch periodically: initially some existing ontologies (GO, RO, MI, in-house ones) are automatically fetched, integrated and merged, producing a core cell cycle ontology. Then, organism-specific protein and gene data are added from UniProt and from the GO Annotation files, generating four organism-specific ontologies. Those four ontologies are merged and more terms are included from an ontology built automatically from the OrthoMCL execution on the cell cycle proteins.

Formats and Queries

CCO is built in the OBOF format with ONTO-PERL and exported to other formats later.[2] CCO is available in: OBOF, RDF, XML, OWL, GML, and DOT. The Semantic Web formats RDF and OWL allow queries on CCO. In a SPARQL endpoint complex queries on the RDF format can be formulated, such as "retrieve all the core cell cycle proteins in S. cerevisiae that are located in the cytoplasm and that have a hydrolysisrelated function".

Relational closures are pre-inferenced in the RDF triple store, by operating SPARUL update queries over CCO and Metarel. This allows for very simple and responsive queries over long chains of relations in CCO.

Finally, during the maintenance phase, a semantic improvement on the OWL version is carried out: Ontology Design Patterns are included using the Ontology Pre-Processor Language. The resulting CCO is designed to provide a richer view of the cell cycle regulatory process, in particular by accommodating the intrinsic dynamics of this process.

References

1. Antezana E, Egaña M, Blondé W, *et al*. The Cell Cycle Ontology: An application ontology for the representation and integrated analysis of the cell cycle process, Genome Biology, 2009, 10:5
2. Antezana E, Egaña M, De Baets B, Kuiper M and Mironov V. ONTO-PERL: An api supporting the development and analysis of bio-ontologies. Bioinformatics, 2008, pp. 885–887.

Applying Biomedical Ontologies on Semantic Query Expansion

Andre Bechara, Maria Luiza M. Campos, Vanessa Braganholo
Informatics Graduate Program (PPGI), Federal University of Rio de Janeiro, Brazil

Abstract

This poster presents an ongoing work on using biomedical ontologies to improve efficiency on information retrieval.

Introduction

The interpretation of a question (or information need) depends, among other things, of a series of lexical-semantic relations that complement and help the cognitive process of answering that information need. Despite this fact, currently used information retrieval mechanisms take few advantages of the semantic interpretation of users' information needs (usually specified through keywords). In most of the cases, those mechanisms are based on keyword matching, and thus are excessively dependant on the query and document terms.

There are several past results[1,2] showing that, in general, information retrieval based on domain knowledge decreases the accuracy of keyword based search engines. We believe this approach deserves further discussion and experimentation, looking for more strong evidences that these negative results can really be generalized. Moreover, there are some questions left unanswered by previous work that our experiment is addressing:

(i) Using a scientific ontology, with formal construction and maintenance processes, such as the OBO ontologies, would produce better results? (ii) Are there more efficient query expansion techniques using available domain knowledge? (iii) Is a scientific ontology complete enough to fulfill the information retrieval researchers' needs, in general?

Semantic Query Expansion

To try to answer some of these questions, we run a query expansion experiment using the Gene Ontology (GO) as domain knowledge. As the document repository, we used an extraction of 10 years of Pub Med publications (from 1994 to 2004), which contains approximately 4.6 millions of documents. This dataset is a test-collection used by the information retrieval community, called Genomic TREC.

Results

To evaluate our ontology-based semantic query expansion technique, we measured the effectiveness of the information retrieval mechanism with and without expansion. In a nutshell, the average result showed an increase of 28% on synonyms relations and a small decrease on other relations.

Our results show a lot of consistence with past related work. In fact, if the expansion strategy does not selectively choose when and how to expand, only synonym relations are worth to be used. However, looking further, it is possible to find several opportunities to try other expansion strategies. For example, the problem with query expansion using generalization/specialization relationships is that, if it is always applied, the bad results are more frequent than the good ones. But, if the strategy is to be selective on when to use these relations for expansion, the increasing on accuracy can be outstanding. As shown by our experiment, there was a query with 98% increment on effectiveness.

Conclusion

We strongly believe that it is premature to assume that semantics-based query expansion is, in general, a recall-enhancing, precision-degrading technique. Our experiments suggest that by using scientific based ontologies (like OBO ontologies) with formal relations, it is possible to increase both recall and precision. Our group is currently revising this first experiment towards a better semantic query expansion strategy.

Acknowledgements

This work was partially funded by CAPES and CNPq research grants 311454/2006-2, 306889/2007-2 and 484713/2007-8.

References

1. Fox E. Lexical relations enhancing effectiveness of information retrieval systems. SIGIR Forum, New York, v.15, n.3,
2. Voorhees E. Query expansion using lexical-semantic relations. In: ACM SIGIR conference on research and development in information retrieval, Proceedings, Dublin:17, p.61–69, 1994

Developing a Mammalian Behaviour Ontology

Tim Beck, John M. Hancock, Ann-Marie Mallon
MRC Harwell, Harwell Science and Innovation Campus, Oxfordshire, UK

Abstract

The use of the Entity + Quality (EQ) model in phenotypic descriptions is dependent on the use of specialised domain ontologies to define the entity under observation. A domain currently lacking a specialised ontology is mammalian behaviour and so the Mammalian Behaviour Ontology is being constructed to address this. The ontology is manually developed and encourages contributions from domain experts. A top-level class distinction is made between individual behaviours and behaviours between two or more individuals.

Introduction

The EUMODIC project (http://www.eumodic.org) is generating a large volume of data from the high-throughput phenotyping of approximately 500 mutant mouse lines. EuroPhenome (http://www.europhenome.org) is the open-access database resource developed to contain this wealth of data. The phenotype data is defined using Open Biomedical Ontoloiges (OBO) derived post-composed terms, which combine Entities (E) and Qualities (Q)[1]. The qualities are consistently defined using the PATO ontology[2], however the entities are defined in alternative domain-specific ontologies depending on the class of the observation (anatomical, biochemical, behavioural etc). In the domain of animal behaviour, there currently exists a small subset of terms in existing OBO ontologies and the broad-scope Animal Behaviour Ontology[3], however a specialised ontology for the description of mouse behaviours is lacking.

A specialised ontology for the description of mammalian behaviours is a pre-requisite for the comprehensive annotation of phenotypes within EuroPhenome; and so work has begun on developing the Mammalian Behaviour Ontology (MBO).

Results

The MBO draws a fundamental distinction between adult behaviours in isolation from other individuals and behaviours between individuals, which are broadly analogous to the Gene Ontology biological process concepts “adult behaviour” (GO:0030534) and “behavioral interaction between organisms” (GO:0051705). This distinction forms the two top-level classes of the ontology (Figure 1). Subsequent child classes have been manually added and define behaviours observed during EUMODIC phenotype screens; those contained within the literature; and those defined as *abnormal* behavioural phenotypes within the Mammalian Phenotype ontology.

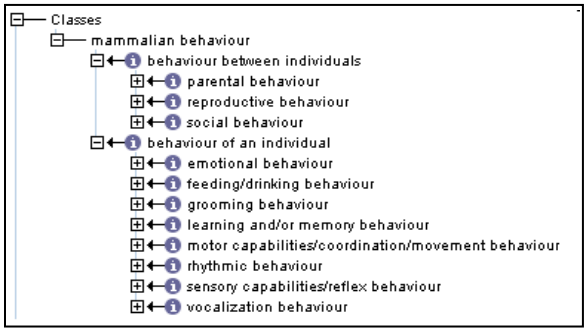


Figure 1. Top level classes of the Mammalian Behaviour Ontology

Conclusion

The MBO will be further developed in collaboration with mouse behavioural experts and will allow for precise definition of behaviour phenotypes using the EQ paradigm. The MBO will be used in conjunction with assay ontologies to define the environmental conditions experienced by the organism when exhibiting a specific behaviour, for example relating a compulsive biting behaviour to handling during the SHIRPA protocol.

Acknowledgements

This research was funded as part of the EUMODIC project (funded by the European Commission under contract number LSHG-CT-2006-037188).

References

1. Beck T, Morgan H, Blake A, Wells S, Hancock JM and Mallon A-M. Practical application of ontologies to annotate and analyse large scale raw mouse phenotype data. BMC Bioinformatics. 2009;10(Suppl 5):S2.
2. Gkoutos GV, Green EC, Mallon AM, Hancock JM and Davidson D. Using ontologies to describe mouse phenotypes. Genome Biol. 2005;6(1):R8.
3. Animal Behaviour Ontology Development – http://ontogenesis.ontonet.org/moin/AnimalBehaviourOntologyDevelopment

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

SNePS as an Ontological Reasoning Tool

Jonathan P. Bona, Stuart C. Shapiro
University at Buffalo, The State University of New York, Buffalo, NY, USA

Abstract

SNePS is a logic-based Knowledge Representation, Reasoning, and Acting system. We have created extensions to SNePS that allow its use as an ontology reasoning system, combining (1) instance data stored in a Referent Tracking[1] *system, (2) facts from domain ontologies such as FMA, and (3) rules from a domain-independent ontological theory. This paper discusses the first of these.*

Introduction

SNePS is a Knowledge Representation, Reasoning, and Acting system that combines elements of network-based, frame-based, and logic-based systems, with models of inference appropriate to each paradigm.[2] SNePSLOG, the logical language of SNePS is higher-order, allowing variables to range over predicates and function symbols, and term-based, allowing proposition-valued terms to be arguments of other functional terms. The software interface that is the subject of this paper connects SNePS to a prototype implementation of a referent tracking system (henceforth "The RTS").[3]

Electronic health systems in the Referent Tracking paradigm improve on previous systems by making unique, explicit, reference to individual particular entities in the world, universals, and the relationships that hold between and among these.[1] Each particular entity to which the system refers has assigned to it a globally unique identifier, which is used to refer to the entity in assertions about it that are stored in the system. Such assertions may also refer to universals and relations defined in external ontologies such as the Foundational Model of Anatomy (FMA), and to concept codes from SNOMED CT and others.

SNePS/RTS Interface

We have implemented an interface that allows SNePS to import data from an online RTS system in real time as it is needed in the course of reasoning, or as the result of a query, and perform inference on that data.

Entries stored in the RTS can be viewed as tuples, each representing an assertion by an agent that a state of affairs was observed at one time to obtain at some time. There are tuples to represent that a particular instantiates a universal, that a particular stands in some relation to other particulars, and more.

For example, the *Particular-to-Universal* tuple: *<id1,2004.03.23 21:37:53, instance of, OBO REL, id2, Face, FMA, 2004.03.23 21:37:53 >* says that the entity with id *id1* asserted on *23 March 2004* at *21:37:53* that the entity with id *id2* instantiates, via the *OBO REL* relation *instance of*, the FMA universal *Face* on *23 March 2004* at *21:37:53*.

These assertions are represented in SNePS as *Asserted(p,t,asn),* meaning that *p* asserted at time *t* that *asn.* Each assertion, *asn*, is represented by a proposition-valued functional term. For example, *Inst(p,u,r,t)* means that entity *p* instantiates universal *u* via relation *r* at time *t*.

The information in the *PtoU* tuple above is represented in SNePSLOG as the formula:

Asserted (id1,
time (2004,03,23,21,37,53),
Inst(id2, universal (Face, FMA),
relation(OBO_REL, instance_of),
time(2004,03,23,21,37,53)))

Conclusion

This interface allows SNePS to access instance data in the RTS *as needed* in the course of answering a query or performing other reasoning tasks. It provides a simple way of querying the RTS system from SNePS. This interface may also be used in combination with SNePS' native reasoning and acting capabilities to perform advanced reasoning tasks with patient data in a referent tracking system. In a similar manner, SNePS can import data from other sources (ontologies, concept systems, etc.), and combine them with data from the RTS.

References

1. Ceusters W and Smith B. Strategies for referent tracking in electronic health records. Journal of Biomedical Informatics. 2006; 39:362–378.
2. Shapiro SC and the SNePS Implementation Group (2008). SNePS 2.7 User's Manual. Department of Computer Science and Engineering, University at Buffalo, The State University of New York, Buffalo, NY.
3. Manzoor S, Ceusters W and Rudnicki R. Implementation of a referent tracking system. International Journal of Healthcare Information Systems and Informatics. 2007;4:41–58.

Organizing Search Results by Ontological Relations

Miao Chen
Syracuse University, Syracuse, NY, USA

Abstract

The study aims to propose a framework to organize retrieved results based on relations between query concepts. For user query containing more than one concept, we can identify semantic relations between the concepts with the help of domain ontology as well as natural language processing (NLP) and machine learning techniques. And then the relations will serve as the criteria of organizing and categorizing search results.

Introduction

In modern information retrieval systems, retrieved documents are primarily organized in two ways: relevance ranking and similarity (or distance) based clustering (Pratt, *et al*., 1999). There have also been trials in representing documents by knowledge-based approaches, i.e. organizing documents by concepts and hierarchical structure of ontologies (Pratt, *et al*. 1999; Chen and Dumais, 2000). However, non-hierarchical relations are seldom taken into consideration in this part of retrieval. To the best of our knowledge, there has been no study on organizing retrieved documents based on ontological relations (both hierarchical and non-hierarchical) between query concepts. Ontology provides comprehensive domain knowledge, including relations between concepts. Therefore ontological relations might help with organizing search results to facilitate user information seeking.

Research Design

The research design includes four primary steps to guide implementation. For experiment, we will use the UMLS ontology and Google search results.

1) Match query terms to ontology concepts and find their relations. For example, query "liver cancer, food" contains two concepts from the UMLS ontology. And the relations between their semantic types are "food causes liver cancer" or "food affects liver cancer";

2) In search result collection, we identify candidate sentences that have both concepts of the query and at least one candidate relation concept;

3) Syntactic level NLP algorithm will be used to parse each candidate sentence in the retrieved documents;

4) Classify candidate sentences into relation categories (determined by query term relations from ontology). Retrieved documents with the same relations between concepts are organized together. Following the example in step1, search results are categorized into two sets, the "cause" set and the "affect" set, based on the possible relations between liver cancer and food (as shown in Figure 1). Under the two categories, results can be further divided into sub-categories according to the UMLS relation structure. For example, the "treat" set is arranged under the "affect" set, the same way as in the ontological relation structure.

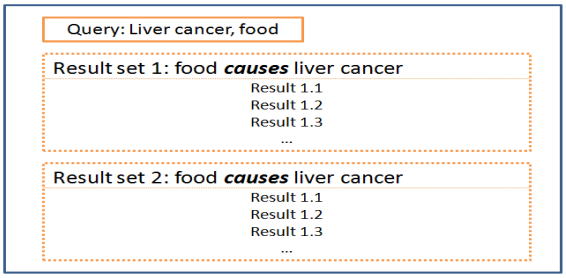


Figure 1. Search results organized by query relations

Evaluation

We will evaluate the new search result organization approach by comparing user satisfaction of two cases: 1) search results organized in relevance ranking, which is a result list from Google result; 2) retrieved results organized based on relations, which is the reorganization work of our study based on Google results. Content analysis will be conducted on the interview data to understand the strengths and weaknesses of relation-based organization from user perspective.

References

1. Chen H and Dumais S. Bringing order to the web: Automatically categorizing search results. Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'00); 2000.
2. Pratt W, Hearst MA and Fagan LM. A knowledge-based approach to organizing retrieved documents. Proceedings of AAAI 1999; 1999.
3. UMLS [Internet]. Maryland: NLM. Available from: http://www.nlm.nih.gov/research/umls/
4. Zamir O and Etzioni O. Grouper: A dynamic clustering interface to web search results. Proceedings of the Eighth International WWW Conference; 1999.

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

NEUROWEB:
A Case-Study of Clinical Phenotype Ontology in the Neurovascular Domain

Gianluca Colombo[1], Daniele Merico[2]
[1]University of Milano-Bicocca, Milan, Italy
[2]Terrence Donnelly Centre for Cellular and Biomolecular Research, Toronto, Canada

Abstract

We present the challenges encountered in the NEUROWEB project, as a case study of ontology for clinical phenotypes. The project aimed at developing an IT-support system for association studies. We identified the clinical phenotypes as the main entity of the ontological model, and we developed a model for their representation.

The Clinical Phenotypes Model

The NEUROWEB Project supports association studies in the neurovascular domain, integrating patient data from different clinical sites. Association studies typically require sets of patients with similar clinical conditions, i.e. with a common *clinical phenotype*. Therefore, we identified the representation of clinical phenotypes as the major modeling commitment. We selected a specific disorder, *Ischemic Stroke*, as a case-study. Ischemic Stroke is an occlusive disorder leading to local brain damage.

Neurovascular pathologies are typically organized taxonomically within clinical guidelines (such as the *TOAST* [1] for Ischemic Stroke) accepted by different clinical communities. For this reason, we initially devised a two-layer ontological model. The top layer (*Top Phenotypes*) was composed by the Ischemic Stroke taxonomy according to the TOAST. The lower layer, the *Core Data Set (CDS)*, was constituted by clinical indicators (a) essential for neurovascular diagnosis, and (b) connected to a specific field in the local clinical databases. Top Phenotypes were deconstructed into CDS elements using logical formulas reminiscent of SQL queries. This model was sufficient to select a pre-defined clinical phenotype and retrieve the corresponding patients from the databases.

To enhance the functionalities supported by the ontological model (flexible/modular definition of clinical phenotypes, integration to external terminologies, integration with genomic resources), the model was extended adding a middle layer, termed *Low Phenotypes*. We followed the following principles: A) set a distinction between *physiological events/processes* and their *diagnostic evidences*; the former entities were essential to establish connections to genomic entities through biological processes; B) set a distinction between the *etiological background* and the *traumatic point-event* characterizing a neurovascular disorder; this choice was suggested, in the specific case of Ischemic Stroke, by the partition between clinical findings (i) pertaining to the occlusion event and its downstream effects, and (ii) the underlying disease (e.g. atherosclerosis, diabetes) leading to the generation of the occluding body; C) set a distinction between (clinical) phenotypes and *anatomical parts* or *topological concepts*; the latter entities are not phenotypes, but they are used for the phenotype formulation.

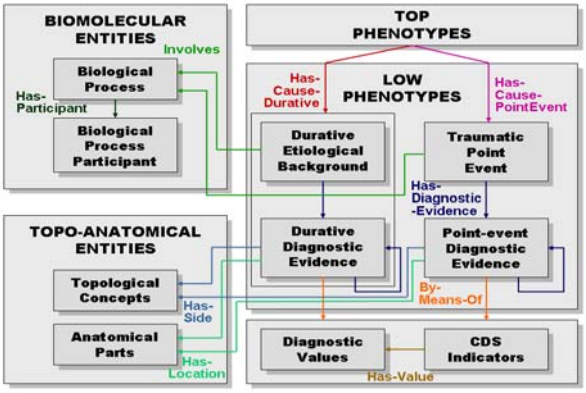


Figure 1. NEUROWEB ontology metamodel.

As a result, a *Top Phenotype* is decomposed into a *Durative Etiological Background* and a *Traumatic Point-Event* through the *Has-Cause-Durative* and *Has-Cause-PointEvent* relations respectively. These entities are connected to their diagnostic evidences via the *Has-Diagnostic-Evidence* relation, and to the *Biomolecular Processes* via the *Involves* relation. The *By-Means-Of* relation connects the *Diagnostic Evidences* to the *CDS* Indicator and its value range (*Diagnostic Value*) required for their assessment.

The NEUROWEB Reference Ontology was implemented in OWL-DL.

References

1. Adams HP Jr., *et al*. Classification of Subtype of Acute Ischemic Stroke, Definition for Use in a Multicenter Clinical Trial, TOAST. In Acute Stroke Treatment, Stroke 1993, 24: 35–41.

