The 11th Annual Bio-Ontologies Meeting

Phillip Lord, Newcastle University
Nigam Shah, Stanford University
Susanna-Assunta Sansone, EMBL-EBI
Matthew Cockerill, BioMed Central

The Bio-Ontologies Meeting has existed as a Special Interest Group meeting at ISMB for more than a decade, making it one of the longest running SIG meetings. The Bio-Ontologies SIG meeting has provided a forum for discussion on the latest and most cutting edge research on ontologies. In this decade, the use of ontologies has become mature, moving from niche to mainstream usage within bioinformatics. As result, this year we are broadening the scope of SIG to include formal and informal approaches to organising, presenting and disseminating knowledge in biology.

July 20, 2008
Colocated with ISMB 2008
Toronto, Canada

Keynote

This year's keynote speaker will be Professor Philip E. Bourne, from the Skaggs School of Pharmacy & Pharmaceutical Sciences, UCSD.

Panel Session

This year's panel session will have the following members:

- Philip E. Bourne, UCSD
- Helen Parkinson, European Bioinformatics Institute
- Matt Cockerill, BioMed Central
- Mark Wilkinson, University of British Columbia

Acknowledgements

We acknowledge the assistance of Steven Leard and all at ISCB for their excellent technical assistance.

We also wish to thank the programme committee for their excellent input and guidance – the programme committee, organised alphabetically is:

Mike Bada, University of Colorado
Judith Blake, Jackson Laboratory
Frank Gibson, Newcastle University
Cliff Joslyn, Pacific National Laboratory
Wacek Kusnierczyk, Norwegian University of Science and Technology
Robin MacEntire, GSK
James Malone, EBI
Helen Parkinson, EBI
Daniel Rubin, Stanford University
Alan Ruttenberg, Science Commons
Susie M. Stephens, Eli Lilly
Robert Stevens, University of Manchester

Talks schedule

Start	End	Authors	Title
8:50	9:00		Introduction
9:00	9:20	R. Kanagasabi *et al*	Ontology-Centric navigation of pathway information mined from text
9:20	9:40	C. Brewster *et al*	Issues in learning an Ontology from Text
9:40	10:00	T. Tong *et al*	GO-WORDS: An entropic Approach to Semantic Decomposition of Gene Ontology Terms
10:00	**10:40**	**Coffee**	
10:40	11:00	T. Beck *et al*	Using ontologies to annotate large-scale mouse phenotype data
11:00	11:20	J. Malone *et al*	Developing an application focused experimental factor ontology: embracing the OBO Community
11:20	11:40	C. Arighi *et al*	TGF-beta Signaling Proteins and the Protein Ontology
11:40	**12:30**	**Phil Bourne**	**Keynote talk**
12:30	13:50	Lunch and Poster Session 1	
13:50	**15:10**	**Panel**	**Philip E. Bourne,** UCSD
			Helen Parkinson, European Bioinformatics Institute
			Matt Cockerill, BioMed Central
			Mark Wilkinson, University of British Columbia
15:10	15:30	A.X.Qu *et al*	Tamoxifen to Systemic Lupus Erythematosus: Constructing a Semantic Infrastructure to Enable Mechanism-based Reasoning and Inference from Drugs to Diseases
15:30	15:50	R. Hoehndorf *et al*	BOWiki: An ontology-based wiki for annotation of data and integration of knowledge in biology
15:50	16:10	W. Xuan *et al*	PubOnto: Open Biomedical Ontology-Based Medline Exploration
16:10	**16:40**	**Coffee**	
16:40	17:00	J.B.L. Bard *et al*	Minimal Information about Anatomy (MIAA): a species-independent terminology for anatomical mapping and retrieval.
17:00	17:20	R. Arp *et al*	Function, Role and Disposition in Basic Formal Ontology
17:20	18:30	Closing Remarks and Poster Session 2	

Posters

No.	Authors	Title
1	Joel Sachs, Cynthia Parr, Lushan Han, Taowei Wang and Tim Finin	Case Studies of Simple Ontologies and Instance Data for Biodiversity Informatics
2	Pascale Gaudet, Petra Fey, Eric Just, Sohel Merchant, Siddhartha Basu, Warren Kibbe and Rex Chisholm	Strain and phenotype annotation at dictyBase
3	Joanne Luciano, Lynn Schriml, Burke Squires and Richard Scheuermann	The Influenza Infectious Disease Ontology (I-IDO)
4	Son Doan, Quoc Hung Ngo, Ai Kawazoe and Nigel Collier	Building and Using Geospatial Ontology in the BioCaster Surveillance System
5	Philippe Rocca-Serra, Ryan Brinkman, Liju Fan, James Malone and Susanna-Assunta Sansone	OBI: The Ontology for Biomedical Investigations
6	Lynn Schriml, Katherine Phillippy, Aaron Gussman, Anu Ganapathy, Sam Angiuoli, Suvarna Nadendla, Victor Felix, Elodie Ghedin, Owen White and Neil Hall	The Evolution of Controlled Vocabularies and Community Development Ontologies in the Frame Work of an Epidemiology Database
7	François Belleau, Peter Ansell, Marc-Alexandre Nolin, Kingsley Idehen and Michel Dumontier	Bio2RDF's SPARQL Point and Search Service for Life Science Linked Data
8	Andrea Splendiani, Scott Piao, Yutaka Sasaki, Sophia Ananiadou, John McNaught and anita burgun	Compositional enrichment of bio-ontologies
9	Mary Shimoyama, Melinda Dwinell and Howard Jacob	Multiple Ontologies for Integrating Complex Phenotype Datasets
10	Indra Neil Sarkar, Ryan Schenk, Holly Miller and Catherine Norton	Identification of Relevant Biomedical Literature Using Tag Clouds
11	Jason Greenbaum, Randi Vita, Laura M. Zarebski, Alessandro Sette and Bjoern Peters	Ontology of Immune Epitopes (ONTIE)

Ontology-centric navigation of pathways mined from text.

Rajaraman Kanagasabai[1], Hong-Sang Low[2], Wee Tiong Ang[1], Markus R. Wenk[2], and Christopher J. O. Baker[1]

[1]Data Mining Dept. Institute for Infocomm Research, & [2]Departmet of Biochemistry, Faculty of Medicine, NUS, Singapore,

ABSTRACT

Motivation: The scientific literature is the primary means for the navigation of new knowledge as well as retrospective analysis across merging disciplines. The state of the art in this paradigm is a series of independent tools that have to be combined into a workflow and results that are static representations lacking sophisticated query-answer navigation tools. In this paper we report on the combination of text mining, ontology population and knowledge representation technologies in the construction of a knowledgebase on which we deploy data mining algorithms and visual query functionality. Integrated together these technologies constitute an interactive query paradigm for pathway discovery from full-text scientific papers. The platform is designed for the navigation of annotations across biological systems and data types. We illustrate its use in tacit knowledge discovery and pathway annotation.

1 INTRODUCTION

A growing number of knowledge discovery systems incorporate text mining techniques and deliver insights derived from the literature. Named entity recognition and relation detection are primary steps. The products of such techniques can take the form of automatically generated summaries, target sentences or lists of binary relations between entities [1] from abstracts for which subsequent networks can be constructed [2] and visualized in graphs [3] in some cases with predefined class directed-layouts [4]. The state of the art in this paradigm is a series of independent tools that have to be combined into a workflow and results that are static representations lacking sophisticated query-answer navigation tools. To enhance the accessibility and search ability of the insights derived from texts these instances of named entities and relations should be associated with descriptive metadata such as ontologies. Recent examples have shown that instantiating ontologies with text segments can be meaningful and useful in knowledge discovery projects [5]. In this paper we go a step further and mine instantiated ontology for transitive relations linking query terms and make this available in the context of (i) tacit knowledge discovery across biological systems; proteins, lipids and disease, and (ii) in mining for pathway segments which can iteratively be re-annotated with relations to other biological entities also recognized in full text documents.