Logical Identity of Digital Files

Primavera De Filippi
European University Institute, Florence, Italy

Abstract

A proper analysis of copyright law necessarily requires accurate identifications of the various subject matters involved. The law provides a series of indications with regard to what constitutes a work of authorship for the purpose of the copyright regime, yet a fundamental question persists: What is it that really constitutes a work? Providing a definition of the term is not as intuitive as it may seem, especially not in the digital environment. As a general rule, the concept of a work is to be distinguished from the expression it has been articulated into and the manifestation be discerned from the item into which it has been embodied. In the digital world, however, the notion of an item may need to be somewhat adjusted so as to be more compatible with the common understanding of what constitutes a digital file.

Introduction

According to the Functional Requirement for Bibliographic Records (FRBR) of the IFLA,[1] any work of authorship can be subdivided into four basic components: the work, the expression, the manifestation and the item. Copyright protection, however, does not apply uniformly to every component of the work. In particular, the use of an item is not regulated by copyright law insofar as it does not infringe the copyright vesting in the work, the expression, or the manifestation thereof. In addition, the doctrine of exhaustion allows for any legitimately obtained item to be freely redistributed without the consent of the copyright owner. In the digital environment, however, the legitimate transfer of digital files may become impossible, as the transfer of a digital work would necessarily produce a new item of the work, which is physically different and yet logically identical to the former. It becomes, therefore, crucial to determine what constitutes a digital file and to identify when two digital files can actually be regarded as being identical.

Identification of Digital Items

An item is generally defined as a tangible carrier of information. It emerges from the fixation of a particular manifestation into a physical object and may only exist as a single instance (e.g. the copy of a book purchased by an individual, the specific digital file downloaded by a particular user, etc).

What appears to be a straightforward concept in the physical world has proven to be a controversial topic in the digital environment.

A digital item is a piece of content that is expressed in a digital format and subsists as a tangible entity on the physical memory of computers or other electronic devices. The transfer of a digital item, however, necessarily involves reproduction. Transfer can only be achieved by generating a copy in a new location and subsequently destroying the original. These two items are likely to assume a completely different physical representation while nonetheless maintaining their distinctive properties as digital items. Although they are physically distinct, from a structural perspective they are fundamentally identical and may therefore be recognized as one single entity. Accordingly, even if two digital entities will never be the same in an absolute sense, there are indeed circumstances where they should be considered, nevertheless, identical for some determined purpose.[2]

Conclusion

For the purpose of copyright law, therefore, an item may refer to any tangible entity which represents a unique exemplary of a particular manifestation of a work, taking into account that so long as the item can be identified as being the same, it is not necessary for the tangible entities to remain the same. An item can however only be properly identified when taking into account the context into which it is to be identified. Although the traditional FRBR framework may seem inadequate for the digital environment, an additional layer could be implemented to introduce the notion of "digital file" as a fictional container employed to determine logical equivalence between digital documents. In this way, the remaining conflicts existing between the definition of physical and digital items could be resolved.

References

1. IFLA (1998) Functional Requirements for Bibliographic Records. In Saur K G (ed.), IFLA Study Group on the Functional Requirements for Bibliographic Records.
2. Paskin N. (2003) On Making and Identifying a Copy. D-Lib Magazine, 9.

Creating a Translational Medicine Ontology

Christine Denney[1], Colin Batchelor[2], Olivier Bodenreider[3], Sam Cheng[4], John Hart[4], Jon Hill[4], John Madden[5], Mark Musen[6], Elgar Pichler[7], Matthias Samwald[8], Sándor Szalma[9], Lynn Schriml[10], David Sedlock[11], Larisa Soldatova[12], Koji Sonoda[13], David Statham[11], Susie Stephens[1*], Patricia L. Whetzel[6], Elizabeth Wu[14]

[1]Eli Lilly, Indianapolis, IN, USA; [2]Royal Society of Chemistry, Cambridge, UK; [3]National Library of Medicine, Bethesda, MD, USA; [4]Boehringer Ingelheim, Ridgefield, CT, USA; [5]Duke University, Durham, NC, USA; [6]Stanford University, Stanford, CA, USA; [7]AstraZeneca, Waltham, MA, USA; [8]DERI Galway, Galway, Ireland & KLI Austria; [9]Centocor R&D, San Diego, CA, USA; [10]University of Maryland, Baltimore, MD, USA; [11]Millennium Pharmaceuticals, Cambridge, MA, USA; [12]University of Aberystwyth, Aberystwyth, UK; [13]Amgen, Thousand Oaks, CA, USA; [14]Alzheimer's Research Forum, Cambridge, MA, USA

Abstract

We, participants in the Translational Medicine Ontology activity of the World Wide Web Consortium's Health Care and Life Sciences Interest Group (http://esw.w3.org/topic/HCLSIG) and members of the National Center for Biomedical Ontology (http://bioontology.org/), are developing a high-level, patient-centric ontology for translational medicine which will draw on existing domain ontologies and allow the integration of data throughout the drug development process.

Introduction

The pharmaceutical industry has historically focused on the development of novel blockbuster drugs. There is now an increasing focus on personalized medicines, requiring the right patients to receive the right drug at the right dose. In order to develop a tailored drug, manufacturers need to identify biomarkers that will indicate how a given patient will respond to a particular treatment. Biomarkers can also be used to demonstrate the comparative effectiveness of drugs, which is increasingly required by payers. Such translational medicine strategies require that traditionally separate data sets from early drug discovery through to patients in the clinical setting be integrated, and presented, queried and analyzed collectively. Ontologies can be used to drive such data integration and analysis; however, at present few ontologies exist that bridge genomics, chemistry and the medical domain.

The Translational Medicine Ontology, an application ontology that bridges the diverse areas of translational medicine, draws on existing domain ontologies where appropriate and will provide a framework centered on less than 50 types of entities.

Goals

The Translational Medicine Ontology will facilitate data integration from diverse areas of translational medicine such as discovery research, hypothesis management, formulation, clinical trials, and clinical research. It will serve as a template for further ontology development, enabling scientists to answer interesting and currently difficult questions more easily, especially those about data that are typically hosted by different functional areas. The ontology will provide a framework for the modeling of patient-centric information, which is essential for tailoring drugs.

Methodology

We have identified a set of 17 roles played by people across health care and the life sciences and collected (1) relevant questions, (2) the entities that those questions involve, and (3) applicable extant domain ontologies.[1] Types of entities include: disease, drug, patient, target, gene, risk, pathway, population, compound, phenotype, and treatment.

Next steps will involve identifying use cases based on those questions, determining which entities to build into the ontology and aligning them with BFO,[2] an upper-level ontology, to aid interoperability between domain ontologies. We will use one use case to test the Translational Medicine Ontology by building a data integration application based on it.

Conclusion

This project seeks to develop a patient-centric application ontology for translational medicine, as a collaborative effort between groups in industry and academia. The presentation will highlight our methodology, work to date, and future steps.

References

1. http://esw.w3.org/topic/HCLSIG/PharmaOntology/Roles
2. http://www.ifomis.org/bfo

* Current address: Johnson & Johnson, Radnor, PA, USA

Accurate Biochemical Knowledge Representation with Precise, Structure-Based Identifiers

Michel Dumontier, Leonid L. Chepelev
Department of Biology, Carleton University, Ottawa, Canada

Abstract

Biochemical ontologies aim to represent biochemical entities and the relations that exist between them in an accurate and precise manner. A fundamental starting point is the use of identifiers that precisely and uniquely identify some biochemical entity. Yet, our current approach for generating identifiers is often haphazard and incomplete. We describe plausible structure-based strategies for biochemical identity, ultimately to generate identifiers in an automatic and curator/database independent fashion, whether it is at molecular level or some part thereof.

Introduction

Accurate biochemical knowledge representation is embodied through the capacity to describe cellular events at all levels of biochemical granularity (organelle, membrane, substance, complex, molecule, molecular region, residue and atom) with their intended meaning wholly and unambiguously preserved. Fundamental to accurate knowledge representation is the ability to refer to real world entities in a precise manner, that is to say, that the identifiers for the entities to be described are readily available and that they consistently refer to the described entity. Having precise identifiers provides the basis by which access to independently generated knowledge becomes possible, at least initially by linked data[1].

Unfortunately, the overwhelming majority of biochemical identifiers are generated in a haphazard manner. For instance, a protein identifier may be generated based on the biological source, the biopolymer sequence, the encoding gene and/or the mRNA transcript from which it was translated. This means that structurally identical entities occurring in different contexts actually have different identifiers. The practice would be akin to giving you an arbitrarily generated name for each place you have been and each activity you have participated in, while maintaining that each of "you" is a different person. At the same time, biochemical entity modifications that have a radical impact on the structure and function are omitted or simply catalogued as annotations of *possible* structural variations for a given biochemical entity.

Results

With a clear need for precise biochemical identifiers, we describe three possible approaches for consideration by the community. The first simply involves the specification of existing chemical identifiers, the second employs a hybrid string-structure format, while the final approach takes advantage of the increased expressivity of the most recent Web Ontology Language (OWL2) to specify structural certainty including negative results as well as structural uncertainty from ambiguous or indeterminate experiments. Mechanisms for unique identification of parts and collections of parts are presented. These unique identifiers can be converted to an explicit representation in a variety of languages including OWL. Thus, not only can these identifiers be generated independently and have a consistently meaning, but they can also in themselves provides all the information about what they describe.

Conclusion

In order to efficiently communicate the results of biochemical experiments and to further integrate experimental information into semantic frameworks, the biochemical community needs a standard for unambiguous and accurate identification of biochemical entities. Just like the widespread implementation of SMILES and InChI representations has tremendously accelerated the pace of chemical research, we believe that the acceptance of a standardized, accurate, and descriptive biochemical identifier scheme within biochemical literature will be inexpensive while opening up countless new opportunities for data integration, reasoning over biochemical knowledge, and streamlined biological research.

Acknowledgements

MD and LLC thank the Canadian Foundation for Innovation and the Natural Sciences and Engineering Research Council of Canada for funding.

References

1. Belleau F, Nolin MA, Tourigny N, Rigault P and Morissette J. (2008) Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. J Biomed Inform. 41, 706–16.

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

An Ontology for RNA Structure and Interaction

Michel Dumontier[1], Jose Cruz-Toledo[1], Marc Parisien[2], François Major[2]
[1]Department of Biology, Carleton University, Ottawa, Canada
[2]Institute for Research in Immunology and Cancer (IRIC), Université de Montréal, Montréal, Canada

Abstract

Increasingly sophisticated knowledge about RNA structure and function requires ontologies to facilitate the integration of information arising from genome sequencing projects, microarray analyses and structure determination experiments. Here, we describe an ontology for nucleic acid composition along with context/model-specific representation of structural features such as sugar conformations, base pairings and base stackings. The ontology supports the goals of the RNA Ontology Consortium towards discovery of new knowledge from independently-published RNA data.

Introduction

The ability to accurately capture biomolecular behavior is critical to our understanding of cellular systems. Ribonucleic acids (RNAs) are essential cellular components with significant roles in protein synthesis and gene regulation. Increasingly sophisticated knowledge about RNA structure and function is being revealed as a result of innovative biochemical investigations such as genome sequencing projects, sequence alignments, microarray analyses, structure determination and RNA SELEX experiments. Yet, our capacity to capture this knowledge by existing representations is limited in at least one important respect. First, RNAML[1], an XML-based exchange format for a select subset of information, is not arbitrarily extensible by users. For instance, the nature of base stacking can be described with a natural language comment as part of the base-stack element, but we cannot specify a machine understandable type (e.g. adjacent stacking or upward stacking).

The use of formal logic-based languages such as RDF/OWL to describe knowledge about RNA structure and interactions provides the means for any researcher to further extend structural and functional annotations of experiments and biological objects in both a machine accessible and de-centralized manner.

Results

Here, we describe an RNA ontology for structure-oriented knowledge using RDF/OWL, Semantic Web technologies, that overcomes the limitations of XML-based approaches such as RNAML. The ontology provides knowledge for nucleic acid composition along with context/model-specific representation of structural features such as sugar conformations, base pairings and base stackings.

We populated the ontology with structural descriptions from the Protein Data Bank along with base pairing, base stacking interactions reported from MC-Annotate. The resulting RNA knowledge base enables powerful question answering over a reasoning-capable OWL-DL system.

Conclusion

This work provides the basis by which other essential RNA structural and functional features may be added. Furthermore, the ontology also enables the accurate representation of highly dynamic and context specific RNA structure interactions. Finally, it supports the aim of the RNA Ontology Consortium[2] "to create an integrated conceptual framework, an RNA Ontology (RO), with a common, dynamic, controlled, and structured vocabulary to describe and characterize RNA sequences, secondary structures, three dimensional structures, and dynamics pertaining to RNA function".

References

1. Waugh A, Gendron P, Altman R, Brown JW, Case D, Gautheret D, Harvey SC, Leontis N, Westbrook J, Westhof E, Zuker M and Major F. RNAML: A standard syntax for exchanging RNA information. RNA, 2002;8(6):707–17.
2. Leontis NB, Altman RB, Berman HM, Brenner SE, Brown JW, Engelke DR, Harvey SC, Holbrook SR, Jossinet F, Lewis SE, Major F, Mathews DH, Richardson JS, Williamson JR and Westhof E. The RNA Ontology Consortium: An open invitation to the RNA community. RNA, 2006;12(4):533–41.

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

Development of an Ontology of Microbial Phenotypes

Michelle Giglio[1], Chris Mungall[2], Peter Uetz[3], Lanlan Yin[3], Johannes Goll[3], Deborah Siegele[4], Marcus Chibucos[1], James Hu[4]

[1]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA
[2]Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[3]The J. Craig Venter Institute, Rockville, MD, USA
[4]Texas A&M University, College Station, TX, USA

Abstract

Phenotypic data are routinely used to elucidate gene and protein function in most organisms amenable to experimental manipulation. However, although phenotype ontologies exist for many eukaryotic model organisms, no standardized system exists for the capture of phenotypic information in bacteria. We propose to build an Ontology of Microbial Phenotypes and use it to annotate the prokaryotic model organism Escherichia coli.

Introduction

Phenotypes are the observable characteristics of an organism that result from the combination of a particular genotype and a particular environment, and thus are a basic and fundamental aspect of the biology of all organisms. The awesome power of genetics is founded on how the phenotypes of mutant genes, alone and in combination, contribute to understanding the biology of affected systems. To fully exploit the power of phenotypes for functional and comparative genomics, the ability to make comparisons across datasets and systems is vital. Making these comparisons either manually or computationally is hindered by the fact that phenotypes are not described consistently for bacteria. Our project aims to develop annotation infrastructure to improve the ability of microbiologists and bioinformaticians to use both existing and new phenotype information and to capture it in a consistent and standardized manner. This will require two key components: 1) an Ontology of Microbial Phenotypes (OMP) that captures phenotype descriptions in a controlled vocabulary, and 2) a set of evidence codes based on extension of the existing Evidence Code Ontology,[1] with links to a database of papers and other resources describing the assays used to "measure" these phenotypes.

Results

We have explored two parallel approaches to building the OMP. Both are pre-coordinated approaches that rely on using the terms in the Phenotypic Quality Ontology (PATO) as a basis for building up phenotype terms.[2] In the first approach we read 100 papers and identified 40 phenotypes described in those papers. We organized the 40 phenotypes into a controlled vocabulary using OBO-Edit.[3] While this effort was not comprehensive, we were able to classify the 40 phenotypes into five superclasses and assign PATO entities and qualities. In the second approach we generated a cross-product between a selection of PATO terms and two GO nodes relevant to microbial phenotypes, "cellular carbohydrate metabolism" and "amino acid metabolism." We found the cross-product generation method to be quite effective in generating large numbers of relevant terms quickly.

Conclusion

The manual and cross-product efforts were undertaken independently and in parallel by separate members of the group to see what, if any, consistency would be achieved. We found that although the concepts captured were similar, the different researchers chose different PATO quality terms to represent the same concepts. The manual curator chose "abnormal," while the person working on cross-products chose "abolished" and "disrupted." The results of this exercise illustrate one reason why the pre-coordinated approach has advantages over the post-coordinated approach. In the post-coordinated approach separate annotators creating phenotype annotations at different points in time may choose different ways of expressing the same concept and thus create inconsistency. In the pre-coordinated approach, one controlled set of PATO terms will be used for term generation, and the fact of storing all the terms in one controlled vocabulary will enforce consistency and uniformity.

References

1. http://www.obofoundry.org/cgi-bin/detail.cgi?id= evidence_code
2. http://bioontology.org/wiki/index.php/PATO: Main _Page
3. Day-Richter J, Harris MA, Haendel M, the Gene Ontology OBO-Edit Working Group and Lewis S. OBO-Edit–An ontology editor for biologists. Bioinformatics. 2007;23(16):2198–2200.

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

Clonal Complexes in Biomedical Ontologies

Albert Goldfain[1], Lindsay G. Cowell[2], Barry Smith[3]
[1]Blue Highway, Syracuse, NY, USA; [2]Duke University Medical Center, Durham, NC, USA; [3]University at Buffalo, Buffalo, NY

Abstract

An accurate classification of bacteria is essential for the proper identification of patient infections and subsequent treatment decisions. Multi-Locus Sequence Typing (MLST) is a genetic technique for bacterial classification. MLST classifications are used to cluster bacteria into clonal complexes. Importantly, clonal complexes can serve as a biological species concept for bacteria, facilitating an otherwise difficult taxonomic classification. In this paper, we argue for the inclusion of terms relating to clonal complexes in biomedical ontologies.

Introduction

Many of the difficulties in classifying bacteria stem from the fact that bacteria are both biological organisms (subject to biological classification) and, in certain circumstances, pathogens (subject to a disease-based classification). The fact that a bacterium can play the role of pathogen is an important ontological fact, but entities should not be classified solely on the basis of the roles they can play. If there is to be a bias in classifying bacteria, it should be a biological bias. This provides a more uniform classification scheme for all biological organisms.

Adopting such a classification brings up the problem of how to treat species at the microbiological scale. Following Mayr[1], we adopt the biological species concept in which differing species are separated by a barrier to gene flow.

Multi-Locus Sequence Typing (MLST) and Biomedical Ontologies

MLST is a popular method for achieving a biological classification for bacteria by using the allelic differences of seven housekeeping genes to determine the degree of relatedness between strains. "Clonal expansion [for bacteria] results from the rise in frequency of a single highly adaptive genotype. These ancestral genotypes subsequently diversify through recombination or mutation to produce minor clonal variants, and hence a 'complex' of closely related strains.[2]" The BURST clustering algorithm uses MLST data to infer this ancestral genotype and assign observed genotypes to clonal complexes. Seven genetic housekeeping loci are selected for a given genotype pool and the ancestral genotype is defined to be "the genotype within the clonal complex that differs from the highest number of other genotypes in the clonal complex at only one locus out of seven [these are called single locus variants (SLV)].[2]" The BURST algorithm succeeds when the assigned ancestral genotypes match the actual ancestral genotype in the phylogeny of the bacteria. BURST output is usable as biological species demarcation when the complexes are genetically isolated as illustrated in Figure 1.

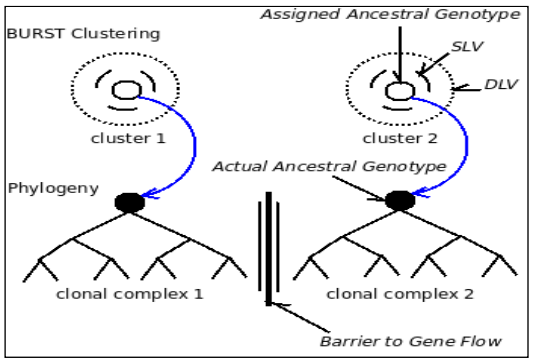


Figure 1. BURST clustering into clonal complexes

This technique does not always produce crisp demarcations between clonal complexes due to horizontal gene transfer. However, any biological taxonomy must tolerate some vagueness and fuzzy borders.

The inclusion of clonal complexes in biomedical ontologies requires the inclusion of several other terms: *clone*, *isolate*, *strain*, *housekeeping gene*, *ancestral genotype*, *recombination*, and *clonal divergence*. The adoption of these terms will yield a more uniform treatment of biotic entities of all sizes and will furnish a sound biological basis for disease ontologies.

Acknowledgements

This work was funded by the National Institutes of Health through Grant R01 AI 77706-01. Smith's contributions were also funded through the NIH Roadmap for Medical Research, Grant 1 U 54 HG004028 (National Center for Biomedical Ontology).

References

1. Mayr E. The Species Category. Toward a New Philosophy of Biology. 1988; 315—334.
2. Feil E and Man-Suen C. The BURST algorithm. http://pubmlst.org/analysis/burst/burst.shtml

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

The Evolution Ontology

Adam M. Goldstein
Darwin Digital Library of Evolution, AMNH, New York, USA

Abstract

Existing ontologies model components of evolution, but none synthesize them or describe the framework of ideas used to conceptualize evolution. The Evolution Ontology (EO) aims to do just this. EO models processes (e.g. natural selection); contexts (e.g. habitats); the entities that undergo evolution; and the theories, methods, and disciplines of evolutionary science. Uses include data curation, data mining, and literature curation, EO's developers working on the latter two for works of Darwin and the Biodiversity Heritage Library.

Motivations

That evolution provides an organizing framework for the life sciences and important disciplines and theories of the behavioral sciences is well-understood. EO aims to model this organizing framework, for literature and data curation and data mining. EO's focus on evolutionary processes will provide life scientists with a capability they presently lack, even in incipient form. At present, no ontology exists—indeed, no scheme of organization whatever exists—for representing the evolutionary process. MeSH,[1] is poor in terms describing evolution. There do exist models of entities at many levels in the hierarchy of biological organization and across taxa, including genes, phenotypes, homologies, and physiological processes.[2,3] Nonetheless, these have yet to be described in a manner that displays their roles in the evolutionary process.

EO aims to provide a way to organize information in a manner most useful for someone asking, of some variant gene or other "unit" of biological variation, Why (or how) did this (the actual) state of variation come to obtain? To answer questions like this, a researcher needs to know which evolutionary processes occurred in the history of the population under study; the "unit'" of variation that is evolving, and how it varies; and the causal background against which the changes have taken place (e.g. a habitat). As the user base of EO grows, researchers will be able to explore data and literature of evolution tailoring queries to reflect features of evolution most relevant to the population under study.

Modeling the concepts of evolution with an ontology is preferable to doing so by means of a subject heading list. The reasons for this are the same in the case of EO as they are in the case of other domains modeled using ontologies: they describe the relationships among the entities modeled, promoting discovery, and machine intelligence can be used to discover properties of the domain that otherwise would remain undetected.

Semantics

EO models evolution on the following schema: "If (and only if) entity *E* evolves in a context *X* by a process *P*, there is some change in a statistical property of *E*, which measures the degree to which a character *C* is present in the population." Processes *P* include natural selection, random drift, speciation, and the like. Context *X* is a generalization of the concept of habitat. An entity *E* is simply any population that can evolve; the character *C* is any heritable property of *E* whose distribution in *E* can change. EO also describes evolution in terms of disciplines, methods, and models and theories.

Literature Curation and Text Mining with EO

EO's developers will use it to organize the works of the Biodiversity Heritage Library.[4] The BHL, having digitized 12,605,478 pages, aims to publish all works for which permission can be obtained in the print collections of a worldwide consortium of libraries of Natural History. The BHL's search tool has rudimentary subject indexing tools, and because of this together with the BHL's massive size, it is virtually useless for information resource discovery. EO is also being used to mine Darwin's *Origin* in order to detect trends and connections in his developing thought.

References

1. NLM. Medical subject headings [home page on the Internet]. Bethesda, MD: NLM; 1999 Sep 1 [cited 2009 Apr 10]. Available from: http://www.nlm.nih.gov/mesh/meshhome.html.
2. Gene Ontology Consortium. The Gene Ontology [home page on the Internet]. Gene Ontology Consortium; 1999 [cited 2009 Apr 10]. Available from: http://www.geneontology.org/.
3. Smith B, Ashburner M, Rosse C, Bard C, Bug W, Ceusters W, *et al*. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007;25:1251–5.
4. BHL. Biodiversity Heritage Library [home page on the Internet]; 2006 [cited 2009 Apr 10]. Available from: http://www.biodiversitylibrary.org/.

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

Uberon: Towards a Comprehensive Multi-Species Anatomy Ontology

Melissa A. Haendel[1], Georgios V. Gkoutos[2], Suzanna E. Lewis[3], Christopher J. Mungall[3]
[1]University of Oregon, Eugene, OR, USA; [2]University of Cambridge, UK;
[3]Lawrence Berkeley National Laboratory, Berkeley CA, USA

Abstract

The lack of a single unified species-neutral ontology covering the anatomy of a variety of metazoans is a hindrance to translating model organism research to human health. We have developed an Uber-anatomy ontology to fill this need, filling the gap between the CARO upper-level ontology and species-specific anatomical ontologies.

Introduction

A number of anatomical ontologies (AOs) exist for specific organisms such as human, mouse and fish, many of which utilize the Common Anatomy Reference Ontology (CARO)[1] to structure the highest-level nodes. However, there is as yet a lack of a unified species-neutral ontology containing representations of embryonic and mature anatomical entities that conforms to OBO Foundry principles, similar to the Gene Ontology (GO) for gene function. Other cross species terminologies exist, such as MeSH, Gemina Anatomy and the Minimum Anatomy Terminology, but these resources are not formal ontologies, utilizing a single relation, and therefore do not provide an adequate substrate for reasoning across species.

Results

Uberon is a preliminary multi-species metazoan anatomy ontology created primarily to fulfil two requirements: (1) support translational research by allowing comparison of phenotypes across species and (2) provide logical cross-product definitions for GO biological process terms. The first version of Uberon was generated automatically by aligning existing species-specific anatomy ontologies (ssAOs) and anatomical reference ontologies, and then partially manually curated. Uberon retains reverse *is_a* links to the ssAOs, such that these can be used in cross-species inferencing and queries. A term is generally included in Uberon if it is a generalization over two or more existing species-specific anatomy terms. For example, `UBERON:dorsal_root_ganglion` subsumes `ZFA:dorsal_root_ganglion`, `MA:dorsal_root_ganglion`, and others. Uberon is homology-independent, and thus contains general terms for analogous structures that have evolved multiple times, such as `eye`. Future versions of Uberon may include evolutionary relationships between structures, along the lines of the TAO and BILA ontologies. Uberon attempts to employ *is_a*, *part_of*, *overlaps*, and developmental relations in the same manner as ssAOs. The current version of the ontology has 2808 terms, and 5110 links between terms, and 9339 links out to other AOs (Table 1), 1643 Wikipedia cross-references, and has been referenced in 682 GO cross-products.

Ontology	Type	Xrefs
FMA	Adult human	2302
MA	Adult mouse	1495
EHDAA	Embryonic human	838
ZFA	Zebrafish	811
TAO	Teleost	755
NIF	Neuroanatomy	701
GAID	Multi-species terms	626
CL	Cell	427
XAO	Xenopus	335
MAT	General	262
FBbt	Drosophila	243
AAO	Amphibian	103
BILA	Bilateria	64
WBbt	C elegans	63
CARO	Upper-level AO	34

Table 1. Number of terms in each AO referenced in Uberon. Each ontology is referenced by its unique ID space (see http://obofoundry.org)

Conclusions

Whilst Uberon is still in its early stages it has so far proven useful as a means of defining terms in the Gene Ontology, and as a means of comparing phenotypic descriptions of genotypic effects across species. Uberon is available from the OBO Foundry site and can be browsed at:
http://berkeleybop.org/obo/UBERON

References

1. Haendel MA, Neuhaus F, Osumi-Sutherland D, *et al*. CARO – The Common Anatomy Reference Ontology. In: *Anatomy Ontologies for Bioinformatics, Principles and Practice*, Burger A, Davidson D and Baldock R (eds.), 2007.

Towards Automatic Classification of Entities within the ChEBI Ontology

Janna Hastings, Paula de Matos, Marcus Ennis, Christoph Steinbeck
European Bioinformatics Institute, Hinxton, UK

Abstract

Biochemical 'small molecules' are a core element of biomedical data. ChEBI provides an ontology of chemical entities with stable unique identifiers and recommended names. Recently, ChEBI introduced direct user submissions, and the size of the database is forecast to grow substantially. Description logics provide a candidate technology for automatic classification of entities. However, as the complexity and size of the ontology increases, the efficiency of available reasoning technology will need to be assessed.

Background

Appearing in a wide variety of contexts, biochemical 'small molecules' are a core element of biomedical data. Chemical ontologies, which provide stable identifiers and a shared vocabulary for use in referring to such biochemical small molecules, are crucial to enable the interoperation of such data. One such chemical ontology is ChEBI (Chemical Entities of Biological Interest), a candidate member ontology of the OBO Foundry. ChEBI is a publicly available, manually annotated database of chemical entities and contains around 18000 annotated entities as of the last release (May 2009). ChEBI provides stable unique identifiers for chemical entities; a controlled vocabulary in the form of recommended names (which are unique and unambiguous), common synonyms, and systematic chemical names; cross-references to other databases; and a structural and role-based classification within the ontology. ChEBI is widely used for annotation of chemicals within biological databases, text-mining, and data integration. ChEBI can be accessed online at http://www.ebi.ac.uk/chebi/ and the full dataset is available for download in various formats including SDF and OBO.

Automated Classification

The selection of chemical entities for inclusion in the ChEBI database is user-driven. As the use of ChEBI has grown, so too has the backlog of user-requested entries. Inevitably, the annotation backlog creates a bottleneck, and to speed up the annotation process, ChEBI has recently released a submission tool which allows community submissions of chemical entities, groups, and classes. However, classification of chemical entities within the ontology is a difficult and niche activity, and it is unlikely that the community as a whole will be able or willing to correctly and consistently classify each submitted entity, creating required classes where they are missing. As a result, it is likely that while the size of the database grows, the ontological classification will become less sophisticated, unless the classification of new entities is assisted computationally. In addition, the ChEBI database is expecting substantial size growth in the next year, so automatic classification, which has up till now not been possible, is urgently required. Automatic classification would also enable the ChEBI ontology classes to be applied to other compound databases such as PubChem.

Description Logic Reasoning

Description logic based reasoning technology is a prime candidate for development of such an automatic classification system as it allows the rules of the classification system to be encoded within the knowledgebase. Already at 18000 entities, ChEBI is a fair size for a real-world application of description logic reasoning technology, and as the ontology is enhanced with a richer density of asserted relationships, the classification will become more complex and challenging. We have successfully tested a description logic-based classification of chemical entities based on specified structural properties using the hypertableaux-based HermiT reasoner, and found it to be sufficiently efficient to be feasible for use in a production environment on a database of the size that ChEBI is now. However, much work still remains to enrich the ChEBI knowledgebase itself with the properties needed to provide the formal class definitions for use in the automated classification, and to assess the efficiency of the available description logic reasoning technology on a database the size of ChEBI's forecast future growth.

Acknowledgements

ChEBI is funded by the European Commission under SLING, grant agreement number 226073 (Integrating Activity) within Research Infrastructures of the FP7 Capacities Specific Programme, and by the BBSRC, grant agreement number BB/G022747/1 within the “Bioinformatics and biological resources” fund.

VO: Vaccine Ontology

Yongqun He[1], Lindsay Cowell[2], Alexander D. Diehl[3], Harry Mobley[1], Bjoern Peters[4], Alan Ruttenberg[5], Richard H. Scheuermann[6], Ryan R. Brinkman[7], Melanie Courtot[7], Chris Mungall[8], Zuoshuang Xiang[1], Fang Chen[1], Thomas Todd[1], Lesley Colby[1], Howard Rush[1], Trish Whetzel[9], Mark A. Musen[9], Brian D. Athey[1], Gilbert S. Omenn[1], Barry Smith[10]

[1]University of Michigan, Ann Arbor, MI, USA; [2]Duke University, Durham, NC, USA; [3]Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, ME, USA; [4]La Jolla Institute for Allergy and Immunology, La Jolla, CA, USA; [5]Science Commons, Cambridge, MA, USA; [6]U.T. Southwestern Medical Center, Dallas, TX, USA; [7]BC Cancer Agency, Vancouver, BC, Canada; [8]Lawrence Berkeley National Laboratory, Berkeley CA, USA; [9]Stanford University, Stanford, CA, USA; [10]University at Buffalo, Buffalo, NY, USA

Abstract

The collaborative, community-based Vaccine Ontology (VO) was developed to promote vaccine data standardization, integration, and computer-assisted reasoning. Currently VO covers a variety of aspects of the vaccine domain, with an emphasis on classification of vaccines and vaccine components, and on host immune response to vaccines. VO can be used for a number of applications, e.g., ontology-based vaccine literature mining through collaboration with the National Center for Integrative Biomedical Informatics (NCIBI).

Introduction

Vaccination is the most important invention to prevent various diseases and improve public health. With extensive vaccine research and clinical usages, it has become challenging to standardize vaccine annotation, integrate information about varied vaccine types, and support computer-assisted reasoning. To address this challenge, we developed the community-based Vaccine Ontology (VO*;* www.violinet.org/vaccineontology).

Results of VO Development

As of June 8, 2009, VO contains 1802 classes, 192 object properties, and 13 datatype properties. Among these terms, 934 classes and 19 properties are assigned VO-specific IDs. In addition, VO includes 38 classes from the Basic Formal Ontology (BFO; www.ifomis.org/bfo) as upper-level framework, 24 terms from Relation Ontology (RO), 37 classes from Ontology for Biomedical Investigation (OBI), and many terms from other ontologies. VO development follows the OBO Foundry principles (obofoundry.org/crit.shtml).

VO has defined 'vaccine' as a 'processed material' that is prepared and used to protect against a pathogen organism or a disease (e.g., cancer). For example, the vaccine Fluvirin has the following hierarchical structure by definition: vaccine -> viral vaccine -> Influenza virus vaccine -> Fluvirin. More than 300 licensed vaccines and vaccine candidates in research or clinical trials have been described in VO. Vaccine components, vaccination protocols, and host responses to vaccination are also major focuses of current VO development.

VO can be used for a number of applications. For example, VO dramatically improves PubMed vaccine literature searching and is being applied to the development of an ontology-based vaccine literature mining system through collaboration with NCIBI. Vaccine-specific immune networks are being investigated using ontology-specific literature mining and advanced statistical methods.

Discussion

VO will include all licensed vaccines in different countries and regions, as well as all possible vaccines in clinical trials and in research for major diseases. Planned future development of the VO will add further details such as clinical trials of vaccine, vaccine surveillance, and safety reports. VO will allow advanced integration and intelligent analysis of large amounts of worldwide vaccine data.

Acknowledgements

This research is supported by a Rackham Pilot Research project at the University of Michigan, and by NIH grants U54-DA-021519 and N01AI40041. Strong support from NCBO, NCIBI, the OBI Consortium, OBO Foundry, and IDO Initiative are acknowledged.

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

Contributions to the Formal Ontology of Functions and Dispositions: An Application of Non-Monotonic Reasoning

Robert Hoehndorf[1,2], Janet Kelso[2], Heinrich Herre[1]
[1]**Institute for Medicial Informatics, Statistics and Epidemiology, University of Leipzig, Germany**
[2]**Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany**

Abstract

We introduce a basic ontology of functions and dispositions. The theory we suggest is developed as module of the General Formal Ontology, and is compatible both with major philosophical theories of biological functions and with most top-level ontologies. The particular focus of the suggested formalism is on the inference of causal relationships from function ascription and the explicit formalization of the normal character of functions using non-monotonic forms of knowledge representation.

The question of functions in biology is a major topic in the philosophy of biology and also continues to be of importance in applied biological sciences. Identifying the functions of proteins, DNA or RNA fragments, as well as the ontogenesis of their functions throughout evolution, remain a central topic of research. Inquiries into the nature of functions can reveal methods for determining the functions of specific entities, for distinguishing multiple functions of one entity, or can provide frameworks for describing the evolution of functions over time.

Theories pertaining to the ontology of *function* range from reductions to causality to explanations based on social ascription. In the context of formal ontology, *function* is investigated in several top-level ontologies, such as BFO and GFO. We provide a formal ontological account of *function* which is compatible with several top-level ontologies. For this purpose, we investigate the difference between dispositions and functions and propose a means to interrelate both kinds of entities using methods from artificial intelligence research, in particular non-monotonic reasoning.

The Ontology of Functions (OF)[1] provides an account of how to represent functions and how to represent their relations to other entities such as processes and objects. The basic assumption is that functional knowledge can be represented and described independently of the realization of function. In the OF, a function structure is described by a label, requirements, a goal and a functional item. The label is a non-formal name or description of the function. The requirement is a situation type whose instances must be present for every realization of the function. The goal is a situation type whose instances describe the states of the world that the function is supposed to cause or otherwise bring about. The functional item describes the role that entities with the function play, selecting all features of the entity that are essential to the function realization.

The theories on function differ in how they analyze the **has-function** relation. However, there are causal facts that should be exhibited by any functional entity in most of these theories: the function-bearing entity must *normally* be able to **cause** the goal of the function given the requirements of the function.

To formalize this observation, we use an additional entity in the ontology of functions. We call this a *disposition*. An individual *e* has the disposition *d* to cause T_{goal} iff *e* **causes** an instance of T_{goal} whenever *e* is placed *in the right circumstances*[2].

We suggest that for every function *F* there is a category *D* of dispositions with the same requirements and goals as *F* such that every individual *e* having a function *f*::*F normally* has a disposition *d*::*D*, and formalize this condition using predicate circumscription[3].

The ontological theory of function we have developed is intended to be compatible both with a wide range of philosophical theories on function and with most upper-level ontologies. It permits the inference of causal relations from function ascription, a feature of particular importance in biological ontologies.

References

1. Burek P. *Ontology of Functions*. PhD thesis, University of Leipzig, Institute of Informatics (IfI), 2006.
2. Hoehndorf R, Kelso J and Herre H. Contributions to the formal ontology of functions and dispositions: An application of non-monotonic reasoning. In *Proceedings of Bio-Ontologies 2009: Knowledge in Biology*, 2009.
3. McCarthy J. Applications of circumscription to formalizing common sense knowledge. *Artificial Intelligence*, 28(1):89–116, 1986.

What's in an 'is a' Link

William R. Hogan
UPMC and University of Pittsburgh, Pittsburgh, PA, USA

Introduction

Several researchers have demonstrated that current medical terminologies and ontologies use relations in inconsistent and ambiguous ways,[1] despite Woods' seminal work that first illustrated the problem.[2]

The goal of the present work is to catalog the different ways in which SNOMED CT uses the *is a* relation. The rationale for creating the catalog is to serve as a basis for systematically improving the semantics of terminologies and ontologies so as to improve their accessibility to machine inference.

Methods

I reviewed the literature to find ontological mistakes that change the interpretation of the *is a* relation, without respect to any particular terminology. I then reviewed the stated relationships table of SNOMED CT, Jan 2009 version, placing them into categories from the literature, and creating new categories when existing categories did not apply.

Results

I found nine categories of misuse of the *is_a* relation in SNOMED CT (Table), eight from the literature and one from my analysis. SNOMED CT had an example of every misuse found in the literature.

Discussion

The January 2009 version of SNOMED CT violates its intended interpretation of the *is_a* relation, which is nearly identical to the definition of Smith et al.[1] I cataloged nine categories of misuse of *is_a*, and found an example of each in SNOMED CT.