* cbaker@i2r.a-star.edu.sg

2 METHODS

The material for our analysis is full text scientific literature. Details and efficiency of our text mining approach, the customization of the ontology to enable the pathway discovery scenario and instantiation of the ontology are detailed below. We also outline pathway discovery algorithms used to facilitate navigation of putative pathways and annotations.

2.1 Ontology Population

Ontology population was achieved through the coordination of content acquisition, natural language processing and ontology instantiation strategies. We employed a content acquisition engine that takes user keywords and retrieves full-text research papers by crawling Pubmed search results. Retrieved collections of research papers were converted from their original formats, to ascii text and made ready for text mining by a customized document converter. Knowledgebase 'instances' are generated from full texts provided by the content acquisition engine using the BioText toolkit. http://datam.i2r.a-star.edu.sg/~kanagasa/BioText/

2.2 Knowledgebase Instantiation

Instantiation comprises of three stages: Concept Instance Generation, for which we also provide a performance evaluation, Property Instance Generation, and Population of Instances. In the context of OWL-DL, Property Instances are assertions on individuals which are derived from relations found in predicate argument structures in mined sentences.

2.2.1 Concept Instance Generation. Concept instances are generated by first extracting the name entities from the texts and then normalizing and grounding them to the ontology concepts. Our entity recognizer uses a gazetteer that processes retrieved full-text documents and recognizes entities by matching term dictionaries against the tokens of processed text, and tags the terms found [5]. The lipid name dictionary was generated from Lipid Data Warehouse that contains lipid names from LIPIDMAPS [6], LipidBank and

KEGG., IUPAC names, and optionally broad synonyms and exact synonyms. The manually curated Protein name list from Swiss-Prot (http://au.expasy.org/sprot/) was used for the protein name dictionary. A disease name list was created from the Disease Ontology of Centre for Genetic Medicine (http://diseaseontology.sourceforge.net). Our normalization and grounding strategy is as follows. Protein names were normalized to the canonical names entry in Swiss-Prot. Grounding is done via the Swiss-Prot ID. For lipid names, we define the LIPIDMAPS systematic name as the canonical name, and the LIPIDMAPS database ID is used for grounding. Disease names are grounded via the ULMS ID.

To evaluate the performance of our named entity/concept recognition we constructed a gold standard corpus of 10 full-texts papers related to the Apoptosis pathway, obtained from our content acquisition engine. We extracted 119 sentences and tagged the mentions of Protein name and Disease name. In these sentences we annotated all valid mentions of the two concepts and built the corpus. To evaluate performance of named entity/concept recognition a corpus without the concept annotations was passed to our text mining engine and the concepts recognized. Our system was evaluated in terms of precision and recall. Precision was defined as the fraction of correct concepts recognized over the total number of concepts output, and recall was defined as the fraction of concepts recognized among all correct concepts.

Table 1. Precision and recall of named entity recognition

Named Entities	Mentions		Precision	Recall
	Target	Returned		
Disease	32	37	0.54	0.62
Lipid	58	25	0.96	0.47
Protein	269	181	0.76	0.51
Micro Average			0.75	0.51

The evaluation of entity recognition shown in Table 1 shows that our text mining achieved performance comparable to that of the state-of-the-art dictionary-based approaches. In our future work, we plan to make use of advanced entity recognition techniques, e.g. fuzzy term matching and coreference resolution, and also train our system on larger corpora, to address these issues.

2.2.2 Property Instance Generation. Object property and Datatype property instances are generated separately. From the Lipid, Protein and Disease instances, four types of relation pairs namely Lipid-Protein, Lipid-Disease, Protein-Protein, Protein-Disease are extracted. For relation detection, we adopt a constraint-based association mining approach whereby two entities are said to be related if they co-occur in a sentence and satisfy a set of specified rules [5]. The relation pairs from the resulting sentences are used to generate the Object property instances. The interaction sentences are instantiated as Datatype property instances. Several other Object property instances are also generated to establish relations between, for example, LIPIDMAPS Systematic Name and its associated IUPAC Name, synonyms and database ID. However, in this case the relation pairs are generated directly from the Lipid Warehouse records requiring no text processing.

2.2.3 Population of Instances. In this step we collect all the concept and property instances generated from the previous two to instantiate the ontology. The concept instances are instantiated to the respective ontology classes (as tagged by the gazetteer), the Object Property instances to the respective Object Properties and the Datatype property instances to the respective Datatype properties. We wrote a custom script using the OWL programming framework, JENA API http://jena.sourceforge.net/ for this purpose.

2.3 Ontology Extension

To facilitate the navigation of pathway information we modified the existing lipid ontology [5] by incorporating Protein concepts under two newly defined superconcepts (i) Monomeric_Protein_or_Protein_Complex_Subunit and (ii) Multimeric_Protein_Complex. This was achieved either by importing protein entities found in Molecule Roles Ontology or by adding the names manually. In total, we incorporated about 48 protein class entities under these 2 concepts. Each protein entity relates to another via the property "hasProtein_Protein_Interaction_with". Each protein entity then relates to a lipid entity via the property "interactsWith_Lipid". These extensions facilitate query of protein-protein interactions derived from tuples found by the text mining of full text documents.

Generic Pathway Algorithm

Let C_{source} = source concept, C_{target} = target concept, and ***P*** = null

If $C_{source} = C_{target}$, output ***P*** as the pathway and stop.

Check if C_{source} = D where D ∈ T = {D, P, R}, and D and R are concept instances related via the object property P.

If C_{target} = R where R ∈ T = {D, P, R}, add the edge $C_{source} \rightarrow T^* \rightarrow C_{target}$ to ***P.***

Else, Let C_{source} be R, and go to Step 1.

Check if C_{source} = R where R ∈ T = {D, P, R}.

If C_{target} = D where D ∈ T = {D, P ,R}, add the edge $C_{source} \leftarrow T^* \leftarrow C_{target}$ to ***P.***

Else, Let C_{target} be D, and go to Step 1.

Fig. 1 Generic pathway algorithm for mining transitive relations.

2.4 Pathway Discovery Algorithm

We implemented a generic pathway discovery algorithm for mining all object properties in the ontology to discover transitive relations between two entities. An outline of this algorithm is presented in Figure 1. Given two concept instances C_{source} and C_{target}, the algorithm seeks to trace a pathway between them using the following approach. First, the algorithm lists all object property instance triples in which the

domain matches C_{source}. Thereafter every listed instance is in turn treated as the source concept instance and the related object property instances explored. This process is repeated recursively until C_{target} is reached or if no object property instances are found. All resulting transitive paths are output in the ascending order of path length. We implement a protein-protein interaction pathway discovery algorithm by adding the following two simple constraints to the generic algorithm: 1) the source and domain concepts are restricted to be proteins, 2) only object property instances of *hasProtein_Protein_Interaction_with* are included.

3 PATHWAY MINING FROM LITERATURE

Knowlegtor [1] is a visual-query navigation platform for OWL-DL ontologies which facilitates the construction of concept level queries from OWL ontology constructs and relays them to a reasoner to query a knowledgebase populated with A-box instances. We have integrated 2 new pathway algorithms into the Knowlegator platform to facilitate literature driven tacit knowledge discovery and apply it here as 'pathway mining'.