This study demonstrates for the first time that (1) common ontological mistakes lead to ambiguity in the interpretation of *is a*, (2) the stated relationships of SNOMED CT are the source of mistakes in the use of the *is a* relation, (3) SNOMED CT has at least one example of every problem with *is a* elucidated from the broader literature.

References

1. Smith B, Ceusters W, Klagges B, *et al*. Relations in biomedical ontologies. *Genome Biol.* 2005;6(5):R46.
2. Woods W. What's in a link? Foundations for semantic networks. In: Bobrow D, Collins A, eds. *Representation and Understanding*. New York: Academic Press; 1975:35–82.

Interpretation	Description	Example
Use-mention confusion	*Is a* states something that is true of the representational unit, not the class itself.	*Acquired disease is_a Navigational concept*
Instance of	*Is a* relates an instance to a class	*Canada is_a North American country*
Some instances play the role of	*Is a* relates class to role played by only some of its instances	*Bacteria is_a Infectious agent*
Epistemological criterion	*Is a* relates a statement about what/how something is known about a class to the class	*Colitis presumed infectious is_a Colitis*
Absence of	*Is a* relates statement that all instances of class are absent to the class itself	*HLA antigen absent is_a HLA antigen*
Part of	*Is a* relates part to whole	*Globin chain is_a Hemoglobin*
Has part	*Is a* relates whole to part	*Albumin bound paclitaxel is_a Albumin*
Is not a	*Is a* connects two ontologically disjoint classes	*Invasive blood pressure is_a Blood pressure*
Set inclusion	*Is a* connects a term designating a set of entities to one member of the set	*Type 1 diabetes mellitus with hypoglycemic coma is_a Type 1 diabetes mellitus*

Table 1. Nine Misuses of the *is a* Relation in SNOMED CT.

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

NIFSTD: A Comprehensive Ontology for Neuroscience

Fahim T. Imam, Sarah M. Maynard, Maryann E. Martone, Stephen D. Larson, Amarnath Gupta, Jeffrey S. Grethe
University of California, San Diego, CA, USA

Abstract

As a core component of Neuroscience Information Framework (NIF) project, NIF Standard (NIFSTD) was envisioned as a set of modular ontologies that provide a comprehensive collection of terminologies to describe neuroscience relevant data and resources. We present here on the structure, design principles and current state of NIFSTD.

Introduction

The NIFSTD is a critical constituent in the NIF project (http://neuinfo.org) to enable an effective concept-based search mechanism against a diverse collection of neuroscience resources. The overall ontology has been assembled in a form that promotes reuse of standard ontologies in biomedical domain, easy extension and modification during its evolution.

Basic Structure and Design Principles

The NIFSTD is constructed according to best practices closely followed by the Open Biological Ontology (OBO) community. It was built in a modular fashion, each covering a distinct orthogonal neuroscience relevant domain (e.g., anatomy, cells, molecules, experimental techniques, digital resources). NIFSTD avoids duplication of efforts by conforming to standards that promote reuse. The modules are standardized to the same upper level ontologies, the Basic Formal Ontology (BFO), OBO Relations Ontology (OBO-RO), and the Ontology of Phenotypic Qualities (PATO). Expressed in OWL-DL, NIFSTD is computationally decidable. NIFSTD follows the single inheritance principle but classes with multiple parents are derivable via automated reasoning on logically defined classes. Entities in NIFSTD are named via unique identifiers and are accompanied by a rich set of annotation properties. NIFSTD reuses object properties from standard ontologies (e.g., OBO-RO) to express the Intra-module and cross-module relations among classes. Within the NIF, NIFSTD is served through an ontology management system called OntoQuest. OntoQuest generates an OWL-compliant relational schema and supports operations for navigating, path finding, hierarchy exploration, and term searching in ontological graphs. We strive to balance between the involvement of the neuroscience community for domain expertise and knowledge engineering community for ontology expertise when constructing the NIFSTD. The wiki version of NIFSTD, NeuroLex (http://neurolex.org) has been developed as the easy entry point for the community to access, edit and enhance the core lexicon.

Current State

We have released the 1.0 version of NIFSTD (http://purl.org/nif/ontology/nif.owl), built upon release 0.5[1]. Version 0.5 was assembled from various external sources and had several shortcomings. Compare the 0.5, Improvements in 1.0 include: reduction of modular dependencies into minimum, a re-engineered import hierarchy to eliminate the redundant imports, elimination of duplicate classes due to multiple imports, normalization of the modules to create cleaner hierarchies, additional module for chemicals (reusing neuroscience relevant terms from CHEBI ontology) and resource types, and enrichment of contents (e.g., additional classes, synonyms, and other annotations) from NeuroLex. Although NIF relies on existing terminologies rather than re-invention, we do provide neuroscience-specific content where required. For example, the NIF cell module has largely been created by the NIF cell working group.

Conclusion

Currently covering about 20,000 concepts including both classes and synonyms, the NIFSTD continues to evolve to incorporate new modules and contents as well as implementing more detailed and useful cross-domain relations that follow ontology development best practices.

Acknowledgement

Supported by a contract from the NIH Neuroscience Blueprint HHSN271200800035C via NIDA.

References

1. Bug W, Ascoli G, Grethe J, Gupta A, Fennema-Notestine C, Laird A, Larson S, Rubin D, Shepherd G, Turner J and Martone M. The NIFSTD and BIRNLex vocabularies: building comprehensive ontologies for neuroscience. Neuroinform.2008;6:175–94

L_{BFO}: Toward an Artificial Language for Ontology Development

Leonard F. Jacuzzo
National Center for Biomedical Ontology, University at Buffalo, Buffalo, NY, USA

The syntax of L_{BFo} represents the initial step toward the creation of a rigorously characterized, recursively defined, artificial language for the sole purpose of ontology development. The underlying idea is that maximally fruitful application of ontology requires accurate representation of reality in accordance with current textbook science. Hence, creating a robust, accurate representation of reality is a fundamental concern.

An ontology represents general types of entities and relations between them. A domain ontology represents the general types and relations for a given domain of research. A top-level ontology represents the general types of entities in any domain of research. Ontologies serve many purposes in computerized collection, management, and storage of data. These applications include enhancement of storage and retrieval in a data system, integration of diverse systems, integration of semantic content on the web, and annotation of publications in a library setting.

Successful application of ontologies has led to the creation of languages with the special purpose of implementing ontologies. A formalized ontology is an ontology expressed in accordance with the grammatical formation rules of an artificial language. Some existing ontology languages have been developed in order to serve specific functions that require expressibility limitations and expression of information in a manner that contributes to human misunderstanding and error. The most potentially detrimental effect is risked when an ontology is constructed in a language designed exclusively for computerized implementation. The result is a skewed representation of salient features of reality. An ontology development language has two purposes: one is to represent reality as accurately and completely as possible, the other is to achieve this in a manner that facilitates computerized implementation: these goals conflict. Validation requires expert human consensus, hence, an ontology should be developed in a language that is understandable to domain experts. However, such a language must be computer tractable, i.e., there must be a correspondence between the information expressed with a sentence and its grammatical structure such that information can be processed on the basis of syntax alone.

L_{BFO} represents the initial step toward the creation of a rigorously characterized, recursively defined, artificial language for the sole purpose of ontology development. The underlying idea is that maximally fruitful application of ontology requires accurate representation of reality in accordance with current textbook science. Hence, creating a robust, accurate representation of reality is a fundamental concern.

L_{BFO} will facilitate providing definitions and characterizations of features of reality in a way conformant with BFO thus ensuring maximal rigor and clarity. Since L_{BFO} is a multi-sorted language, L_{BFO} has resources to represent the ontological categories found in BFO and the universals defined in their terms in an economical and at the same time user-friendly way.

Capitalized variables range over universals, while lower-case variables range over individuals. Universal constants are upper-case. Individual constants are lower-case. The syntax of L_{BFO} also distinguishes in a straightforward manner between variables for continuants, processes, and times. The syntax of L_{BFO} contains precisely expressed grammatical-formation rules, so that its variables cannot be combined in a manner that results in category errors. The predicates of L_{BFO} are such that the ontological category from which terms representing entities can be taken as arguments is specified in advance. Sentences which express category errors are not grammatically correct in L_{BFO}.

L_{BFO} can serve as a bridge between domain experts, knowledge engineers, and implementation languages. The semantic apparatus of an FOL system serves as the basis for the models developed for implementation languages such as OWL and RDF. FOL is also a segregated dialect of Common Logic so there is a link to that international standard; hence, there is potential to develop middle-ware that maps L_{BFO} to the variety of implementation languages that exist both now and in the future.

Though there is much work to be done in perfecting L_{BFO}, this first step in the process provides hope for achieving the goal of facilitating maximally accurate, rigorous representations of general features of reality.

Acknowledgements

This work was supported by NIH grant U54 HG004028.

NeuroLex.org: A NIF Standard Ontology-Based Semantic Wiki for Neuroscience

Stephen D. Larson, Sarah M. Maynard, Fahim Imam, Maryann E. Martone
University of California, San Diego, San Diego, CA, USA

Abstract

Bridging the domain knowledge of a scientific community and the knowledge engineering skills of the ontology community is still an imperfect practice. Within the field of neuroscience, we have tried to close this gap by presenting an ontology through the medium of a wiki where each page corresponds to a class. By opening it to the World Wide Web, we have made the process of maintaining a ~20,000 concept neuroscience ontology (NIFSTD), more collaborative.

Introduction

The neuroscience community needs its basic domain concepts organized into a coherent framework. Ontologies provide an important medium for reconciling knowledge into a portable and machine readable form. For many years we have been building community ontologies for neuroscience, first through the Biomedical Informatics Research Network and now through the Neuroscience Information Framework projects (http://neuinfo.org). These projects resulted in the construction of a large modular ontology, constructed by importing existing ontologies where possible, called NIFSTD[1]. One of the largest roadblocks that we encountered was the lack of tools for domain experts to view, edit and contribute their knowledge to NIFSTD. Existing editing tools were difficult to use or required expert knowledge to employ. By combining several open source technologies related to semantic wikis and NIFSTD[1], we have created NeuroLex.org, the first semantic wiki for neuroscience.

Methods

NeuroLex.org is built on top of the open source Semantic Mediawiki platform[2]. This allows classes, properties, and instances to be represented within a wiki which is easily editable and allows the content of that wiki to be exported as OWL. Semantic Mediawiki makes querying the ontology via properties or class hierarchy very straightforward. In addition, we have incorporated tools such as Semantic Forms, which allow the ontology classes to be edited as a form rather than as a wiki page with special text mark-up. Some of the fields in the form support autocomplete which allows users to populate those fields with other classes from the ontology.

Results

NeuroLex.org has evolved into a powerful platform for collaboratively maintaining and extending the NIFSTD ontology. We have been able to incorporate user feedback and create custom views of the ontology content with very rapid turnaround. Table 1 shows some key metrics that we have collected on its usage. The content contributed to the Neurolex is not directly added to NIFSTD, but is incorporated into the NIFSTD OWL file by a knowledge engineer after curation by the NIF ontology group.

Contributing neuroscientists	~12
Average edits per weekday	~25
Average hits per weekday	~220
% increase hits 01/09 – 04/09	~870%

Table 1. Key metrics for usage of NeuroLex.org.

Conclusion

We conclude that the Semantic Mediawiki is a good starting point for the collaborative maintenance of ontologies. Other groups are also using a similar approach, e.g., BioMedGT[3]. While we are still working through some issues**,** e.g., synchronizing the NIFSTD with the content on NeuroLex, exporting and importing OWL, and bulk uploading concepts, we believe that semantic wikis are a good tool for providing community contribution and feedback to projects like NIF.

Acknowledgements

With thanks to the Mediawiki project, the Semantic Mediawiki Project, Ontoprise, Yaron Koren, and Anders Larsson.

References

1. Bug WJ, Ascoli G.A, *et al.* (2008) The NIFSTD and BIRNLex vocabularies: Building comprehensive ontologies for neuroscience. Neuroinformatics 6, 175–194
2. Semantic Mediawiki platform: http://semantic-mediawiki.org.
3. BioMedGT: http://biomedgt.nci.nih.gov

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

The Role of Bio-Ontologies in Data-Driven Research: A Philosophical Perspective

Sabina Leonelli
ESRC Centre for Genomics in Society, University of Exeter, UK

Abstract

This project aims to reach a philosophical understanding of the role played by theory in the practices of data dissemination and re-use that characterise data-driven research. Bio-ontologies have the potential to play the epistemic role of theories in this context, insofar as they (1) express the knowledge underlying data-driven research and (2) guide such research towards future discoveries.

Introduction: Data-Driven Research

Up to the second half of the 20th century, biological data were largely produced as evidence to support given hypotheses. The activity of data gathering has since become increasingly automated and technology-driven. It is argued that the extraction of knowledge from automatically generated data may constitute a new, 'data-driven' approach to scientific methods.[1] This project examines the characteristics and significance of data-driven research from a philosophical perspective. If data-driven research constitutes a distinctive mode of knowledge production, how can it be characterised and how does it relate to hypothesis-driven research? To answer these questions, I focus on the epistemic role played by bio-ontologies in facilitating data re-use.

The Epistemic Roles of Bio-Ontologies

In their quality of classification systems, bio-ontologies fulfil a representational and a heuristic role in data-driven research.

(1) *Bio-ontologies constitute representational maps of the biological knowledge underlying data-driven research.*

Their aim is to represent what is currently known about biological entities or processes, in order to further the study of those entities and processes through coordination among research projects and exchange of relevant data, protocols and materials.[2] Thanks to their precisely defined terms, bio-ontologies explicitly formulate knowledge that is taken to be widely assumed, yet is usually dispersed across publications and research groups. These maps need not be true or universal;[3] rather, they need to capture the assumptions and practices underlying the successful sharing and re-use of biomedical data.[4]

(2) *Bio-ontologies constitute a network of theoretical hypotheses guiding data-driven research towards future discoveries.*

The definitions assigned to bio-ontology terms are modified depending on research developments. At the same time, the adoption of specific terms and definitions shapes how data are used in new research contexts.[5] They inform their users' understanding of how phenomena are defined beyond their own field.[6] They also define the evidential scope of the datasets classified and distributed through databases.[7]

Conclusion: Bio-Ontologies as Theories

The definitions used to disambiguate bio-ontology terms play the epistemic role traditionally assigned to hypotheses: they are descriptions of phenomena that are relied upon when planning new research, are open to further testing and modified on the basis of new findings. Bio-ontologies are biological theories suitable for the discovery of new facts.

Acknowledgments

This research is funded by the Economic and Social Research Council (ESRC). Thanks also to Mary Morgan and the Gene Ontology team.

References:

1. Bell G, Hey T and Szalay A. (2009) Beyond the data deluge. *Science* 323: 1297–1298.
2. Leonelli, S. (forthcoming 2010) Documenting the emergence of bio-ontologies. *History and Philosophy of the Life Sciences.*
3. Giere, RN. (1999) *Science Without Laws*. University of Chicago Press, p.27.
4. Leonelli S. (2008) Bio-ontologies as tools for integration. *Biological Theory* 3, 1: 8–11.
5. Leonelli S. (2009) Centralising labels to distribute data: The regulatory role of genomic consortia. In Atkinson P, Glasner P and Lock M (eds.). *The Handbook for Genetics and Society.* London: Routledge.
6. Leonelli S. (forthcoming 2010) Packaging data for re-use: Databases in model organism biology. In Howlett P and Morgan MS (eds). *How Well Do 'Facts' Travel.* Cambridge, MA: CUP.
7. Leonelli S. (2009) On the locality of data and claims. *Philosophy of Science* 76, 5.

Developing an Application Ontology for Annotation of Experimental Variables (Experimental Factor Ontology)

James Malone, Tomasz Adamusiak, Ele Holloway, Helen Parkinson
European Bioinformatics Institute, Cambridge, UK

Abstract

The Experimental Factor Ontology (http://www.ebi.ac.uk/efo) is an application focused ontology modelling the experimental factors in ArrayExpress[1]. The ontology has been developed to increase the richness of the annotations that are currently made in the ArrayExpress[1] repository, to promote consistent annotation, to facilitate automatic annotation and to integrate external data. The methodology employed in the development of EFO involves construction of mappings to multiple existing domain specific ontologies, such as the Disease Ontology and Cell Type Ontology.

Methodology

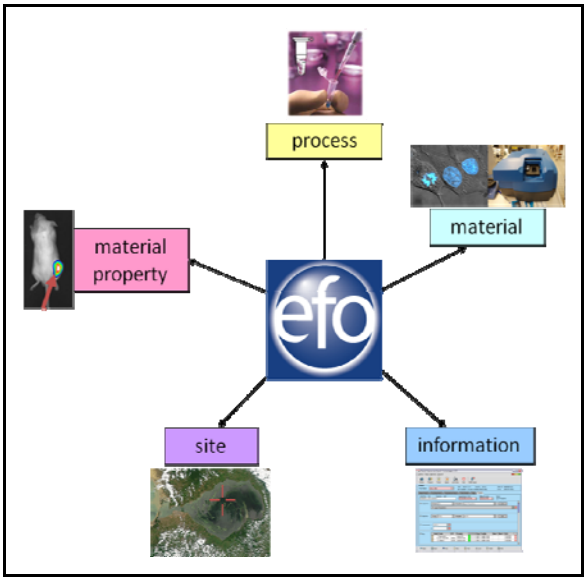


Figure 1. EFO axes.

EFO is organised around five main axes (Figure 1). Manual curation and semi-automated text mining are used to map EFO classes to other bioontology efforts that exist in the domain (Figure 2).

EFO has prototyped the use of agent technology to automate some aspects of ontology validation. All the supporting ontology tools are available on the http://efo.sf.net website.

The driving use case in developing EFO is based on the need for annotating experimental data in ArrayExpress[1]. These include, query expansion, data visualisation/integration, and nonsense detection.

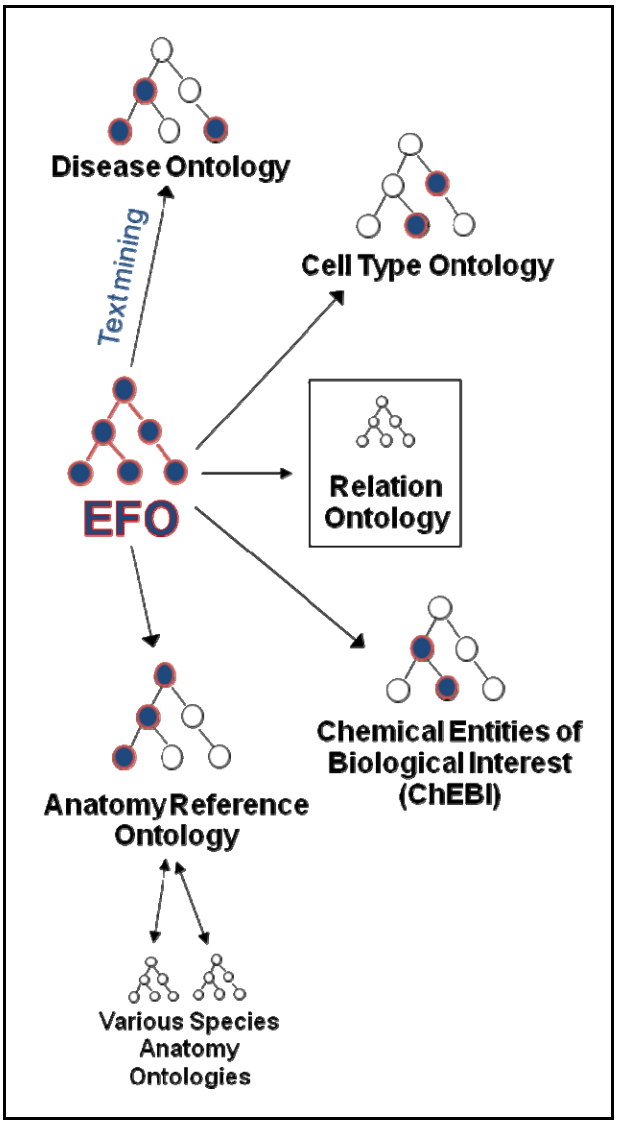


Figure 2. Mapping EFO classes to external resources.

Acknowledgements

The authors are funded in part by EC grants: FELICS (no. 021902), EMERALD (no. LSHG-CT-2006-037686), GEN2PHEN (no. 200754) and by EMBL.

References

1. Parkinson H, *et al*. ArrayExpress update – from an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Res. 2009 Jan;37(Database issue):D868–72.

Ontology Mapping of PATO to YATO for the Improvement of Interoperability of Quality Description

Hiroshi Masuya[1], Nobuhiko Tanaka[1], Kazunori Waki[1], Tatsuya Kushida[2], Riichiro Mizoguchi[3]
[1]RIKEN BioResource Center, Tsukuba, Japan; [2]NalaPro Technologies, Inc, Tokyo, Japan; [3]Osaka University, Ibaraki, Japan

Abstract

To facilitate broad interoperability for phenotype information between different ontological frameworks, we developed a reference ontology, PATO2YATO_ Quality, with the careful mapping of terms of PATO which is a quality ontology commonly used for biological phenotype annotation to the latest top-level ontology, YATO, which represents advanced modeling of quality-related concepts. This represents of interrelationships among quality-related concepts to provide fully integration of qualitative values and quantitative values obtained from phenotyping experiments and advanced representation of more detailed quality description.

Introduction

The description of qualities is a core issue for the integration of biological phenotype information. The Phenotype Quality Ontology (PATO) provides a practical basis for the integration of phenotype information across species. Typically, it is used for "entity plus quality" (E+Q) annotation of experimental parameters and parameter values[1]. However, there are multiple different methods of quality description recommended by the different top-level ontologies[2].

Recently, Yet Another Top-level Ontology (YATO) has been developed[3,4]. YATO represents not only ordinal quality descriptions covered by DOLCE and GALEN but also advanced quality descriptions not covered them. For the realization of more broad interoperability and advanced quality description of phenotypic quality using PATO terms, we developed a reference ontology called "PATO2YATO_Quality".

Results

We have worked out mapping of terms in the PATO2YATO framework by the careful examination with the helps of flags for subset of "attribute slim" and "value slim" embed in OBO format file of PATO as the remains of previous version. In this ongoing work, we currently have mapped about 500 terms of version 1.132 of PATO (quality_v1.132.obo) to YATO: (UpperOntology090112.ont: http://www.ei.sanken.osaka-u.ac.jp/hozo/onto_library/upperOnto.htm).

In PATO2YATO_Quality, quality-related concepts (dependent entities) are arranged as two hierarchies, "Quality type" and "Quality value", both of these are essential for an ontologically correct description of a change in quality. Furthermore, it allows systematic integration of numerical scales values and detailed representation such as <patient_1, diarrhea, yes> and <tail of mouse_1, short, severe>. We have started discussion with PATO developers to establish certain interoperability between two ontologies.

PATO2YATO_Quality is available at: http://www.brc.riken.jp/lab/bpmp/Ontologies/PATO2YATO/P2Y_Quality.html. Its OWL version, exported from Hozo, will be available soon.

Conclusion

We worked out mapping of PATO terms to the YATO framework, and successfully represented both the advanced meaning of each concept and the interrelationships among them.

Acknowledgements

With thanks to the Dr. Georgios V. Gkoutos for kindly sending us post-coordinated MP library, and for meaningful discussion.

References

1. Gkoutos GV, Green EC, Mallon AM, Blake A, Greenaway S, Hancock JM and Davidson D. Ontologies for the description of mouse phenotypes. Comp Funct Genomics. 2004:5:545–551.
2. Aranguren ME, Antezana E, Kuiper M and Stevens R. Applying ontology design patterns in bio-ontologies, Proc. of 16th International Conference:2008:LNAI 5268:7–16.
3. Mizoguchi R. Yet Another Top-level Ontology: YATO, Proceedings of the 2nd Interdisciplinary Ontology Meeting, 2009: in press.
4. Masuya H and Mizoguchi R. Toward fully integration of mouse phenotype information, Proceedings of the 2nd Interdisciplinary Ontology Meeting, 2009: in press.

Ontology Relating Human Neurodegenerative Disease to Associated Animal Model Phenotypes

Sarah M. Maynard[1], Lisa L. Fong[1], Stephen D. Larson[1], Asif Memon[1], Nicole Washington[2], Chris J. Mungall[2], Maryann E. Martone[1]
[1]University of California, San Diego, La Jolla, CA, USA
[2]Berkeley Lab, Lawrence Livermore National Labs, Berkeley, CA, USA

Abstract

We have developed a multi-scale ontology and knowledge base from literature and light and electron microscopic observations of animal models and human disease. We created a model for phenotype descriptions, formalized in Web Ontology Language (OWL), which draws entities from the Neuroscience Information Framework ontologies (NIFSTD, http:www.neuinfo.org/about/vocabularies.shtml) and Ontology of Phenotypic Qualities (PATO, http:obofoundry.org/egi-bin/detail.cgi?id=quality). The knowledge base has been loaded into the Open Biological Data Ontologies Database (http://www.berkeleybop.org/OBDUI) which statistically compares models and human disorders.

Introduction

Neurodegenerative diseases have a wide and complex range of biological and clinical symptoms. While neurodegenerative diseases share many pathological features, they each contain unique signatures in targeted cellular and subcellular structures. Animal models are key to translational research, yet typically replicate only a subset of disease features that may be only indirectly related to the human disease. Additionally, a given animal model may also map onto more than one condition. We employ formal descriptions of structural phenotypes associated with neurodegenerative disorders and animal models to provide a more effective means of matching animal models to diseases. We focus on multi-scale anatomical data, providing relationships among structural phenotypes at different scales.

Results

Creating ontologies for diseases is a significant informatics challenge because of the complex nature of disease[1]. Rather than focusing on a disease process, we focus phenotypes: any observable or measurable feature associated with an organism. Each phenotype is constructed from the combination of biological entities from the Neuroscience Information Framework (NIF, http://neuinfo.org) and qualities from the Ontology of Phenotypic Qualities (PATO) (Figure 1).

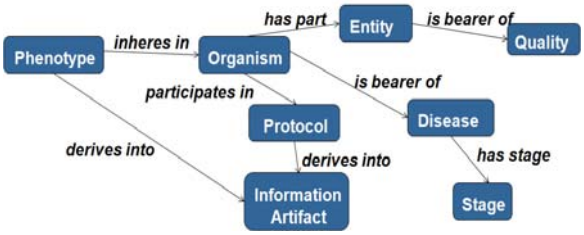


Figure 1. Phenotype Template.

We generated phenotype statements from observations made by basic and clinical researchers in literature and observations of animal models made from biological images. Human disease phenotypes are defined classes in the Neurodegenerative Disease Phenotype Ontology (NDPO, http://www.ccdb.ucsd.edu/NDPO/1.0/NDPO.owl). We collect instances in the Phenotype Knowledge Base (PKB, http://www.ccdb.ucsd.edu/PKB/1.0/PKB.owl).

Phenotypes were loaded into the Open Biological Data repository (OBD) which conducts statistical comparisons of phenotypes using information content and semantic similarity. For example, OBD captures the similarity between a human Lewy Body and an alpha-synuclein inclusion in a mouse.

Conclusion

Using NIF and PATO we have provided a template for phenotype descriptions that can be applied to neurological diseases. By using a consistent model for description, we can aggregate data from multiple animal models into a common data model (OBD), facilitating comparative analysis[2]. Through underlying ontologies, we can provide some of the necessary knowledge to bridge descriptions made in animal models from basic research and descriptions of pathological features in clinical preparations.

References

1. Gupta A, Ludascher B, Grethe JS and Martone ME. Towards a formalization of disease-specific ontologies for neuroinformatics. Neural Networks 2003;16:1277-1292.
2. Mungall CJ, Bruggner RV, Washington N and Lewis S. The OBD Database. *(in preparation)*

Towards a Modular Ontology for Annotating Structured Imagery Reports: Early Experiments in Bone and Joint Diseases

Sonia Mhiri[1,3], Sylvie Despres[2], Ezzeddine Zagrouba[3]
[1]University of Paris 5 - CRIP5, Paris, France
[2]University of Paris 13, LIM&BIO, Bobigny, France
[3]High Institute of Computer Science - SIIVA, Ariana, Tunisie

Abstract

We aim to build a modular ontology to assist radiologists to annotate their osteoarticular case reports. In this paper, we introduce the major outlines of our work: a description model of radiologists annotation viewpoints, the modular ontology building approach by reusing existing ontologies and early experiments with a prototype of a specific annotation tool.

Introduction

In medical imagery, considerable structured reports are produced in a digital form (doc, html, xml, dicom sr...). To take advantage of this expansion and to improve radiologists diagnostics, ontology-based annotation tools can be proposed.

Outlines of work in progress

A description model of radiologists annotations viewpoints:

Firstly, we consider for the radiologist the most relevant annotation viewpoints of its structured reports (textual information and images). We propose six abstract levels: patient context (name, old, weight, types of modality...), visual descriptors (color, texture, form, spatial characteristics...), technics (area of interest, segmented zone...), anatomy (organs structure...) pathological results (diseases, signs...) and recommendations[1].

Ontology reuse to build our modular ontology:

In medecine, ontologies are not widely used in annotation tools[2]. This is why we aim to build a modular ontology that will allow radiologists to annotate their imagery reports with the respect of the viewpoints description model. Because no existing ontologies can meet our needs, our approach consists in: the evaluation of an existing ontology in bone and joint domain, the modularization of this ontology in accordance with the proposed description model (Figure 1) and an adequate extraction tool, the development from scratch of the relevant missed knowledge, the enrichment of several parts by reusing existing ontologies and finally trying to unify six modules in a particular way.

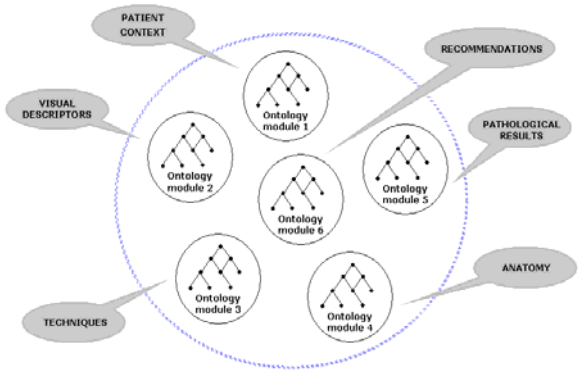


Figure 1. A sketch of our modular ontology

Early Experiments

We developed a prototype of a bilingual annotation tool which can load and visualize the modules extracted from a specific modular ontology. After charging an existing report (dicom sr or html), many functionnalities are offered: a multiaxial loading of owl ontologies (related to the six abstract levels), a visualization of interesting modules according to an arborescence view, a textual description for each ontology element, multi-selected text annotations, image panel tools for selection, showing previous existing annotations to help the radiologist, deleting or modifying existing annotations, etc.[1]

Perspectives

Several questions arise around our modularization building method. A classification of existing ontologies is imperative. Keeping a check on the reusing possibilities and then on their heterogeneities are also initial research tracks.

References

1. Mhiri S and Despres S. Ontology Usability via a Visualization Tool for the Semantic Indexing of Medical Reports (Dicom SR). 3rd Symposium for Medicine and Health Care (Graz). 2007; 412–417.
2. Mhiri S, Despres S and Zagrouba E. Ontologies for the semantic-based medical image indexing: An overview. IKE Conference (Las Vegas). 2008;311–317.

Phenoscape: Ontologies for Large Multi-Species Phenotype Datasets

Peter E. Midford[1], Paula Mabee[2], Todd Vision[3], Hilmar Lapp[3], Jim Balhoff[3], Wasila Dahdul[2], Cartik Kothari[3], John Lundberg[4], Monte Westerfield[5]
[1]University of Kansas, Lawrence, KS, USA; [2]University of South Dakota, Vermillion, SD, USA;
[3]US National Evolutionary Synthesis Center, Durham, NC, USA;
[4]Academy of Natural Sciences, Philadelphia, PA, USA; [5]University of Oregon, Eugene, OR, USA

Abstract

The Phenoscape project is developing ontologies and tools to integrate morphological and genomic data to address comparative questions in evolutionary biology. We are currently curating 81 publications describing ~5000 phenotypic characters in 4,000 species of Ostariophysian fishes, and will be making our database of ontology-based annotations concurrently with this meeting via a web-based interface at http://kb.phenoscape.org.

Introduction

Until recently, biological ontologies have either focused on single model organisms, or, like the Gene Ontology, attempted to span the tree of life.. Phenoscape is a project to develop ontologies and a database to describe the phenotypes of members of the the Ostariophysi, a large group (>9,000 species) of teleost fish. Ultimately, the database, ontologies and associated tools will allow us to apply reasoning to queries over zebrafish mutant phenotypes and "evolutionary" phenotypes across the Ostariophysi.

Ontologies

We have built two ontologies: the Teleost Anatomy Ontology (TAO) and the Teleost Taxonomy Ontology (TTO). We also constructed an ontology of taxonomic ranks and contributed terms to several existing OBO ontologies (e.g., Evidence Codes, PATO, etc.).The TAO is a multispecies ontology of anatomical terms. It was derived from, and is regularly synchronized with, the zebrafish anatomy ontology that is maintained by the Zebrafish Information Network (ZFIN). We have added, with input from the ichthyological community, over 400 terms since the TAO was cloned from the ZFIN anatomy ontology in September 2007. The TTO is an ontology of taxonomic names of groups within the teleost fishes. It includes all species and genera from the Catalog of Fishes database and additional taxa mentioned in papers we curated.

Curation of evolutionary phenotypes requires use and extension of the OBO phenotype ontology (PATO), evidence code ontology (ECO), relation ontology (RO), and spatial ontology (SO).

Curation

We have selected an initial set of 81 papers for curation, based on the availability of phenotype data in evolutionary character matrices with the added goal of covering as many Ostariophysan species as possible. Curation has been performed both by Phenoscape personnel and domain experts in the ichthyological community.

The curation process consists of identifying and requesting necessary additions to the ontologies, followed by annotating reported evolutionary phenotypes of species in the paper using the Entity-Quality (EQ) syntax. We will construct over eight million EQ annotations from the initial 81 papers.

EQ statements are constructed with the Phenex tool, an enhanced and extended version of Phenote. Phenote is also used to construct statements of homology (identity of structures in different species by common descent).

Knowledge-Base and Webservices

Our annotations are stored in a database based on the OBD database schema with the mutant phenotypes, genes and other data from the zebrafish community database (zfin.org). Because both data sets include ontology-based phenotypes, they can be integrated in the Phenoscape web interface. Examples of queries immediately possible are finding candidate genes underlying the evolution of morphological characters and searches to discover similar phenotypes among different taxa. A publicly available web interface and services will be available, concurrent with this meeting at http://kb.phenoscape.org.

Conclusion

The Phenoscape project has developed an ontology-based generalizable system for addressing questions that span the domains of developmental and evolutionary biology. It makes possible examining phenotype evolution at a large scale.

Acknowledgements

We thank NSF DBI 0641025, NIH HG002659 and the National Evolutionary Synthesis Center (NESCent) #EF-0423641 for funding.

Adding *Complex*-ity to the Protein Ontology

Darren A. Natale[1], Cecilia N. Arighi[2], Judith A. Blake[3], Carol J. Bult[3], Peter D'Eustachio[4,5], Gopal Gopinath[5], Cathy H. Wu[1,2]
[1]Georgetown University Medical Center, Washington, DC; [2]University of Delaware, Newark, DE; [3]The Jackson Laboratory, Bar Harbor, ME; [4]New York University School of Medicine; [5]Cold Spring Harbor Laboratories

Abstract

The Protein Ontology (PRO) provides a formal representation of protein objects within the OBO Foundry, consisting both of descriptions of these objects and of the relationships between them. Here we describe upcoming developments in PRO; namely, the inclusion of species-specific protein forms, expansion of the ID space, and—in conjunction with the Gene Ontology and the pathway databases Reactome and MouseCyc—the extension to protein complexes.

Introduction

OBO Foundry ontologies are organized along the dimensions of granularity (molecule to population) and relation to time (objects, qualities, processes). Within this scheme, the Protein Ontology is a representation of protein objects at the single molecule level of granularity, treating the protein molecules themselves rather than some property of the molecules (such as function, location, types of post-translational modification, etc.). Such properties are instead handled by other ontologies such as the Gene Ontology (GO)[1] and the Protein Modification Ontology (PSI-MOD);[2] PRO provides the objects to which such properties can be attached.[3,4] PRO encompasses a sub-ontology of proteins based on evolutionary relatedness (ProEvo), and a sub-ontology of the multiple protein forms produced from a given gene locus (ProForm) (Figure 1).

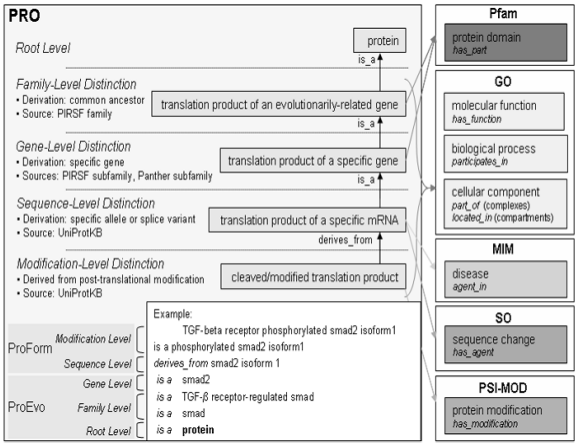


Figure 1. PRO organization and inter-ontology connections.

To date the curation of PRO focused on providing a deep hierarchy and annotation for proteins encoded by a small number of genes from humans and mice only, where the translation products from mRNA splice isoforms that were deemed equivalent in both species were given a single species-neutral term. PRO also did not address the associations that proteins make in living cells.

Recently the Protein Ontology Consortium hosted a workshop focusing on the user community and how best to maximize the usefulness and adoption of PRO. The outcome pointed to three needs: (1) a greatly expanded ID space for genes not deeply annotated, including those from other model organisms; (2) species-specific terms to supplement the species-neutral terms; and (3) terms denoting specific protein complexes to supplement the neutral terms in GO. In this presentation we describe our preliminary and forthcoming work to address these needs.

Acknowledgements

PRO is funded by NIGMS / NIH grant 1 R01 GM080646-01.

References

1. Blake JA, Harris MA. The Gene Ontology (GO) project: structured vocabularies for molecular biology and their application to genome and expression analysis. Curr Protoc Bioinformatics. 2008 Sep;Chapter 7:Unit 7.2.
2. Montecchi-Palazzi L, Beavis R, Binz PA, Chalkley RJ, Cottrell J, Creasy D, Shofstahl J, Seymour SL, Garavelli JS. The PSI-MOD community standard for representation of protein modification data. Nat Biotechnol. 2008 Aug;26(8):864-6.
3. Natale DA, Arighi CN, Barker WC, Blake J, Chang TC, Hu Z, Liu H, Smith B, Wu CH. Framework for a protein ontology. BMC Bioinformatics 2007; 8 Suppl 9:S1.
4. Arighi CN, Liu H, Natale D, Barker WC, Drabkin HJ, Blake J, Smith B, Wu CH. TGF-beta Signaling Proteins and the Protein Ontology. BMC Bioinformatics 2009; in press.