In order to mine the instantiated ontology for the existence of one or more pathways between user-specified proteins the graphical features of Knowlegator permit users to drag two protein onto the query canvas and then invoke a search for transitive relations between these two concepts (Figure 2). Results from this search are returned as a list of possible pathways each of which can rendered on the query canvas as a chain of labeled concepts and instances illustrating the linkage between the selected starting entities (Figure 3). These pathways traverse multiple relation and data types, namely, protein, lipid and disease names as well as provenance data i.e. individual sentences and document identifiers. Parent concept names are rendered along with instance level names. By using a wide range of relations a deeper search for tangible relations between entities is facilitated. This is however beyond the scope of pathway analysis and more in line with identifying evidence sources and illustrating causality or participants in a disease context. There exists however many such paths through the instantiated ontology and the user's navigation experience may become tedious, in particular when the user is confronted with significant sources of new material. Within the process of knowledge discovery a more intuitive approach is the iterative overlay of new material on top of existing knowledge that is queried in the first stage of the analysis. In this vain we illustrate the overlay of lipid-protein interaction information on top of the protein-protein interaction information displayed in the initial pathway discovery step. Knowlegtor facilitates use of the second pathway algorithm for users who wish to apply specific constrains on the pathway they hope to find, be it based on protein-protein interactions, protein-disease or protein-lipid interactions. Figure 4 shows a pathway query for a protein-protein pathway with associated protein-lipid interactions, the results for which are

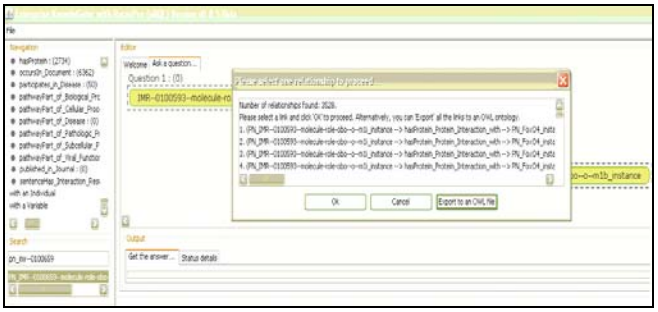


Fig. 2 A tacit knowledge query in Knowlegator, searching for links between two proteins in Apoptosis signaling.

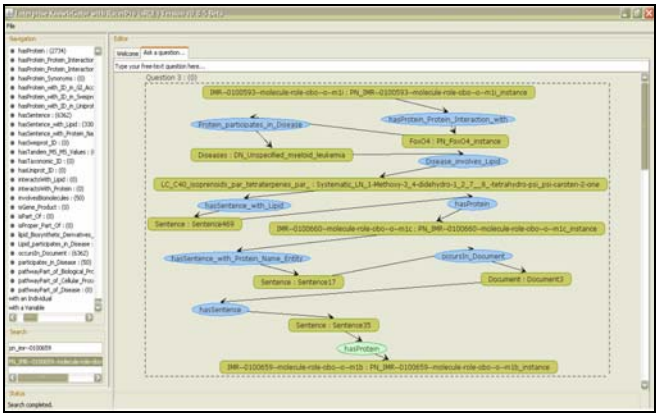


Fig. 3 Results of a query to the instantiated ontology using PI3K kinase and P53 which both play roles in apoptosis signaling.

provided for the PI3-Kinase to BIM II fragment of the apoptosis pathway (Figure 5). In our initial trials we considered the following 3 known protein pathways:

1. PI3K --> Akt(Akt1; Akt2; Akt3) --> Bad(Bad)
   --> Bcl-xl(Bcl-xL:) --> Bid(Bid) --> Bax(Bax)
2. PI3K --> Akt(Akt1;Akt2;Akt3) --> Mdm2(Mdm2)
   --> p53(p53) --> Puma(Puma) --> Bim(Bim)
3. PI3K --> Akt(Akt1;Akt2;Akt3) -->FoxO1(FoxO1)
   --> FasL(FasL) --> FasR(FasR) --> Caspase-3

where PI3K is (PI3-kinase p85-alpha subunit; PI3-kinase p85-beta subunit; PI3-kinase p110 subunit alpha; PI3-kinase p110 subunit beta; PI3 kinase p110 subunit delta).

For each pathway, we ran our pathway discovery algorithm with first protein as the source concept and the last protein as the target concept. There were > 1000 paths linking the soure and target proteins and the correct pathway was identified as the path that had the identical P2P interactions as in the known pathway and in the same order, stipulated by our domain expert. With our system we were able to identify 2 out of 3 known pathways exactly. The third pathway was found except for the segment PN_FoxO1 => PN_FasL => PN_FasR. Instead our algorithm identified a shorter path "PN_FoxO1 => PN_FasR indicating there system is able to pick up nuances beyond common 'textbook' pathways.

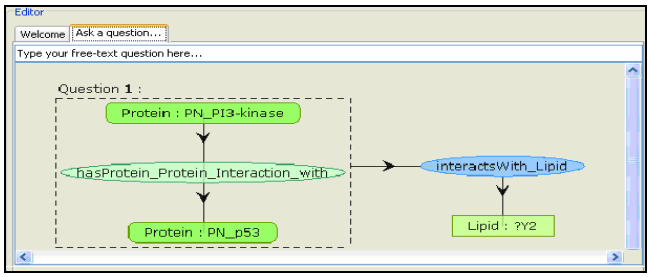


Fig. 4 shows the query for an apoptosis pathway fragment involving PI3K kinase and P53 and for lipid-protein annotations.

The URL below links to three sample pathways found using the transitive query - protein interaction algorithm.
http://datam.i2r.a-star.edu.sg/~kanagasa/pathway/index.html

4 DISCUSSION

The challenge we address in our scenario is the aggregation of tuples of normalized named entities from full text documents and the provision of these tuples as an interactive query resource for pathway discovery. The overall workflow has a series of exchangeable components which make it an attractive solution. In future we plan to evaluate the benefits to the overall system of exchanging different components, such as the information extraction engine which exports tuples to a tagged XML file. In the current system ontology population takes an average of 1 min/document, mining the protein-protein pathways takes on average of 45 secs, and the automated verification of a known pathway took less than 1 sec. In addition to generating content and modifying the ontology to support the instantiation of protein-protein interactions, we have deployed two data mining algorithms within the Knowlegator platform. With Knowlegator's drag and drop query paradigm users can generate cross-discipline paths or stepwise extensions to existing known pathways by adding annotations or alternate paths e.g. that include lipids. Moreover the results can be returned with concept labels as well as instance names to enhance the semantics of knowledge discovery output. We envision that users of our approach would have a specific set of pathways in mind from a given biological domain and specify a body of literature to be mined and from which relevant information would be instantiated to the ontology. Thereafter these users would navigate outwards from known pathways selecting to augment them with information which is beyond their domain expertise. Moreover the ontology captures sentence provenance so that as users can verify new information that they were not previously aware of. Whilst this is preliminary work it shows that mining literature sources in the context of existing knowledge domain can support scientists engaged in knowledge discovery around pathways. As we move forward with this paradigm we acknowledge that we become more dependent on domain experts for precise requirements, (pathways and corresponding corpora) and for verification of the value added by the system in their discovery process, which in some contexts is subjective.

Pathway Fragment: PI3-Kinase to BIM II

PN_PI3-kinase_p110_subunit_alpha => PN_Akt1

PathWithLipid: PN_PI3-kinase_p110_subunit_alpha => Systematic_LN_ethanoic_acid => PN_Akt1

PathWithLipid: PN_PI3-kinase_p110_subunit_alpha => Systematic_LN__5Z_7E_par_-_3S_par_-9_10-seco-5_7_10_19_par_-cholestatrien-3-ol => PN_Akt1

PathWithLipid: PN_PI3-kinase => Systematic_LN_Paclitaxel => PN_Akt1

PN_Akt1 => PN_Mdm2

PathWithLipid: PN_Akt1 => Systematic_LN_ethanoic_acid => PN_Mdm2

PathWithLipid: PN_Akt1 => Systematic_LN_GalNAca1-3_Fuca1-2_par_Galb1-3GlcNAcb1-3Galb1-3GlcNAcb1-3Galb1-4Glcb-Cer => PN_Mdm2

PN_Mdm2 => PN_p53

PathWithLipid: PN_Mdm2 => Systematic_LN_1-Methoxy-3_4-didehydro-1_2_7__8_-tetrahydro-psi_psi-caroten-2-one => PN_p53

PathWithLipid: PN_Mdm2 => Systematic_LN_ethanoic_acid => PN_p53

PathWithLipid: PN_Mdm2 => Systematic_LN_2-methyl-propanoic_acid => PN_p53

PathWithLipid: PN_Mdm2 => Systematic_LN_GalNAca1-3_Fuca1-2_par_Galb1-3GlcNAcb1-3Galb1-3GlcNAcb1-3Galb1-4Glcb-Cer => N_p53

PN_p53 => PN_Puma

PathWithLipid: PN_p53 => Systematic_LN_1-Methoxy-3_4-didehydro-1_2_7__8_-tetrahydro-psi_psi-caroten-2-one => PN_Puma

PathWithLipid: PN_p53 => Systematic_LN_Phorbol => PN_Puma

PN_Puma => PN_Bim

PathWithLipid: PN_Puma => Systematic_LN_1-Methoxy-3_4-didehydro-1_2_7__8_-tetrahydro-psi_psi-caroten-2-one => PN_Bim

PathWithLipid: PN_Puma => Systematic_LN_Phorbol => PN_Bim

Fig. 5 shows the results of a query for a pathway fragment involving PI3K kinase and P53 along with lipid-protein annotations.

ACKNOWLEDGEMENTS

A*STAR (Agency for Science and Technology Research, Singapore). National University of Singapore's Office of Life Science (R-183-000-607-712), the Academic Research Fund (R-183-000-160-112) and the Biomedical Research Council of A*STAR (R-183-000-134-305).

REFERENCES

[1] Jang H, etal. BioProber: software system for biomedical relation discovery from, Conf Proc IEEE Eng Med Biol Soc. 2006;1:5779-82

[2] Chen H. and Sharp BM, Content-rich biological network constructed by mining PubMed abstracts, BMC Bioinformatics 2004, 5:147-160.

[3] Garcia O, etal.., GOlorize: a Cytoscape plug-in for network visualization with Gene Ontology-based layout and coloring, Bioinformatics. 2007 Feb 1;23(3):394-6.

[4] Douglas SM, Montelione GT, Gerstein M, PubNet: a flexible system for visualizing literature derived networks, Genome Biology 2005, 6:R80

[5] Baker C.J.O., etal. Towards ontology-driven navigation of the lipid bibliosphere, *BMC Bioinformatics,* vol. 9(Suppl 1), 2008

[6] Sud M., etal. Subramaniam S: LMSD: LIPID MAPS structure database. Nucleic Acids Res 2007, 35:D527-D532.

Issues in learning an ontology from text

Christopher Brewster, Simon Jupp§, Joanne Luciano¶, David Shotton# Robert Stevens§§, and Ziqi Zhang*

**University of Sheffield, §University of Manchester, ¶Harvard University, #University of Oxford*

ABSTRACT

Ontology construction for any domain is a labour intensive and complex process. Any methodology that can reduce the cost and increase efficiency has the potential to make a major impact in the life sciences. This paper describes an experiment in ontology construction from text for the Animal Behaviour domain. Our objective was to see how much could be done in a simple and rapid manner using a corpus of journal papers. We used a sequence of text processing steps, and describe the different choices made to clean the input, to derive a set of terms and to structure those terms in a hierarchy. We were able in a very short space of time to construct a 17000 term ontology with a high percentage of suitable terms. We describe some of the challenges, especially that of focusing the ontology appropriately given a starting point of a heterogeneous corpus.

1 INTRODUCTION

Ontology construction and maintenance are both labour intensive tasks. They present major challenges for any user community seeking to use sophisticated knowledge management tools. One traditional perspective is that once the ontology is built the task is complete, so users of ontologies should not baulk at the undertaking. The reality of ontology development is significantly different. For some large, widely used ontologies, such as the Gene Ontology (Ashburner et al. 2000), a manual approach is effective even if very expensive. For small, scientific communities with limited resources such manual approaches are unrealistic. This problem is all the more acute as research in many areas, including the life sciences, is moving to an e-science industrialised paradigm.

The work presented in this paper concerns the semi-automatic construction of an ontology for the *animal behaviour* domain. The animal behaviour community has recognised the need for an ontology in order to annotate a number of data sets. These data sets include texts, image and video collections. In a series of workshops[1], an initial effort has been made to construct an ontology for the purposes of applying annotations to these data sets. The current Animal Behaviour Ontology (ABO) has 339 classes and the top level structure is shown in Figure 1.

While considerable effort has already gone into the construction of the Animal Behaviour Ontology, its limited size raises the important question as to whether it is more appropriate to slowly build an ontology entirely by hand, and have its potential expansion led by user demand, or whether to rapidly build a much larger ontology based on the application of a variety of text processing methods, and tidy or clean the output. With community engagement comes growth, but there is a question of stimulating engagement through some critical mass of useful ontology. The former approach is the standard approach and has been used successfully in cases such as the Gene Ontology, but becomes more challenging as the size and complexity of the ontology increases. On the other hand, while much has been written about automatic ontology learning, most such work has been undertaken in non-biological domains, or in rather abstract contexts (Cimiano et al. 2005; Brewster et al. 2007; Navigli and Velardi 2004). Although such research is called "ontology learning" in reality, given the limitations of Natural Language Processing, the outputs have been structured

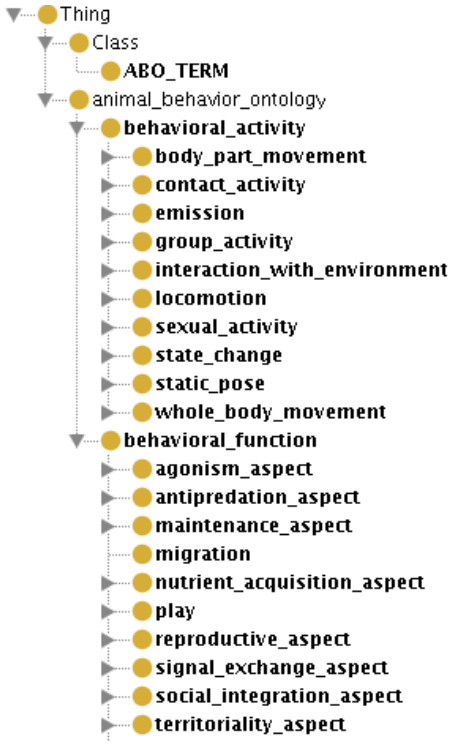


Figure 1 Top level terms in the Animal Behaviour Ontology

* To whom correspondence should be addressed.