OBI: Ontology for Biomedical Investigations

The OBI Consortium, http://purl.obofoundry.org/obo/obi

Abstract

OBI, the Ontology for Biomedical Investigations, is being engineered by a set of domain experts encompassing a wide array of biomedical science disciplines. The scope of this ontological effort is to close a gap in coverage in resources available to annotate scientific experimental practice and make it coincide with the evidence-based biology paradigm.

Introduction

The OBI consortium endorses the OBO Foundry principles.[1] Those guidelines have positively impacted the work. First, by encouraging an open and inclusive approach, the OBI group proactively seeks partners and may act as an accretion point, avoiding work fragmentation. Second, by insisting on documentation, working practices are made explicit for a decentralized yet consistent development. OBI aims at representing various experimental processes (investigation, study, assay), the study design, the protocols and instrumentation used, the material used, the data generated and the type of analysis performed on the data. OBI supports the consistent annotation of biomedical experiments regardless of the particular field of study.

Results

OBI selected the Basic Formal Ontology (BFO)[2] as its upper-level ontology, and as a result is being developed following 3 main axes: *bfo:Process* covering assays and information processing, *bfo:MaterialEntity* encompassing instrument and other materials and *bfo:DependentContinuant*, (with children such as quality, role and disposition) which holds entities used to possibly qualify elements of the first two dimensions. OBI uses the Ontology Web Language (OWL)[3] and the Protégé editor[4] as development environment and is organized as a series of working groups tackling specific sub-domains.

Procedures have been devised to ensure consistent work across branches. Thus, OBI has agreed on a naming convention for representational artifacts, a minimal set of metadata to supply when submitting terms or creating classes and methods both for merging branch outputs and for cross referencing OBO foundry sister ontologies (e.g. CHEBI, CL, GO). OBI is currently being evaluated against competency questions and use cases collected from its members.

Biomedical experimental processes can involve numerous sub-processes, where each step can involve various material entities e.g., whole organisms, organ sections, cell culture, cell pellets, devices. Material entities realize distinct roles given the context of the process they are used in e.g. study subject role, host role, specimen role, patient role; and distinct functions e.g. measuring, separating, environment controlling. Use cases are employed to demonstrate how to model entities and their relations in OBI in order to describe experimental processes such as a blood glucose measurement assay, or a vaccine protection experiment.

Conclusion

OBI provides an approach to represent biological and clinical investigations in an explicit and integrative framework, which facilitates computational processing and semantic web compatibility.

Acknowledgements

The OBI consortium is (in alphabetical order): Ryan Brinkman, Bill Bug, Helen Causton, Kevin Clancy, Christian Cocos, Mélanie Courtot, Dirk Derom, Eric Deutsch, Liju Fan, Dawn Field, Jennifer Fostel, Gilberto Fragoso, Frank Gibson, Tanya Gray, Jason Greenbaum, Pierre Grenon, Jeff Grethe, Yongqun He, Mervi Heiskanen, Tina Hernandez-Boussard, Philip Lord, Allyson Lister, James Malone, Elisabetta Manduchi, Luisa Montecchi, Norman Morrison, Chris Mungall, Helen Parkinson, Bjoern Peters, Matthew Pocock, Philippe Rocca-Serra, Daniel Rubin, Alan Ruttenberg, Susanna-Assunta Sansone, Richard Scheuermann, Daniel Schober, Barry Smith, Larisa N. Soldatova, Holger Stenzhorn, Chris Stoeckert, Chris Taylor, John Westbrook, Joe White, Trish Whetzel, Stefan Wiemann, Jie Zheng.

References

1. Smith B, *et al.*, Nature Biotechnology 25, 1251–1255, 2007.
2. Basic Formal Ontology (BFO), http://www.ifomis.org/bfo
3. Ontology Web Language (OWL), http://www.w3.org/TR/owl-guide/
4. Protégé, http://protege.stanford.edu

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

Virtual Fly Brain: An Ontology-Linked Schema of the *Drosophila* Brain

David Osumi-Sutherland[1], Mark Longair[2], J. Douglas Armstrong[2]
[1]FlyBase, Cambridge University, Cambridge, UK
[2]School of Informatics, Edinburgh University, Edinburgh, UK

Abstract

Drosophila neuro-anatomical data is scattered across a large, diverse literature dating back over 75 years and a growing number of community databases. Lack of a standardized nomenclature for neuro-anatomy makes comparison and searching this growing data-set extremely arduous.

A recent standardization effort[1] has produced a segmented, 3D model of the Drosophila brain annotated with a controlled vocabulary. We are formalizing these developments to produce a web-based ontology-linked atlas.

This well-defined gross anatomy provides a substrate for defining neuronal types in the ontology according to where they fasciculate and innervate. Neuronal types are also classified in the ontology according to neurotransmitter released, lineage and function. The resulting ontology provides both a vocabulary for annotation and a means for integrative queries of neurobiological data.

Introduction

The *Drosophila* brain can be crudely described as consisting of tracts (bundles of axons and dendrites lacking synapses) and neuropil domains (discrete regions generally containing whole terminal arbors, in which axons and dendrites synapse with each other). The names of tracts and domains and details of their gross connectivity are currently being standardized in an atlas based on labeled 3D image stacks[1].

Common criteria for classification of neurons include: which tracts their axons are bundled (fasciculated) in; which neuropil domains they innervate; their morphology; their function; their circuit position; what neurotransmitter they release; their lineage. Large amounts of data pertaining to these classifications exist, but have not been standardized or integrated in any query-able resource.

Results

Terms referring to elements of the gross architecture of the *Drosophila* brain are defined with reference to a standard 3D segmented model generated from image stacks of the *Drosophila* brain. See http://fruitfly.inf.ed.ac.uk/brain/ for online tools to explore this reference brain and its relationship to terms in the Drosophila anatomy ontology.

With the gross architecture of the brain defined in this way, we can use the resulting terms to define neuronal classes. General mereological relationships (part-hood, connectedness) are useful in defining relations between neuronal classes and gross anatomy, but are not sufficient to capture biologically important details (e.g.- fasciculation, innervation). We are developing relations to capture this information. For example:

fasciculates_with – relation between a neuron and an axonal or dendritic bundle that its axon or dendrite is part of (implies mereological **overlap**[2])

innervates – relation between a neuron and a structure in which it has a synapse (implies mereological **overlap**[2]).

We are also developing methods to record neurotransmitter released (classified according to the chemical ontology, CHEBI) and function (classified according to gene ontology biological process terms) and systems for asserted classification of neurons according to morphology and circuit position. We are recording lineage using existing relations.

Conclusions and Aims

The combined ontology and models of the *Drosophila* brain will provide the basis for an online atlas providing links between images and ontology terms and allowing OWL-DL based queries for neuronal classes using the criteria described here.

References

1. Strausfeld N, *et al*., 2007 http://www.hhmi.org/janelia/conf-002.html
2. Smith B. *et al*., 2005 Genome Biol. 6(5): p. R46.

A Bayesian Hierarchical Model to Derive Novel Gene Networks from Gene Ontology Fingerprints

Tingting Qin, Lam C. Tsoi, Andrew Lawson, Jim W. Zheng
Medical University of South Carolina, Charleston, SC, USA

Abstracts

We developed a Bayesian hierarchical model to identify gene networks based on the similarity score generated from comparing the gene ontology fingerprints of gene pairs. Genes in this network were assumed to have similar biological functions that can be indicated by their ontology fingerprints. Our results indicate that different pathways show consistent score threshold that allow us to distinguish biological relevant gene—gene connections in the network.

Methodology

The enrichment of each gene ontology (GO) term among PubMed abstracts linked to a human gene was computed to construct the ontology fingerprint for all human genes[1]. The biological relevance between every gene pair was then measured by a similarity score generated from comparing the ontology fingerprints of the two genes. We developed a Bayesian approach to model the biological relevance of the similarity score in order to develop gene networks, and we used WinBUGS to compute the posterior distributions of the parameters in the model (Figure 1). We applied our model to evaluate genes in the KEGG pathways[2] in order to study the properties of gene networks within biological pathways. The log gene-gene similarity score (y) was modeled as a mixture normal distribution representing similar and dissimilar genes, as defined by threshold c_k. The jump parameter α_i represents the biological coherence of gene i within biological pathway. μ^* was estimated as the mean score of dissimilar genes.

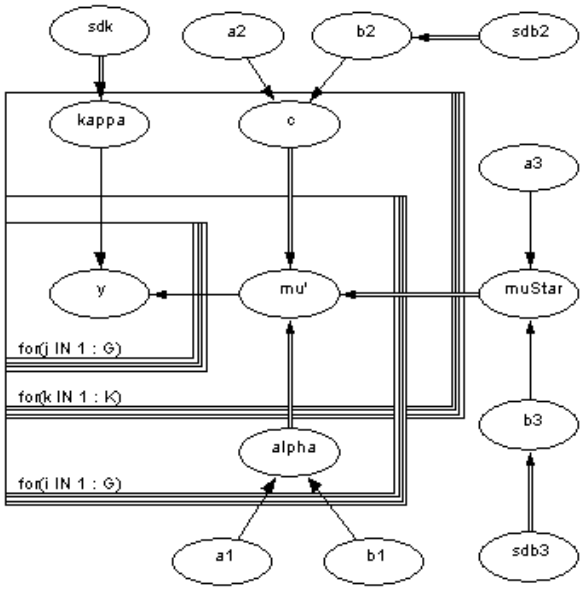


Figure 1. Illustration of Bayesian model

Results

We were able to distinguish similarity scores among genes belonging to the same KEGG pathway from those among randomly picked, irrelevant genes. Moreover, the results show that there is a trend of consistent score threshold across different biological pathways, indicating that there might be a standard threshold to separate biological coherent genes from dissimilar genes by using ontology fingerprints. As a result, we may be able to utilize this information to infer new genes for biological pathways. Different posterior α values were also observed for different genes, which could give us insight about the degree of biological coherence of a particular gene in the pathway, as well as the role or biological importance of the gene in gene networks. The ontology fingerprints can then be used to further identify the biological relevance of each gene, known or inferred, in the pathway.

Conclusion

Applying a Bayesian hierarchical model to analyzing the similarity scores derived from comparing two genes' ontology fingerprints provide a novel approach to investigate gene networks. Our model suggests a consistent threshold of the similarity scores among all KEGG pathways, which could be used as an indicator to distinguish the genes within KEGG pathways from those that are irrelevant.

References

1. Tsoi LC, *et al*. Evaluation of genome-wide association study results through development of ontology fingerprint. *Bioinformatics*. 2009, 25(10):1314–20
2. Ogata H, *et al*. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 1999; 27(1):29–34

Letting the Cat out of the Bag: OBO and the Semantic Web

John E. Rose
The Winooski Foundation, Cortland, NY, USA

Abstract

The tools developed to work with an expanding number of biomedical ontologies are all directed "inward," back toward the scientists The Semantic Web is an effort to assign meaning to data available on the internet resulting in a universal platform of knowledge exchange. For the Open Biomedical Ontologies to become a part of the Semantic Web, a more "outward-looking" tool is required. A prototype, AmiGA, based on 21 ontologies with over 600,000 terms has been constructed.

Introduction

Two primary approaches to combine the structure of the Semantic Web with content on the scale of Wikipedia[7] have developed: (1) to start over using a structured vocabulary to build content with machine-accessible semantics or (2) to extract structure from existing content. Semantic MediaWiki[1] represents the first approach, and dbpedia[2] the second.

A third approach, lying somewhere between these extremes is also possible: create a scaffold based on structured vocabularies (e.g., OBO[3] ontologies) and let users add to it in a wiki format.

Implementation

User input is stored in a separate database from that of the ontology information using a schema similar to, but much simpler than that of MediaWiki. Users are only allowed to edit the "Details" section of each web page. Definitions and navigation graphs are automatically generated from the OBO-based database and are not accessible to users.

Semantic MediaWiki has a feature called "semantic data" that allows embedding property-value pairs on a page and accessing them on other pages. AmiGA uses the OBO feature "property_value" in this way.

Each term in the database has a web page with a navigation graph generated by Graphviz[4]. The graph displays all terms that are related to the current term directly by an OBO relation. The user may browse the various hierarchies imposed by the relations.

Similarly to Wikipedia, so-called "disambiguation" pages are constructed for identical term names in different ontologies. These pages contain the definition and the navigation graph for each term.

AmiGA currently searches OBO Term names, Term definitions and user-edited content and returns results for each in separate panels. Searches can be restricted to the *is_a*, *develops_from* and *part_of* hierarchies.

Problems

Disambiguation pages bring up issues regarding the various ontologies. Some of the relate a term to the corresponding GO term via an *xref*, yet have a different, yet equivalent, definition. The OBO file format allows terms with no *name* tag, so that such terms will have to relate directly to the GO. However, there is still a need for disambiguation pages, because some ontologies may use the same term, but with unrelated definitions.

Quite a few new time-related relations are proposed. Adopting all of these would lead to a proliferation of semantically identical, but lexically different terms. The currently used *develops_from* relation provides a surprisingly substantial amount of structure.

There is no standard way to get the name of the ontology from the database. Sometimes the only indication of an ontology's name is in the file name.

References

1. Wikipedia: the Free Encyclopedia, http://en.wikipedia.org/wiki.
2. Auer S and Lehmann J. PDF Document: What have Innsbruck and Leipzig in common? Extracting Semantics from Wiki Content. In Franconi, *et al*. (eds.), Proceedings of European Semantic Web Conference (ESWC'07), LNCS 4519, pp. 503–517, Springer, 2007. http://wiki.dbpedia.org.
3. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, OBI Consortium, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL and Lewis S. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology. 2007;25: 1251– 255.
4. Junger M and Mutzel P. Graph Drawing Software (Mathematics and Visualization). Symposium on Graph Drawing 2001, Vienna.

Ontology Integration: Bridging Bioinformatics to Clinics

Sirarat Sarntivijai[1], Yongqun He[1,2], Matthias Kretzler[1,3], Brian D. Athey[1,4]
[1]The National Center for Integrative Biomedical Informatics and the Center for Computational Medicine and Bioinformatics, [2]Department of Microbiology and Immunology, [3]Department of Internal Medicine, and [4]Department of Psychiatry, University of Michigan, Ann Arbor, MI, USA

Abstract

Most biomedical ontologies have been created by individual laboratory of origin and thus promote information diversity even when describing similar elements in the same domain. Ontology mapping remains a challenging issue. While annotations are immediate use of ontology, there is increasing awareness of computations for translational informatics. Knowledge discovery from the proposed model of ontology integration is demonstrated in vaccine ontology mapping, and identification of disease factors of Diabetic Nephropathy. This work is funded by NIH grant U54 DA021519 for the NCIBI and R01 LM008106.

Introduction

Biomedical ontology (bio-ontology) was first created at the laboratory of origin out of the needs for systematic annotation. Therefore, computing with logical reasoning embedded in individual bio-ontology can be challenging due to the divergence of individualism, especially when mapping multiple bio-ontologies for knowledge discovery[1]. While such reusability and interoperability for knowledge transfer and discovery should be promoted, working with multiple bio-ontologies requires a sophisticated operating model that can overcome the issues of structural definition discrepancy of ontologies describing similar elements in the same domain, incomplete and error-prone information within an ontology, and bridging across different information layers. The framework proposed here utilizes graph matching theory, natural language processing, and ontology alignment to create a novel approach of ontology integration that drives ontology processing forward from annotations to computations, to translations for the next-generation translational informatics.

Results

By mapping and integrating bio-ontologies of different biological layers from molecular genotype to molecular phenotype to clinical phenotype, we demonstrate that bio-ontology processing plays an important role in knowledge discovery. More than 10 bio-ontologies or controlled biomedical vocabulary systems are used in our approach (Figure 1). Examples of use cases given in this study are integration of vaccine ontology[2] in health care research, and using ontology integration to identify key disease factors of Diabetic Nephropathy.

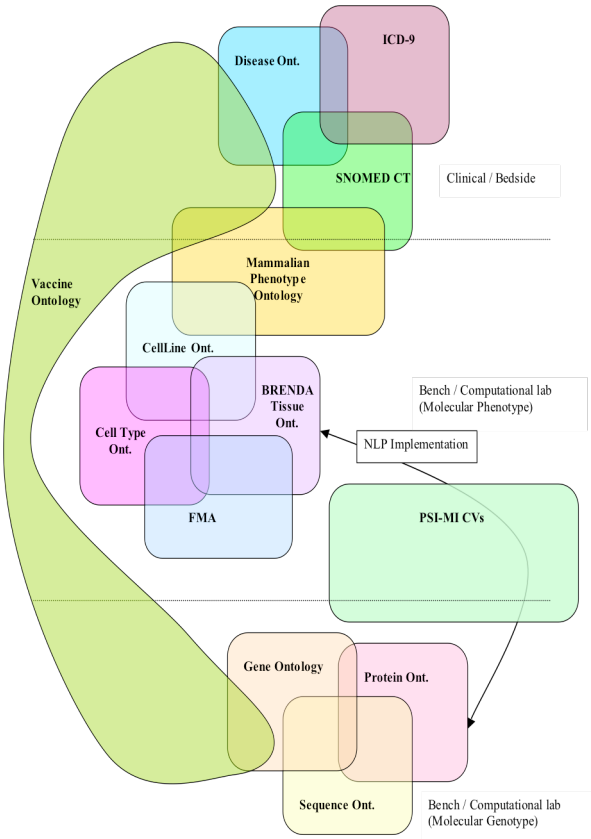


Figure 1. Overlaying ontologies for knowledge discovery by ontology integration

Acknowledgements

This research is funded by NIH grant U54 DA021519 for the National Center for Integrative Biomedical Informatics and R01 LM008106.

References

1. Bodenreider O. Biomedical Ontologies in Action: Role in Knowledge Management, Data Integration and Decision Support. Yearb Med. Inform. (2008);:67–79.
2. Vaccine Ontology (VO): http://www.violinet.org/ vaccineontology. Last accessed April 10, 2009.

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

BFO/DOLCE Primitive Relation Comparison

A. Patrice Seyed
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA

Abstract

This paper examines the primitive relations (dependence, quality, and constitution) of the BFO and DOLCE upper ontologies, employed in developing domain ontologies of the biomedical sciences. The strengths in both upper ontologies are examined, which sets a framework for developing a common upper ontology that utilizes the assets of both.

Introduction

Cross-domain reasoning with data can be achieved through successful integration of domain with upper ontology types. Basic Formal Ontology (BFO)[1] and Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE)[2] are two widely used upper ontologies, especially for the development of ontologies in the biomedical sciences. BFO is based in realism, whose primitive relations are defined in the Relation Ontology (RO)[3]. DOLCE attempts to capture ontological categories presupposed by natural language and commonsense. The purpose of this paper is to provide a comparison of the primitive relations defined for BFO and DOLCE. Note that BFO uses the terms continuant and occurrent – while DOLCE uses endurant and perdurant – to denote entities that are wholly present in an instant of time and those that have temporal parts, respectively.

Dependence and Quality

Specializations of dependence include 'function of' and 'role of', the domain of which are internally and externally grounded realizable entities, respectively[4]. This is not the case for 'quality of', since qualities are not dependent on a process or activity to be manifested. DOLCE defines 'quality of' as a relation between a quality, and another quality, endurant, or perdurant. As with BFO, a quality cannot be present unless the particular it inheres in is also present. However unlike BFO, in DOLCE this relation can hold between two qualities, or between qualities and occurrents. BFO observes that describing qualities as inhering in events is convenient for explanation, but represents knowledge and not ontological reality. DOLCE also includes a relation 'quale of' holding between qualities and qualia, the latter of which only exists as a reflection of human cognition.