[1] For further details cf. http://ethodata.comm.nsdl.org/

Language Processing, the outputs have been structured vocabularies organised in taxonomic hierarchies. This might be considered a major defect if it were not that a) most ontologies are used for labelling/annotation purposes rather than for computational inference, and b) a hierarchically structured vocabulary based on the actual terminology used by a community is a major step towards the creation of a formal ontology. Thus in our view, the construction of formal ontologies of the type needed for driving semantic applications should be considered to involve a significant manual step following the automated process (Luciano and Stevens 2007; Stevens et al. 2007).

In the research reported here, we chose to see how far we could go in the context of limited resources. We approached the challenge as being one to construct a controlled or structured vocabulary as quickly as possible, with minimal effort, and then allow subsequent efforts to clean up the output of this exercise. At one level, we have tried to assess how much effort is worth investing and what is the balance of cost and benefit. A greater understanding of what is the best and most effective methods will in the longer term not only facilitate the creation of useful ontologies for scientific domains with limited resources, but will also facilitate the growing issue of maintenance and upkeep of ontologies as a whole.

2 METHODOLOGY

2.1 The Data Set

It has been argued elsewhere that the only effective way to build representative ontologies for a given domain is through the use of text corpora (Brewster, Ciravegna, and Wilks 2001), and in our case we were able to have access to a considerable corpus of journal articles from the journal *Animal Behaviour*, published by Elsevier. This consisted of articles from Vol 71 (2006) to Vol 74 (2007), containing 623 separate articles. We were given access to text, PDF and XML versions together with a corresponding DTD. We used the XML version for the procedures which are described below.

2.2 From text to ontology

1. Clean text was extracted from the XML files. Using the information from the structured markup, we excluded all author names, affiliations and addresses, acknowledgements, and all bibliographic information, except for the titles of the cited papers.

2. A number of stop word lists and gazetteers were used to further remove noise from the data. We excluded person names as noted above and also through the use of a gazetteer, animal names based on a short list derived from the LDOCE[2], and place names using another gazetteer.

3. A lemmatizer was used to increase coverage (Zhou, Xiaodan Zhang, and Hu 2007). In some cases this generated some noise due to imperfections in the lemmatizer but overall it reduced data sparsity.

4. Five different term extraction algorithms were applied as described in (Ziqi Zhang et al. 2008). The chosen term recognition algorithms were ones that selected both single and multi-word terms as we believe that desirable technical terms are of both sorts. The algorithms were applied to each subsection of the journal article as well as to the whole. This allowed us to look at the terms from different sections of the articles (abstract, introduction, materials and methods, conclusion, etc.). as we aimed to build an ontology of animal behaviour, the terms found exclusively in the "Materials and Methods" section were removed from further consideration. Such terms are the subject of a different ontology.

5a. We then used a set of regular expressions to filter the candidate terms. A regular expression was constructed that looked for terms that ended in *behaviour, display, construction, inspection*, etc. It also included some very generic regular expressions looking for terms that ended in *-ing* and *–ism*. The regular expression used for term selection is available on the website accompanying this research[3].

5b. The step described in 5a. involved quite specific domain knowledge. To have an alternative procedure that does not involve any domain knowledge, we used a voting algorithm to rank the terms and weight them for distribution across the corpus. This was calculated by taking the mean rank for each term and multiplying by the document frequency. From the resulting rankings terms were selected for the subsequent steps (to parallel those extracted by the regular expression).

6a. There are a number of methods that can take a set of terms and try to identify ontological (taxonomic) relations between the terms (Cimiano, Pivk, Schmidt-Thieme, and Staab 2005; Brewster 2007). Most methods suffer from low recall. So in our approach we chose to use the method used in the literature with highest recall – string inclusion. This means that a term A B ***IS_A*** B, and A B C ***IS_A*** (B C and A C) ***IS_A*** C. The resulting ontology was saved in the Web Ontology Language (OWL).

6b. The same method as 6a. was applied to the output of 5b.

7a. and 7b. The resultant ontologies were then filtered for their top level terms i.e. children of THING. A technique used extensively in the ontology learning community is that of using lexico-syntactic patterns (or Hearst patterns (Hearst 1992)) to either learn or test for a candidate ontological relation (Brewster et al. 2007). In this case, we tested each top

[2] The Longman Dictionary of Contemporary English. Our thanks to Louise Guthrie for providing this.

[3] http://nlp.shef.ac.uk/abraxas/animalbehaviour.html

level term in each ontology as to whether it was a kind of *behaviour, activity* or *action* using the Internet as an external resource. Thus we constructed phrases such as the following: *"behaviours such as biting"* (found) or *"behaviours such as dimorphism"* (not found).

3 RESULTS

Figure 2 Partial subtree from ontology at Step 6a.

A total of 64,000 terms were extracted from the whole corpus of 2.2 million words. From this the regular expression extracted 10,335 terms. These included animal behaviour terms, but also included non-animal behaviour terms. The regular expression was designed to capture a large number of terms such as *begging, foraging, dancing, grooming, burrowing, mating*. Due to its crudity it also picked up non-behavioural terms with similar endings: *-bunting, -herring, dichromatism, dimorphism*.

The ontology produced by Step 6a. resulted in an artefact of 17776 classes, of which 1295 classes are top level (i.e. direct children of `OWL:THING`). The ontology produced by Step 6b. from the 10,335 terms selected by the voting algorithm in step 5b. resulted in an artefact of 13,058 classes, of which 2535 classes were top level. The ontologies mentioned here are available on the web site accompanying this paper[4]. A screen shot of the sub tree concerning *call* from ontology 6a. is shown in Figure 2.

The filtering process described in Step 7a. resulted in 383 top level terms being removed leaving 912 immediate descendants of `OWL:THING`. Top level classes that were filtered out by this method included terms such as *stocking referencing, holding, attraction, time, schooling, movement, pacing, defending, smashing, loading, matricide*. The parallel process in 7b. resulted in 649 top level classes being removed, leaving 1886.

A sample of the terms excluded by step 5a. has been evaluated by a biologist (Shotton). Of the 56,000 terms excluded, a random sample of 3140 terms were manually inspected. Of these 7 verbs and 42 nouns were identified as putative animal behaviour-related terms. These included terms such as *forage, strike, secretion, ejaculate, higher frequency yodel, female purring sound,* etc. The low number of significant excluded terms shows that our approach has a *Negative Predictive Value* of 0.98, and a *Recall* of 0.905. We have yet to determine the precision of this approach due to the need for large scale human evaluation of the selected terms.

4 DISCUSSION

A key challenge in the process of learning an ontology from texts is to identify the base units, i.e. the set of terms which will be used as labels in the ontology's class hierarchy. This problem has been largely ignored in the NLP ontology learning literature. The problem of constructing an ontology from a data set such as the one we were using is that in effect there are a number of different domain ontologies represented in the text. In the case of our corpus from the journal *Animal Behaviour*, there existed terms reflecting *experimental methods, animal names, other named entities (places, organisations, people),* etc in addition to behaviours. Such domains are obviously pertinent to animal behaviour (there are species specific behaviours), but the terms exclusively from these domains belong to separate ontologies. The linking together of these separate domains within one ontology is a further step in the process of ontology building.

In order to construct an ontology of animal behaviour from such a heterogeneous data set, one must focus the term selection as much as possible. In order to do this we used first a manually constructed set of regular expressions, an approach which is dependant on domain expertise. As an alternative, for the sake of comparison, we selected the same number of terms using the term recognition voting approach. The ontology generated by this latter approach re-

[4] http://nlp.shef.ac.uk/abraxas/animalbehaviour.html

sulted in less complexity because it included fewer multi-word terms, which using our string inclusion method had generated further intermediate concepts and a richer hierarchy when using the terms identified by regular expressions.

Our initial evaluation of the terms excluded by the regular expressions shows that very few of the omitted terms were significant from an expert's perspective. Our approach will tend to high recall and low precision so there are certainly a significant number of terms included that would need subsequent manual exclusion. A brief consideration of Figure 2 shows a number of terms that would need to be excluded: *g_call, lower call*, etc.

Nevertheless, the resulting ontologies, especially after filtering the top level terms, contains a large number of useful taxonomic fragments even if there is quite a lot of noise. Part of the principle of our approach, as noted in the Introduction, is that it is far easier to collect a large set of potentially significant ontological concepts automatically and then eliminate the noise than to slowly build up a perfectly formed but incomplete set of concepts but which inevitably will exclude a lot of important domain concepts. Such an artefact is far from a formal ontology but is nonetheless useful as a step towards a taxonomic hierarchy for the annotation of research objects, and as a stepping-stone to a more formal ontology. While we still have to undertake a full evaluation, initial assessments indicate the ontologies derived using the regular expressions are cleaner and of greater utility.

The limitations of our approach may be summarised as follows: a) there is a certain amount of noise in the resulting ontologies (which we specify more precisely in future work), b) some effort is involved in *focussing* the ontology produced (i.e. to exclude terms that properly belong to another domain/ontology), c) the result is only taxonomic – the use of string inclusion implies an ISA hierarchy although careful inspection shows that this is not always the case.

The significance of our approach is that it is very quick and easy to undertake. The results produced are very useful, both in themselves as a knowledge discovery exercise in a scientific domain, and as a stepping stone to a more rigorous or formal ontology. The very low effort involved in the process means that this type of data collection could be used in all cases when building ontologies from scratch. We also propose this approach as being a significant tool in ensuring ontologies are up to date and are current with the terminology of a domain.

Future work will include applying the full Abraxas methodology (Brewster et al. 2007) to construct the richest possible structure from the existing ontology. We plan a more extensive evaluation of the noise present i.e. terms that should be excluded. At a more fundamental level, we need to consider how appropriate it is to use terms derived from a corpus for the building of an ontology in contrast to a formally and rigorously hand built ontology.

ACKNOWLEDGEMENTS

We would like to thank Anita de Ward of Elsevier for making the text available from the Journal *Animal Behaviour*. This work was supported by the AHRC and EPSRC funded Archeotools project (Zhang), the Companions project (www.companions-project.org) (IST-FP6-034434) (Brewster), and the Sealife project (IST-2006-027269) (Jupp).

REFERENCES

Ashburner, M. et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25, no. 1:25-29.

Brewster, Christopher. 2007. Mind the Gap: Bridging from Text to Ontological Knowledge. Department of Computer Science, University of Sheffield.

Brewster, Christopher, Fabio Ciravegna, and Yorick Wilks. 2001. Knowledge Acquisition for Knowledge Management: Position Paper. In Proceeding of the IJCAI-2001 Workshop on Ontology Learning, Seattle, WA http://www.dcs.shef.ac.uk/~kiffer/papers/ontolearning.pdf.

Brewster, Christopher et al. 2007. Dynamic Iterative Ontology Learning. In Recent Advances in Natural Language Processing (RANLP 07), Borovets, Bulgaria.

Cimiano, Philipp, Aleksander Pivk, Lars Schmidt-Thieme, and Steffen Staab. 2005. Learning Taxonomic Relations from Heterogeneous Sources of Evidence. In Ontology Learning from Text: Methods, Evaluation and Applications, Frontiers in Artificial Intelligence, IOS Press http://www.aifb.uni-karlsruhe.de/WBS/pci/OLP_Book_Cimiano.pdf.

Hearst, Marti. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In Proceedings of the Fourteenth International Conference on Computational Linguistics (COLING 92), Nantes, France, July 1992.

Luciano, Joanne S, and Robert D Stevens. 2007. e-Science and biological pathway semantics. BMC Bioinformatics 8 Suppl 3:S3.

Navigli, Roberto, and Paula Velardi. 2004. Learning Domain Ontologies from Document Warehouses and Dedicated Websites. Computational Linguistics 30, no. 2:151-179.

Stevens, Robert et al. 2007. Using OWL to model biological knowledge. Int. J. Hum.-Comput. Stud. 65, no. 7:583-594.

Zhang, Ziqi, Jose Iria, Christopher Brewster, and Fabio Ciravegna. 2008. A Comparative Evaluation of Term Recognition Algorithms. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC08), Marrakech, Morocco.

Zhou, Xiaohua, Xiaodan Zhang, and Xiaohua Hu. 2007. Dragon Toolkit: Incorporating Auto-learned Semantic Knowledge into Large-Scale Text Retrieval and Mining. In Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI) http://www.dragontoolkit.org/dragontoolkit.pdf.

GO-WORDS: An Entropic Approach to Semantic Decomposition of Gene Ontology Terms

*Tuanjie Tong, Yugyung Lee, and Deendayal Dinakarpandian**

School of Computing and Engineering, University of Missouri-Kansas City, Kansas City, Missouri 64110, USA

ABSTRACT

The Gene Ontology (GO) has a large and growing number of terms that constitute its vocabulary. An entropy-based approach is presented to automate the characterization of the compositional semantics of GO terms. The motivation is to extend the machine-readability of GO and to offer insights for the continued maintenance and growth of GO. A prototype implementation illustrates the benefits of the approach.

1 INTRODUCTION

The underlying motivation of the work described in this paper is to map annotations based on the Gene Ontology (GO) (Ashburner, et al., 2000) to a semantic representation that exposes the internal semantics of GO terms to computer programs. The Gene Ontology (GO) views each gene product as being a structural component of a biological entity, being involved in a biological process, and as having a molecular function. These three dimensions of component (C), process (P) and function (F) are hierarchically refined into several thousand subconcepts or GO terms for a fine-grained description of gene products, and ultimately a representation of collective biological knowledge. The machine-readability of GO is based on explicit IS-A or PART-OF relations between different GO terms (Fig. 1). The representation of each GO term in terms of a phrase in English is primarily meant for human readability, and not machine-readability (Wroe, et al., 2003) (Fig. 1). For example, while both humans and computer programs can understand that 'Folic Acid Transporter Activity' is one kind of 'Vitamin Transporter Activity," only a human reader can appreciate that proteins annotated with 'Folic Acid Transporter Activity' actually *transport* the vitamin *folic acid*. In other words, the compositional semantics embedded within each GO term is not currently accessible by computer programs; each term *per se* is effectively a black box or meaningless string of characters to computer programs.