Constitution

Constitution is a more general sense of composition – which denotes 'is made of' – and helps describe particulars that are naturally in flux[5]. *x* constitutes *y* when there are properties of *x* (e.g., heartbeat) which are accidental to *x* (e.g., body) but essential to *y* (e.g., person)[6]. DOLCE includes constitution as a primitive relation, but according to BFO two things cannot exist at the same time and space. However a relation in BFO similar to constitution is 'role of'.

Conclusion

BFO holds that qualities can only inhere in continuants, and entities that are only available through the human perceptual lens are not bona fide, falling to subjectivism. DOLCE applies the notion of quality more liberally, and allows for objects of thought to be basic units of its ontology. Future work should investigate how entities of a conceptualist-centric upper ontology can fit into the theory and hierarchy of a realist one, in a manner that does not contradict its philosophical underpinnings.

Acknowledgements

With thanks to Dr. Barry Smith for discussions on the primary topics of formal ontology, Kelly Graham for discussions on biomedical domain entities, and Joel Potter for his input in preparing this paper.

References

1. Smith B. The basic tools of formal ontology. In Formal Ontology in Information Systems. 1998; 19–28.
2. Masolo C, Borgo S, Gangemi A, Guarino N and Oltramari A. WonderWeb deliverable D18 ontology library. 2003.
3. Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL and Rosse C. Relations in biomedical ontologies. Genome Biology, 2005; 6(5).
4. Arp R and Smith B. Realizable entities in basic formal ontology. ISMB Workshop on Bio-Ontologies, 2008.
5. Simons P. Parts: A study in ontology. Oxford Clarendon Press. 1987.

Multiple Ontologies for Integrating Complex Phenotype Datasets

Mary Shimoyama, Melinda Dwinell, Howard Jacob
Medical College of Wisconsin, Milwaukee, WI, USA

Abstract

There has been an emergence of multiple large scale phenotyping projects in the rat model organism community as well as renewed interest in the ongoing phenotype data generated by thousands of researchers using hundreds of rat strains worldwide. Unfortunately, this data is scattered and is neither described nor formatted in a standardized manner. A system to integrate complex phenotype data from multiple sources and facilitate data mining and analysis is being developed using multiple ontologies.

Introduction

The potential value of integrating phenotype data from multiple sources (different laboratories, varying techniques to measure similar phenotypes, multiple strains) is enormous. Presented here is a data integration system for complex phenotype data from both large-scale and individual experiments and the taxonomy and ontologies that provide the backbone of this format. RGD along with Mouse Genome Informatics (MGI) (Blake et al, 2009) and the Animal QTL Database (Hu and Reecy, 2007) is developing a Vertebrate Trait Ontology to represent morphological states and physiological processes to be used to annotate quantitative trait loci (QTL) and other data. RGD has also used the Mammalian Phenotype Ontology (Smith et al, 2005) for several years to indicate the relationship of genomic elements to abnormal phenotypes. The Vertebrate Trait Ontology represents what is being assessed, and the Mammalian Phenotype Ontology represents the conclusion that was made. The system presented here represents what was done to measure the trait in order to reach the conclusion. Because of the close relationship among these ontologies, care is being taken to ensure compatibility and similarity in structure using the phenotype properties in the Phenotypic Quality Ontology (PATO) for guidance. (http://www.bioontology.org/wiki/index.php/PATO:Main_Page)

Data Format and Ontologies

Standardization of data types and relationships used to define the phenotype experiment and resulting data, and the ontologies to be used to standardize descriptive fields are being developed. For phenotype data, the major informational components include Researcher, Study, Experiment, Sample, Experimental Conditions and Clinical Measurement. A Rat Strain Taxonomy has been developed to standardize this information and provide the relationships among strains to allow investigators to retrieve and analyze phenotype data for strains that are related genetically. Two important aspects of a phenotype measurement include 1) what was measured and 2) how it was measured. The Clinical Measurement Ontology and the Measurement Method Ontology are being developed to standardize this information. In addition an Experimental Conditions ontology is under construction to allow integration of data measured under various conditions.

Pilot Study Results

Cardiovascular and biochemistry phenotype data from two major datasets have been integrated using the Rat Strain Taxonomy and the three phenotype related ontologies. A prototype data mining tool http://rgd.mcw.edu/rgdweb/ has also been developed that provides the user with options to begin a search with strains or any of the ontologies and make subsequent filter choices from the other ontologies. Choices presented to the user are restricted to those for which data is available and query tracking functions are provided to alert the user to the number of results being returned and the query choices made.

References

1. Blake JA, Bult CJ, Eppig JT, Kadin JA and Richardson JE. Mouse Genome Database Group, 2009Nucleic Acids Res. Jan;37:D712–9.
2. Hu ZL and Reecy JM. Animal QTLdb: Beyond a repository. A public platform for QTL comparisons and integration with diverse types of structural genomic information, 2007, Mamm Genome, Jan;18(1):1–4.
3. Smith CL, Goldsmith CA and Eppig JT. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information, Genome Biol. 2005 6(1):R7.

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

Modeling Cardiac Rhythm and Heart Rate Using BFO and DOLCE

Lynda Temal, Arnaud Rosier, Olivier Dameron, Anita Burgun
U936 INSERM University of Rennes 1, IFR140, Rennes, France

Abstract

This paper presents an application ontology for modeling cardiac rhythm and its anomalies such as tachycardia and bradycardia. We use BFO and DOLCE as ontological reference framework in order to compare their impact on ontology design.

Introduction

Managing cardiac rhythm disorders usually involves implanted cardiac devices (artificial pacemakers and/or implantable cardioverter-defibrillators) to treat arrhythmias (bradycardias or tachycardias). Such devices send many remote alerts about arrhythmias to physicians, who have to assess their relevance and emergency level. In the AKENATON project, we aim at improving alert management by shifting from strictly device-centered follow-up to perspectives centered on the patient. This requires reasoning capabilities relying on domain knowledge. We designed an application ontology based on the following principles: *i)* the ontology has an applicative purpose and will be used to reason on real data. *ii)* we used a foundational ontology as a reference framework to guarantee ontological commitment, competency, and reasoning capabilities. We model cardiac rhythm and heart rate, and compare the implications of using BFO[1] and DOLCE[2] as upper-level ontologies.

Results: Use Case and Ontology Design

Patients may present different types of arrhythmia whose caracterization refers to the cardiac frequency (*heart rate* may be *fast* or *slow*) during the arrhythmia episode.

Atrial fibrillation is usually irregular and its mean frequency is referred to as arrhythmia heart rate. *Ventricular tachycardia* is typically regular although it may present faster or slower parts during the same episode. The fastest frequency is then referred to as the episode frequency. Knowledge about *Arrhythmia* and its inherent *Heart rate* together with other clinical information are necessary to assess the associated risk level. Since frequency is an important piece of information in both cases, it is important to model it correctly in our ontology. This work focuses on representing the taxonomical positions and the relations involving *Heart*, *Arrhythmia* and *Heart rate*. We compared how BFO (Figure 1) and DOLCE (Figure 2) influenced the ontology design, and studied to what extent both met our application requirements.

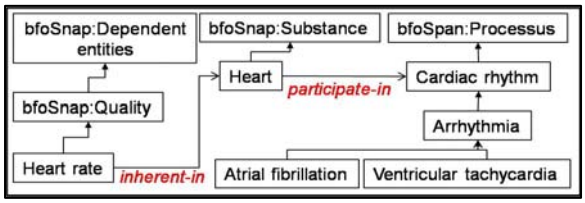


Figure 1. Cardiac rhythm and heart rate in BFO

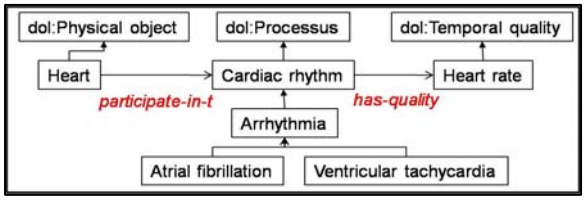


Figure 2. Cardiac rhythm and heart rate in DOLCE

Discussion and Conclusion

According to its *realist* approach, BFO (Figure 1) uses no qualities for *Perdurants* (i.e. entities that unfold in time). *Heart rate* is therefore not inherent in an *Arrhythmia.* Instead, *Heart rate* is inherent in *Heart* which participates in such *Arrhythmia*. Thus, determining whether an Arrhythmia is fast or slow according to a heart rate is not straightforward from BFO. Conversely, DOLCE, because of its cognitive bias (Figure 2) recognizes qualities associated with Perdurants. *Heart rate* is inherent in *Cardiac rhythm* which has for participant a *heart*. Hence, DOLCE allows to associate for each kind of arrhythmia the appropriate heart rate with the exact semantic of how the measure is made. The heart rate of an atrial fibrillation is the mean frequency, whereas the heart rate of a ventricular tachycardia is the fastest frequency of the episode. Thus, DOLCE is directly suitable for characterizing the exact semantic associated to the frequency of different kinds of cardiac rhythm. This property is used for inferring when arrhythmia is fast or slow, which is necessary for computing an associated risk level.

References

1. Grenon P and Smith B. SNAP and SPAN: Prolegomenon to geodynamic ontology. In Spatial Cognition and Computation, 2004.
2. Masolo C, Borgo S, Gangemi A, Guarino N and Oltamari A. WonderWeb deliverable D18 ontology library (final). Tech. report, ISTC-CNR, 2003.

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

NPO: Ontology for Cancer Nanotechnology Research

Dennis G. Thomas, Rohit V. Pappu, Nathan A. Baker
Washington University in St. Louis, St. Louis, MO, USA

Abstract

We present the design and development of NanoParticle Ontology (NPO). The ontology is implemented in the Ontology Web Language (OWL). The domain terms in NPO currently represent entities, which describe knowledge about physical, chemical, and functional properties of nanoparticles characterized in cancer nanotechnology research. The upper-level of NPO is formed using terms from the Basic Formal Ontology (BFO).

Introduction

In cancer nanotechnology research, there are diverse types of nanoparticles being developed and tested for applications in cancer diagnostics and therapeutics (NP-CDTs). These nanoparticles are diverse in their chemical composition, properties and application. The chemical composition of these nanoparticles can be varied in many combinatorial ways, which result in the development of as many types of nanoparticles. Small variations in the chemical composition cause drastic changes in the physical, chemical and functional properties of nanoparticles. Experiments performed to characterize the properties of these nanoparticles generate large volumes and diverse types of data. To efficiently share and use this data, and to further the application of nanotechnology to cancer treatment, supports for a common vocabulary and informatics methods are required.

We have developed an ontology, called the **N**ano**P**article **O**ntology (NPO), to provide a common vocabulary and the knowledge framework for enabling interdisciplinary discourse, and annotation of NP-CDT data in order to facilitate the sharing and semantic integration of data for reuse, analysis and inferencing of the data.

Results

We developed the NPO using well-defined design principles in OWL. Public releases of NPO are available through BioPortal (http://tinyurl.com/npo-bioportal). The current version (2009-04-02) of the NPO contains 919 terms and 21 associative relationships. The domain terms in the NPO were first obtained from the literature and other controlled vocabularies / ontologies (e.g., GO, ChEBI, NCI Thesaurus). These terms were organized into a taxonomic hierarchy starting with BFO terms at the upper-level of the NPO (see Figure 1). The domain terms represent different types of entities related to the description of NP-CDTs. These include entities which describe: (1) material entities that are synthesized, characterized and distinguished at the nanoscale (1-100 nm) size range; (2) material entities that are distinguished at the molecular structure level; (3) physical sites in a material entity; (4) surface of a material entity; (5) quality or property inhering in a material entity; (6) role of material entity at the molecular level; (7) type of stimulus for activating nanoparticle function, and response to stimulus; (8) tumor targeting methods; (9) functions of molecular entities that are realized as processes; (10) biological processes; and, (11) chemical linkages and interactions in a nanoparticle.

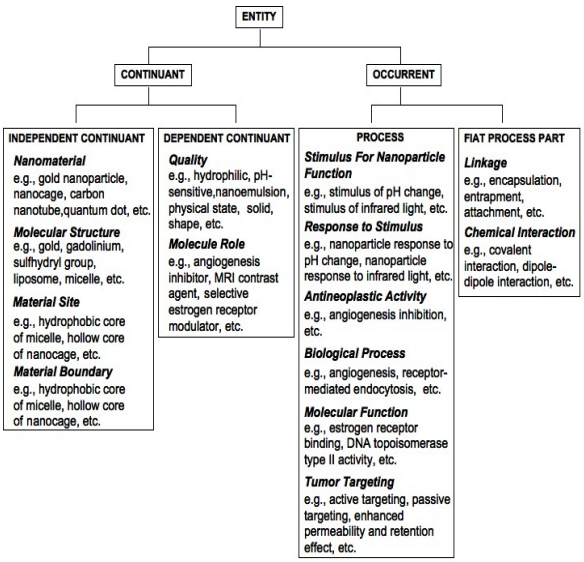


Figure 1. Example showing the BFO classification of domain terms

Conclusion and Future Directions

We have laid the foundation for future growth of NPO. Future work involves curation of existing terms and extension of NPO for supporting caNanoLab (http://gforge.nci.nih.gov/projects/calab/) database curation activities and data annotation.

Acknowledgements

Work funded by NIH grants U54 CA119342 and U54 HG004028. The authors would like to thank Daniel Rubin, Sharon Gaheen, Liz Hahn-Dantona, Frank Hartel, and Gilberto Fragoso for their helpful comments.

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

OCRe: An Ontology of Clinical Research

Samson W. Tu[1], Simona Carini[2], Alan Rector[3], Peter Maccallum[4], Igor Toujilov[5], Steve Harris[6], Ida Sim[2]
[1]Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA
[2]Division of General Internal Medicine, UCSF, San Francisco, CA, USA
[3]Department of Computer Science, University of Manchester, Manchester, UK
[4]Cancer Research UK Cambridge Research Institute, Cambridge, UK
[5]University College London Cancer Institute, London, UK
[6]Oxford University Computing Laboratory, Oxford, UK

Abstract

Querying data and meta-data across clinical trials and observational studies is difficult because of the lack of semantic and terminology standards for describing the design and analysis of human studies. The Ontology of Clinical Research (OCRe) is an ontology created to fulfill this need. OCRe allows the indexing of human studies across multiple study designs, interventions/exposures, outcomes, and health conditions. With such indexing, investigators interested in the evidence pertaining to a particular question (e.g., what is the effect of A on B in people with C) will be able to locate relevant research studies more easily across disparate data sources.

Human studies – encompassing both interventional and observational studies – are the most important source of evidence for advancing health science. These studies are expensive, logistically complex, and labor intensive to design, perform, and analyze. Important tasks for clinical and translational research include searching for studies that involve particular designs, interventions, or outcomes. Such queries are currently difficult to execute because there is no standard terminology or information model for the design of human studies and because clinical terms are not standardized across studies. To address these difficulties, we created the Ontology of Clinical Research (OCRe), a formal OWL ontology that represents the entities and relationships related to the design and analysis of human studies.

We conceptualize a study as a real-world entity whose properties and components parts evolve during the life cycle of the study. At the design stage, the artifacts of study consist mostly of documents (i.e., informational entities) that spell out the scientific hypothesis being studied, the design of the study, and planned activities of the study. At the execution phase, participants of a study carry out activities that result in a body of collected data. In the analysis phase, investigators transform the data and perform statistical analysis on them, resulting in publications and other artifacts.

The current OCRe ontology focuses on the design stage of studies. It is a modular ontology of clinical investigation that includes (1) a representation of the structure of human studies and associated entities (e.g. persons and organizations), (2) informational entities, such as study protocols, eligibility criteria, the specification of outcome variables and the statistical methods used to analyze them, (3) terms for characterizing and classifying study designs (e.g., how control groups are defined helps to characterized a parallel group study), and (4) bindings to standard terminologies, such as SNOMED CT and the NCI Thesaurus. We reused ontologies and information models that have already covered relevant domains. For example, the modeling of the schedule of activities in OCRe is imported from BRIDG.[1] The objective of having such a rich model is to permit queries across multiple types of studies that cannot now be performed.

OCRe has been subjected to initial formative evaluation, in which we annotated published clinical studies with OCRe terms and verified that we can query the repository of studies to select for studies that satisfy specific criteria. This has included cancer clinical trials annotated as part of the UK Medical Research Council funded CancerGrid project (http://www.cancergrid.org). Within the US National Institutes of Health CTSA Human Studies Database Project, it will be evaluated and further developed as a key component of a federated multi-centre database of human studies.

Acknowledgments

The work on OCRe was supported in part by R01-LM06780 and MRC-G0100852.

References

1. Fridsma DB, Evans J, Hastak S and Mead CN. The BRIDG Project: A Technical Report. J Am Med Inform Assoc 2008;15:130–7.

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

A Collaborative Framework for Ontology Development

Tania Tudorache, Natasha Noy, Mark A. Musen
Stanford Center for Biomedical Informatics Research, Stanford, CA, USA

Abstract

We present a collaborative platform for editing and browsing ontologies in distributed environments that provide facilities for discussions, change tracking, provenance, policy control, simultaneous editing and querying. The platform can be accessed by Collaborative Protégé – an extension of the Protégé tool that provides rich support for editing, and by WebProtégé – a lightweight web-based version of the ontology editor that is implemented as a web portal. All software is open-source and available at: http://protege.stanford.edu.

The Collaborative Framework

As ontologies become more prevalent in many domains, such as bio-medicine, they evolve as dynamic products of collaborative development rather than artifacts produced in a closed environment of a single research group. In this poster, we will present a collaboration framework that our group has developed to support the collaborative ontology development. In designing the framework, we have been inspired by the popularity of Web 2.0 applications and have borrowed some of their most popular features as we describe in the following paragraphs. A user may access the framework through a desktop client, called Collaborative Protégé, through a web-client, WebProtégé, or from other applications.

The collaborative framework provides several collaboration functionalities that are exposed in the two clients, and that can be invoked by any other application using the Java API. Some of the main features are: support for simultaneous editing of ontologies, policy control (read, write, etc.), change tracking and provenance information (who did what and when), attaching different types of notes to entities in the ontology, live chat, integrated search on other terminologies and ontologies to support mappings, etc. Additional features that are already available in Protégé can also be accessed by the collaboration framework, such as structured comparison of different versions of an ontology, mappings, visualization, reasoning, and many others. The collaborative framework is pluggable and new functionalities can be added very easily. Two users working simultaneously in Collaborative Protégé and WebProtégé will see immediately each other changes.

Collaborative Protégé

We have developed Collaborative Protégé[1] as one of the clients of the Collaborative Framework. Collaborative Protégé enables users who develop an ontology collaboratively to hold discussions, chat, annotate ontology components and changes – all as an integral part of the ontology development process.

Users are able to attach notes to ontology parts (e.g., "*ToDo: check class synonyms for class Disease*"). Privileged users may assign tasks to other users, or may start change proposals (e.g., "*Adjust definition of this class*") to which other users may express their agreement or disagreement by voting. The framework tracks all the changes made in the ontology, so that a full history of an ontology component (e.g., a class) is available together with the provenance information. Search and filtering of all notes and other annotation types are also available through the user interface. Our collaborators have also found the integrated chat functionality to be very useful when they needed to send short message with internal links to ontology parts.

WebProtégé

An alternative client of the Collaborative Framework is WebProtégé[2] – an open source lightweight ontology editor for the Web that supports the collaborative development process. In designing the user interface, we took inspiration from well known portals, such as MyYahoo and iGoogle. We refer to each component in the user interface as a *portlet* (e.g., Class tree portlet, Notes portlet). Users can easily customize the appearance of the interface using drag-n-drop. Users can also show or hide specific tabs, or add other portlets to a tab via toolbar buttons. In this way, a project can customize the user interface of WebProtégé in a straightforward way based on different criteria (user's expertise, role, etc.). In a similar way to Collaborative Protégé, users may add notes to ontology parts. All edits in the web client are tracked together with provenance information.