It has been estimated that about two-thirds of GO terms (Ogren, et al., 2004) contain another GO term as a substring within it. For example, the GO term 'Transporter Activity' is a substring of several GO terms such as 'Vitamin Transporter Activity' and 'Biotin Transporter Activity.' In other words, many GO terms are combinations of distinct semantic units, as opposed to being a completely new concept. The compositional nature of GO terms has the side effect of resulting in a combinatorial increase in the size of GO. For example, 'Folic Acid' appears in 12 different GO terms like 'Folic Acid Transport,' 'Folic Acid Binding,' and 'Folic Acid Transporter Activity.' Similarly, the vitamin Biotin appears in 23 GO terms, including 6 terms identical to that

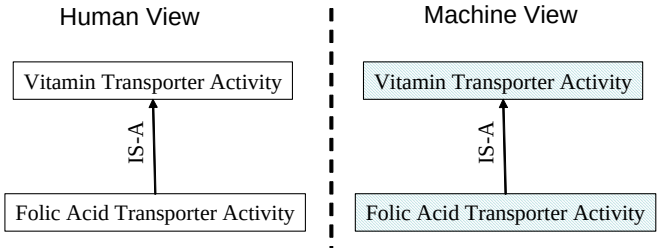


Fig. 1. The internal semantics of GO terms are visible to humans but not to computer programs

for Folic Acid except for the replacement of 'Folic Acid' with 'Biotin,' e.g., 'Biotin Transport,' 'Biotin Binding' and 'Biotin Transporter Activity.' This phenomenon has been one of the motivating factors behind the GO Annotation Tool (GOAT) (Bada, et al., 2004) and the Gene Ontology Next Generation (GONG) project (Wroe, et al., 2003), which suggested having multiple intersecting hierarchies, with a proposed evolution towards a DAML+OIL representation. Reasons for studying the compositional nature of GO are to suggest missing relations (Mungall, 2004; Ogren, et al., 2004), suggest new terms (Lee, et al., 2006; Ogren, et al., 2004), increase computability of GO (Doms, et al., 2005; Ogren, et al., 2004; Wroe, et al., 2003), and for providing models for GO-based analysis of natural language processing of text (Blaschke, et al., 2005; Couto, et al., 2005; Doms and Schroeder, 2005).

One way to discretize GO is to represent it as a language consisting of progressive concatenation of tokens in the form of regular expressions. An example of this is *Obol* (Mungall, 2004), a language that exploits the regularity of GO term names to represent it in Backus-Naur format. However, this is applicable to only a subset of all GO terms.

In this paper, we use an entropic approach for the analysis of regularity of GO term nomenclature. We show how this

*To whom correspondence should be addressed.

may be used to detect sets of GO terms sharing similar semantics. The decomposition of GO terms presented here also suggests a way to minimize the complexity of GO.

2 METHODS

The general principle is to find clusters of GO terms sharing similar semantic structure. Entropy (see below) is used to find GO terms that share consistent location of a specific token (word) within them. Each cluster is evaluated and a corresponding semantic rule created.

Analysis of position-dependent conservation of GO tokens

Each GO term (version Feb 16th, 2006), including synonyms, was tokenized on white space into a sequence of individual words. For example, the GO term "L-amino acid transport" is tokenized as "L-amino" + "acid" + "transport." Entropy (Shannon 1950) is used to measure the regularity in location of each token within all GO terms:

$$EP_t = \sum_{i=1}^{l} - p_i^t \log p_i^t$$

where EP_t is the positional entropy of token t, l is the length (in number of tokens) of the longest GO term or synonym that token t is observed to occur in, and p_i^t is the probability of finding token t at position i. If the logarithm is in base 2, then entropy can be quantified in terms of bits. Recognizing that gene product and molecule names embedded in GO terms consist of a variable number of tokens, we choose to note the position of each token relative to *both* the beginning and end of each GO term. For example, the token "acceptor" almost always occurs at the end of a GO term (with the sole exception of the term "electron acceptor activity"). Thus, it is uniformly the first term when counted from the end of a GO term, with a resulting low positional entropy of 0.08 with respect to the end (EPE). In contrast, this token has a highly variable position when counted from the start of a GO term (as many as 15 different locations) resulting in a high positional entropy (EPS) value of 3.3. If we focus only on an EPS value, we would miss its positional conservation, i.e., tendency to occur at the very end of GO terms.

Since Shannon entropy is based only on proportions, it does not distinguish between token distributions like [1, 1] (token found once at the first position, and once at the second) and [100, 100] (token found a hundred times each at the first and second positions). Both would yield an entropy value of 1 bit even though there are only 2 occurrences of the former and 200 of the latter. To distinguish between such tokens, the absolute numbers of occurrence at a given distance from either the start or end of GO terms are also recorded. The calculated entropies are then 'normalized' (NEP) by adding 0.1 to the calculated value and dividing by the total number of occurrences. Division of the entropic value by the total number of occurrences yields lower values for a higher number of token occurrences. The addition of 0.1 bit helps to distinguish between tokens having an entropy of zero but differing in their frequency of occurrence within GO terms. For the above examples, this would yield values of (0.1/2 = 0.05) and (0.1/200 = 0.0005) respectively, thus yielding a lower NEP value (implying higher degree of positional conservation after correction for more frequent occurrence) for the more frequent token.

Semantic mapping rule generation

Tokens with low positional entropy, high number of occurrences or low normalized positional entropy are used as a starting point for the generation of rules. For each such token, the corresponding set of GO terms is verified for semantic uniformity and a corresponding rule generated. This takes minimal time as the majority of terms in a set follow the same pattern. For example, 'binding' is a token that has much lower entropy when measured from the end (0.28 bit) than from the beginning (2.16 bits). The vast majority, 1544 out of 1597, of GO terms containing the token 'binding' end

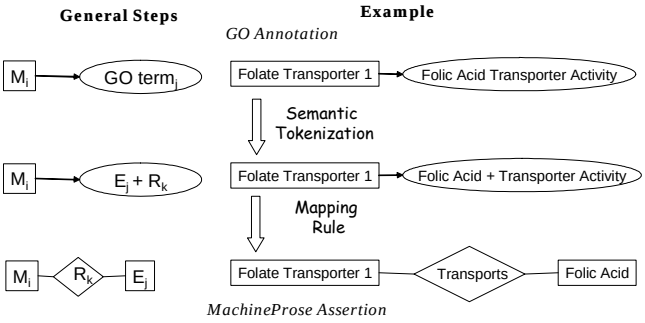


Fig. 2. Mapping a GO annotation to a discretized triplet. The general procedure is shown on the left together with a specific example on the right

with it. 1524 of these are of the general form 'Entity' + 'binding' where 'Entity' represents one or more tokens in succession representing a single concept. The Entity most often specifies a molecule, and sometimes a structural component. The 20 exceptions include terms like 'Protein domain specific binding' and 'regulation of binding.' Thus, the discretizing rule applicable to gene products {Mi} annotated with these GO terms may be stated as 'Mi binds Entity.' In other words, each corresponding GO term (e.g. Zinc Binding) is decomposed into a relational term (e.g. Binds) and the embedded concept (e.g. Zinc). Thus, if the protein "40S ribosomal protein S27" is annotated with the GO term 'Zinc Binding,' then the corresponding discretized semantic form is '40S ribosomal protein S27 Binds Zinc.' Fig. 2 summarizes the general procedure with another example. Triplets of this form correspond to MachineProse assertions (Dinakarpandian, et al., 2006) and can contribute to an incremental knowledge-base distinct from paper publications.

3 RESULTS & DISCUSSION

GO-WORDS browser

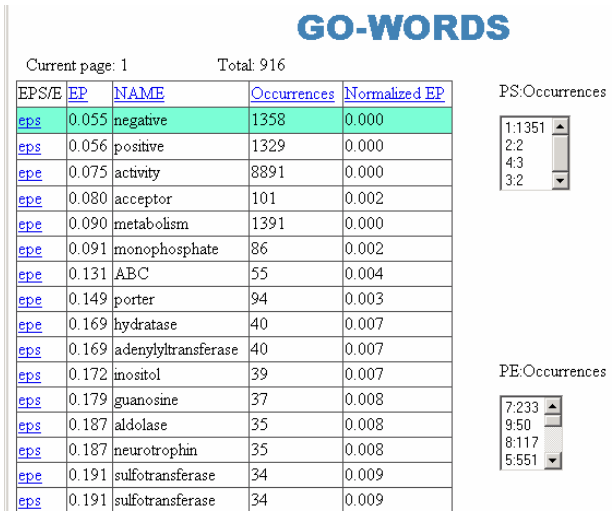


Fig.3. Browser for analyzing tokens/words found within GO terms. Columns 2 and 5 are measures of positional variation of each token within GO terms, column 1 indicates whether position in each row is with respect to the beginning or end of corresponding GO terms, column 3 shows name of token, and column 4 shows number of GO terms it is found in.