References

1. http://protegewiki.stanford.edu/index.php/Collaborative_Protege/
2. http://protegewiki.stanford.edu/index.php/WebProtege

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

A Linguistic Approach to Aligning Representations of Human Anatomy and Radiology

Pinar Wennerberg[1], Manuel Möller[2], Sonja Zillner[1]
[1]Siemens AG, Munich, Germany; [2]DFKI, Kaiserslautern, Germany

Abstract

To realize applications such as semantic medical image search different domain ontologies are necessary that provide complementary knowledge about human anatomy and radiology. Consequently, integration of these different but nevertheless related types of medical knowledge from disparate domain ontologies becomes necessary. Ontology alignment is one way to achieve this objective. Our approach for aligning medical ontologies has three aspects: (a) linguistic-based, (b) corpus-based, and (c) dialogue-based. We briefly report on the linguistic alignment (i.e. the first aspect) using an ontology on human anatomy and a terminology on radiology.

Linguistic-based Medical Ontology Alignment

Semantic medical image search as approached by MEDICO[1] research project relies on ontology based semantic annotation of the image contents and patient data for an efficient search. Retrieving heterogeneous information (i.e. images and text) from a single access point requires the data to have been previously integrated appropriately. The integration task can be addressed by aligning the ontologies (i.e. *ontology matching*[2] or *alignment.*) that are used for the annotation e.g. Foundational Model of Anatomy (FMA) and Radiology Lexicon (RadLex). Our ontology alignment approach[3] has 3 aspects: (a) the linguistic analysis, (b) corpus analysis and (c) user interaction. The linguistic aspect suggests exploiting the information rich concept labels in the medical ontologies to discover further relations. The context information aspect based on corpus analysis assumes that ontology concepts from different ontologies with similar meaning will have similar contexts in the corpus so that the concept similarity follows from the context similarity. Finally, the user interaction aspect conceives of an alignment process as an interactive dialogue between the user and the system.

The linguistic alignment proposes to use rules for detecting the syntactic variants of the ontology concept labels to discover semantic relations e.g., equivalence and hyperonymy. In this way the initial ontologies are augmented with new concepts to be aligned subsequently. For example, the concept label 'blood in aorta' (noun preposition noun) can be transformed to 'aorta blood' using the rule: [noun1 preposition:in noun2 → noun2 noun1]. We annotated both FMA and RadLex concept labels with part-of-speech (POS) information i.e. we assigned the words their lexical categories. Eventually, for the most frequent prepositions (e.g. in, of, to) we generated 924 FMA and 135 RadLex variants (i.e. semantic equivalents) using the previous rule. Initial matching was exact string match between the concept labels of FMA and RadLex, which yielded 1147 correspondences. Additional 62 were found using the generated variants. We also generated hyperonyms (superconcepts) as additional concepts for subsequent matching. For all the multi-word concept labels from FMA and RadLex, where the last noun in the concept label is preceded by at least one or more successive adjectives, an adjective from the beginning of the multi-word concept label was omitted repeatedly until one adjective+noun combination was left. Each newly generated concept label in this way was added to the original as its hyperonym and used for matching. Eventually, we generated for FMA 1504 and for RadLex 902 hyperonyms that we incorporated in the alignment process. Currently, we are at the process of evaluating the validity of the semi-automatically generated variants with our clinical expert. Next step will be the implementation of the context-based alignment aspect.

Acknowledgements

This research has been supported in part by the THESEUS Program in the MEDICO Project, which is funded by the German Federal Ministry of Economics and Technology under the grant number 01MQ07016. The responsibility for this publication lies with the authors. Special thanks to Kamal Najib of Siemens AG for implementing the tests and to our clinical partner Dr. Alexander Cavallaro of the University Hospital Erlangen.

References

1. Möller M, Regel S and Sintek M. "RadSem: Semantic Annotation and Retrieval for Medical Images", In: Proc. of the ESWC 2009.
2. Euzenat J and Shvaiko P. Ontology Matching. June 2007, Springer-Verlag.
3. Wennerberg P. Aligning Medical Ontologies for Clinical Query Extraction. In: Proc. of EACL 2009, PhD Symposium, Athens, Greece.

BioPortal: Ontologies and Integrated Data Resources at the Click of the Mouse

Patricia L. Whetzel[1], Nigam H. Shah[1], Natalya F. Noy[1], Benjamin Dai[1], Michael Dorf[1], Nicholas Griffith[1], Clement Jonquet[1], Cherie Youn[1], Chris Callendar[2], Adrien Coulet[1], Daniel L. Rubin[1], Barry Smith[3], Margaret-Anne Storey[2], Christopher G. Chute[4], Mark A. Musen[1]
[1]Stanford University, Stanford, CA, USA; [2]University of Victoria, Victoria, Canada;
[3]State University of New York at Buffalo, Buffalo, NY, USA; [4]Mayo Clinic, Rochester, MN, USA

Abstract

BioPortal (http://bioportal.bioontology.org) is an open repository of biomedical ontologies that provides programmatic and web-based access to ontologies developed in OBO, OWL, Protégé frames, and RDF. Features include browsing, searching, and visualization of ontologies. Searching of integrated data resources is also possible through ontology-based indexing of biomedical resources with BioPortal ontologies.

Introduction

A variety of ontology repositories exist, however they differ by either method of ontology content collection or ontology formats supported.[1-6] BioPortal is an open repository of biomedical ontologies that store ontologies developed in various formats, that provides for automatic updates by user submissions of new versions, and that provides access via Web browsers and through Web services.

BioPortal Content and Functionality

The ontology content of BioPortal covers a wide range of subject matter such as anatomy, phenotype, imaging, chemistry, and experimental conditions. BioPortal supports ontologies in OBO, OWL, Protégé frames, and RDF. Metadata collected for each ontology include keywords, version information, release date, and ontology author contact information. BioPortal also supports filters of the ontology content such as limiting the view to OBO Foundry ontologies.[7]

BioPortal users can browse and search the ontologies, submit new versions of the ontologies in the repository, comment on any ontology (or portion of an ontology) in the repository, add a review of the ontology, describe their experience in using the ontology, or make suggestions to ontology developers. The focus on enabling members of the community to contribute actively to BioPortal content and to increase the value of that content to other users distinguishes BioPortal from other ontology repositories. Another key feature of BioPortal is the ability to query biomedical data resources such as ArrayExpress, the Gene Expression Omnibus (GEO), and ClinicalTrials.gov through the annotation and indexing of these resources with ontologies in BioPortal.[8,9]

Conclusion

BioPortal not only provides investigators, clinicians, and developers "one-stop shopping" to programmatically access biomedical ontologies, but also integrates data from various biomedical resources.

Acknowledgements

This work was supported by the National Center for Biomedical Ontology, under the roadmap-initiative from the National Institutes of Health [grant U54 HG004028].

References

1. http://swoogle.umbc.edu
2. http://watson.kmi.open.ac.uk/Overview.html
3. http://olp.dfki.de/ontoselect?wicket:bookmarkablePage=wicket-0:de.dfki.ontoselect.Home
4. http://www.daml.org/ontologies
5. http://www.schemaweb.info
6. http://www.ebi.ac.uk/ontology-lookup
7. Smith B, Ashburner M, Rosse C, *et al*. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007; 25:1251–55.
8. Shah NH, Jonquet C, Chiang AP, Butte AJ, Chen R and Musen MA. Ontology-driven indexing of public datasets for translational bioinformatics. BMC Bioinformatics. 2009;10, Suppl 2:S1.
9. Jonquet C, Shah NH and Musen MA. The Open Biomedical Annotator. AMIA Summit on Translational Bioinformatics, p. 56–60, March 2009, San Francisco, CA, USA.

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

Improvement of PubMed Literature Searching Using Biomedical Ontology

Zuoshuang Xiang, Yongqun He
University of Michigan, Ann Arbor, MI, USA

Abstract

PubMed articles are annotated using the Medical Subject Headings (MeSH) to increase search efficiency. However, MeSH contains limited information on many biomedical domains (e.g., vaccines). Biomedical ontologies may be used to improve PubMed searching capability. This study demonstrates that Vaccine Ontology (VO) can be used to significantly improve PubMed searching efficacy in the vaccine domain. The recall and precision of the ontology-based literature mining approach are analyzed and discussed.

Introduction

MeSH is the controlled vocabulary of medical and scientific terms that are used by biomedical scientists to manually index articles in the PubMed literature database. The MeSH terminology has been used in PubMed to improve literature searching. However, MeSH does not cover many biomedical domains (e.g., vaccine) well. An ontology represents the consensus-based controlled vocabularies of terms and relations which are logically formulated to promote intelligent information retrieval and modeling. We hypothesize that ontology-based PubMed search will significantly improve literature search efficacy. To test this hypothesis, we apply the Vaccine Ontology (VO; http://www.violinet.org/vaccineontology) to search for PubMed literature associated with *Brucella* vaccines. *Brucella* is an intracellular bacterium that causes brucellosis, the most common zoonotic disease worldwide.

Results

A user case study is to search "live attenuated *Brucella* vaccine" in PubMed. As of June 10, 2009, a direct PubMed search of this string of keywords returns 58 papers (or PubMed hits). VO includes 16 *Brucella* vaccines (including 4 licensed vaccines) that have phenotypes of 'live' and 'attenuated'. We developed an algorithm that recursively searches for all ontology labels and synonyms of the class '*Brucella* vaccine' and all its subclasses in VO. The phenotypes 'viable' (synonym: live) and 'attenuated' are used for filtering out unqualified *Brucella* vaccines. All names are assembled into a searching string for a PubMed keywords search. VO is designed based on OBO Foundry principles. Each subclass in VO has an '*is_a*' relationship with its parent class. This ensures that all subclasses (e.g., *Brucella* RB51) can be included when a parent class (e.g., "*Brucella* vaccine") is searched.

PubMed Search Keywords	Hits	True	Precision
live attenuated *Brucella* vaccine	58	55	95%
Consider "live attenuated *Brucella* vaccine" in VO:			
Brucella (RB51 OR SRB51)	182	182	100%
Brucella (strain 19 OR S19)	537	510	95%
Brucella Rev. 1	145	144	99%
B. suis (strain 2 OR S2)	56	12	21%
Brucella bacA mutant vaccine	1	1	100%
Other 12 live attenuated *Brucella* vaccines in VO	62	59	95%
Total (unique ones)	763	695	91%

Table 1. Enhanced literature search using VO.

Our search using this VO-based method significantly increased the recall of searching "live attenuated *Brucella* vaccine" by 13 fold (698/55) compared to the searching without VO (Table 1). The precision of the searching remains high (91% vs 95%). Using "strain 19" or "strain 2" instead of (strain 19) or (strain 2) dramatically improves precision to >95% (not implemented in Table 1). Our study shows that inclusion of synonyms (e.g., strain 19 vs. S19) can improve searching recall. However increased recall and precision can only be achieved with well-assigned names for labels and synonyms. For example, the search for the term '*Brucella abortus bacA* mutant vaccine' returns higher searching precision than '*Brucella abortus bacA* mutant' without lose of recall. On the other hand, "*Brucella* RB51" is better than "*Brucella abortus* RB51" for obtaining higher recall without lose of precision. This approach was also successfully applied to other types of *Brucella* vaccines and vaccines against other pathogens. A web server (http://www.violinet.org/pubvo) is currently under development using our ontology-based literature mining method.

Conclusion and Discussion

Bio-ontologies (e.g., VO) can be used to improve literature searching. Ontology term naming is important for improved literature search.

Acknowledgements

This research is supported by a Rackham Pilot Research grant at the University of Michigan.

CORRECTED VERSION

Logical Implications for Regulatory Relations Represented by Verbs in Biomedical Texts

Sine Zambach
Roskilde University, Roskilde, Denmark

Abstract

Relations used in biomedical ontologies can be very general or very specific in respect to the domain. However some relations are used widely in for example regulatory networks. This work focuses on positive and negative regulatory relations, in particular their usage expressed as verbs in different biomedical genres and the properties of the relations.

Introduction

In the research area of biomedical ontologies, the work with formal relations has recently reached a level where integration in a larger project is possible[1]. Using a thorough analysis of the actual logic implications of the relations has been suggested[2].

With the opportunity of using new DL-formalism $\mathcal{EL}+$ and reasoning tool[3], CEL, relations can be treated as modules with complex inclusions themselves.

This study is concerned with the two relations, positive and negative regulation relations. The relations have been investigated in corpora and in relation to their logical implications. For an easy reading we call them *stimulates* and *inhibits*.

Results

Inhibit and stimulate are relations that – in a biochemical pathway actually contains a special kind of inheritance, e.g. if *x* inhibits *y* and *y* stimulates *z*, then you can deduce that *x* stimulates *z* as formulated in FOL using *stim* for stimulation and *inh* for inhibition:

$$\forall x(A(x)\rightarrow\forall y(B(y)\wedge stim(x,y)\wedge\forall z(C(z)\wedge stim(y,z)\rightarrow stim(x,z))))$$
$$\forall x(A(x)\rightarrow\forall y(B(y)\wedge inh(x,y)\wedge\forall z(C(z)\wedge inh(y,z)\rightarrow stim(x,z))))$$
$$\forall x(A(x)\rightarrow\forall y(B(y)\wedge inh(x,y)\wedge\forall z(C(z)\wedge stim(y,z)\rightarrow inh(x,z))))$$
$$\forall x(A(x)\rightarrow\forall y(B(y)\wedge stim(x,y)\wedge\forall z(C(z)\wedge inh(y,z)\rightarrow inh(x,z))))$$

These properties can also be expressed in the $\mathcal{EL}+$ language and are called *complex role inclusions*[3]: *inhibit o stimulate* $\subseteq$ *inhibit* etc and the new CEL-module can handle axioms like this.

The verb frequencies in biomedical corpora are ranked in figure 1. This experiment reveals that verbs representing the regulatory relations has a special use in for example Medline abstracts and biomedical patents compared to the British National Corpus (BNC). Whereas the 10 most frequent verbs from BNC has a low rank in all corpora, the regulatory relations seems to be more specific for the biomedical texts.

Conclusion

Corpora analysis indicates that the words representing positive and negative processes have an important role in biomedical texts which should be investigated further. In addition to this, an implementation in DL is suggested such that these relations can be expressed in a way that facilitates *e.g.* reasoning. Files can be found at: *ruc.dk/~sz/ICBO09/relations.*

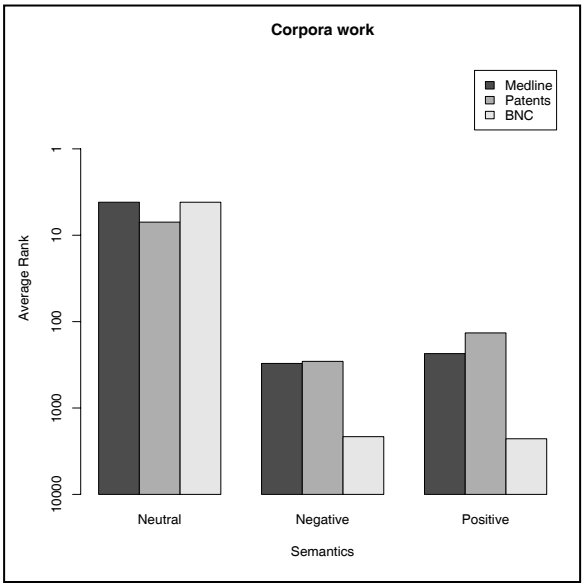


Figure 1. Average rank of regulatory relations for biomedical patents, Medline abstracts and BNC.

References

1. Smith B, *et al*. Relations in biomedical ontologies. Genome Biology 2005, 6:R46.
2. Smith B and Rosse C. The role of foundational relations in the alignment of biomedical ontologies, Medinfo. 2004, 444–448.
3. Baader F, Lutz C and Suntisrivaraporn B. CEL – a polynomial-time reasoner for life science ontologies. Lecture Notes for Computer Science, Springer Berlin / Heidelberg, 2006, 287–291.

ICBO: International Conference on Biomedical Ontology
July 24-26, 2009 ▪ Buffalo, New York, USA

Annotation-Based Meta-Analysis of Microarray Experiments

Jie Zheng, Junmin Liu, Elisabetta Manduchi, Christian J. Stoeckert, Jr.
Center for Bioinformatics, Department of Genetics,
University of Pennsylvania School of Medicine, Philadelphia, PA, USA

Abstract

We are developing software applications to perform meta-analysis of microarray experiments based on standardized experiment annotations aiming to identify similar experiments and cluster experiments. The applications were tested on files obtained from the ArrayExpress public repository. Annotation terms were used to compute experiment dissimilarities to find experiments related to a query experiment. These applications may motivate efforts of bench biologists to better annotate experiments.

Introduction and Methods

Integrating data to address a scientific problem of interest is a major challenge. Meta-analysis of experiments based on standard annotations to identify similar experiments will facilitate such data integration. Standardized microarray experiment annotations can be generated using the MGED ontology (MO)[1]. We are developing software applications to extract the annotation components covering the biological intent and context of experiments and then generate dissimilarity measures between experiments to find related experiments. Annotated experiments were obtained from ArrayExpress[2] in the MAGE-TAB format[3] and used to test the applications.

Annotation Components Used:

- *StudyName* (free-text)
- *ExperimentDesignType* (MO terms)
- *ExperimentFactorType* (MO terms)
- *ExperimentFactorValue* (free-text or measurement or ontology terms)
- *Organism* (ontology terms)
- *BiomaterialCharacteristics* (ontology terms)
- *TreatmentType* (MO terms)

Proposed Dissimilarity Measures: The dissimilarity between two experiments was computed based on the overlap of the two corresponding annotation term sets for each annotation component and weighted averaging across components.

Gold Standards for Experiments: A list of gold standards for experiments about glucose responsive genes and insulin secretion in islets was obtained based on keyword searches of ArrayExpress. Three experiments in this list were used as query experiments to find related experiments obtained from a total of 6632 experiments. The remaining experiments from the same list were used as positive controls. A list of negative controls was compiled too.

Results

The first software module to retrieve annotations from MAGE-ML or MAGE-TAB files has been developed. A second software module has been built to generate dissimilarity measures between a query experiment and a collection of target experiments, based on the extracted annotation terms. We tried various combinations of 0-1 weights to include/exclude annotation components to optimize the scores between the positive (resp. negative) controls and the query experiments.

The query experiment with the richest annotations led to the best results in finding closely related experiments in ArrayExpress when the free-text annotation fields (StudyName and ExperimentFactorValue) were not included. Including these fields improved the ability to distinguish true positive and true negative experiments using the two query experiments with fewer annotations. This indicates that the free-text fields provide extra information for experiments with fewer standardized annotations. Including all annotation components except for Organism and TreatmentType gave good results. However, this does not mean that these should be the weights of choice for all possible applications.

Conclusion and Future Work

1. Meta-analyses based on annotation can help to find closely related experiments. The more richly annotated query experiment gave better results in identifying similar experiments from our test set.
2. A suitable weight assigned to each annotation component can improve the dissimilarity measures.
3. To further improve the comparison of annotation terms, we will harmonize terms using meta-thesauri, tag controlled terms from free-text annotations, and apply ontological relationships to refine the dissimilarity measures.

Acknowledgements

This study is funded by NIH Grant, R21-HG004521.

References

1. Whetzel, *et al.* Bioinformatics. 2006; 22(7): 866.
2. Parkinson, *et al.* Nucleic Acids Res. 2007;35 (Database issue):D747.
3. Rayner, *et al.* BMC Bioinformatics 7:489.