Tokenizing GO resulted in a 9152 unique tokens from a total of 37,403 terms (20115 canonical + 17288 synonym terms). Each token occurred 13.7 times on average. The most frequent token was found to be "activity," occurring a total of 8891 times. In contrast, almost half the tokens (4204), e.g. "xylem," occurred only once. We implemented a browser (Fig. 3) to analyze *position-wise* frequencies and entropy of GO-terms. EP stands for entropy. The suffix S, as in EPS indicates that positions were counted from the beginning of the string, whereas the suffix E, as in EPE, indicates that positions were counted backwards from the end of the string. The prefix N indicates normalization (see Methods above). Each token was analyzed using multiple metrics. For example, Table I shows that the token 'negative' has the lowest positional entropy because it occurs most of the time at the beginning of a GO term (1351 out 1358 occurrences with a corresponding EPS=0.055, and normalized EPS=0.00004). In contrast, the token 'oxidoreductase' (not shown) has the highest positional entropy (EPE=3.854;NEPE=0.019) because its 212 occurrences are spread over 29 different positions within GO terms like '*oxidoreductase* activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced pteridine as one donor, and incorporation of one atom of oxygen.' Clearly, it is potentially easier to map GO terms containing the token 'negative' than 'oxidoreductase' to a machine-readable representation.

The GO-WORDS browser is a useful tool to gain insights into the composition of GO terms. With respect to this paper, we focused on using it mainly to create semantic mapping rules. Thus, tokens with low values of NEPE (observed range=0.00004 – 0.562) (Table I) and a large number of occurrences were used to select GO terms for semantic mapping to an assertion representation.

Given a token and position either from the beginning or end of a string, the GO-WORDS browser lists all GO terms and synonyms that share the token at a given position. For example, the token 'transporter' occurs second from the end (517 out of 650) in GO terms like the following:

name: L-ornithine transporter activity
name: S-adenosylmethionine transporter activity
exact_synonym: S-adenosyl methionine transporter activity
name: adenine nucleotide transporter activity
name: spermine transporter activity
name: sulfite transporter activity

Table I. Tokens with lowest normalized positional entropy

Token	Normalized Entropy	Token	Normalized Entropy
activity	nepe=0.000	dehydrogenase	neps=0.001
negative	neps=0.000	cell	nepe=0.001
positive	neps=0.000	complex	nepe=0.001
metabolism	nepe=0.000	metabolism	neps=0.001
activity	neps=0.000	receptor	neps=0.001
binding	nepe=0.000	biosynthesis	neps=0.001
regulation	neps=0.000	transporter	nepe=0.001
of	neps=0.000	binding	neps=0.001
of	nepe=0.000	formation	neps=0.002
biosynthesis	nepe=0.000	ligand	nepe=0.002
pathway	nepe=0.000	catabolism	neps=0.002
regulation	nepe=0.001	transport	nepe=0.002
formation	nepe=0.001	cell	neps=0.002
anabolism	nepe=0.001	synthesis	neps=0.002
synthesis	nepe=0.001	acid	nepe=0.002
differentiation	nepe=0.001	proliferation	nepe=0.002
catabolism	nepe=0.001	acceptor	nepe=0.002
receptor	nepe=0.001	degradation	neps=0.002
breakdown	nepe=0.001	exocytosis	nepe=0.002
degradation	nepe=0.001	anabolism	neps=0.002

The general pattern for the above examples is "Entity transporter activity." Thus, the mapping rule applicable to gene products {Mi} annotated with these GO terms may be stated as 'Mi transports Entity,' where entity is presumed to be the prefix of 'transporter activity.' This assumption is true in 420 of the 440 cases. Exceptions to the rule include terms like "siderophore-iron (ferrioxamine) uptake transporter activity" and "transporter activity." In the former, only a subset of the prefix of "transporter activity" represents an

Entity, i.e, the word 'uptake' doesn't conform to the same pattern. The latter is the parent term representing the abstract concept of 'transporter activity.'

The GO token entropic measure helps in clustering terms that share a token at the same relative position. Based on the general patterns 'Entity binding' and 'Entity transporter activity,' 23780 and 903 annotations respectively were mapped to discretized triplets. However, the entropic analysis is based on the naïve assumption that each token represents a concept. In reality, names of entities often consist of a variable number of words strung together, e.g., lipoprotein lipase. Measuring the positional entropy of a token from either end helps mitigate this problem to an extent, but only to an extent. In particular, GO terms where the token of interest is flanked by entities of variable length will not show a peak in the positional distribution. Further, since it is based purely on a textual approach (no prior semantics), manual verification is required to find sub-concepts that are made up of contiguous tokens.

4 CONCLUSION

This paper has presented and addressed the advantages of a discretized triplet representation of GO annotations and a partially automated approach for doing so. In future, we intend to extend the approach to the entire Gene Ontology, combine information from other sources, and devise a sophisticated search interface that shall incorporate the MachineProse relation ontology (Dinakarpandian, et al., 2006).

The number of terms in GO has been rapidly growing since its inception (Ashburner, et al., 2000). The total number of terms has grown from 4507 in 2000 to more than 20,000 in Feb 2006 (Gene Ontology Consortium). One reason is a richer description, but redundancy of nomenclature is also a factor. As GO is continuously revised (terms becoming obsolete, renamed and rearranged), maintaining its semantic integrity is quite challenging. This paper suggests an approach to a leaner GO that is both people and machine friendlier by allowing annotations to be built from reuse of semantically defined building blocks. This would lessen the growth rate of GO, with the resultant smaller size helping in ensuring uniformity and semantic consistency of GO. The benefits would be easier maintenance of GO and higher semantic transparency. In the interim, a triplet view of GO annotations offers a pragmatic solution. A potential advantage is to facilitate searches specified as a set of triplets, occupying the middle ground between a natural language interface and a keyword-based one.

Since a large number of entities within GO are general or specific names of molecules, extracting the embedded molecular ontology would be a useful adjunct. Using other ontologies like ChEBI (ChEBI) and completed mappings between GO and other ontologies (Johnson, et al., 2006) would be useful in this regard.

ACKNOWLEDGEMENTS

We would like to acknowledge funding support for this work from the University of Missouri-Kansas City Research Board to DD (FRG 2006) and from the University of Missouri Research Board to YL (UMRB 2005).

REFERENCES

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat Genet*, **25**, 25-29.

Bada, M., Turi, D., McEntire, R. and Stevens, R. (2004). Using Reasoning to Guide Annotation with Gene Ontology Terms in GOAT. SIGMOD Record, **33**(2).

Blaschke, C., Leon, E.A., Krallinger, M. and Valencia, A. (2005) Evaluation of BioCreAtIvE assessment of task 2, *BMC bioinformatics*, **6 Suppl 1**, S16

ChEBI Chemical Entities of Biological Interest. http://www.ebi.ac.uk/chebi/

Couto, F.M., Silva, M.J. and Coutinho, P.M. (2005) Finding genomic ontology terms in text using evidence content, *BMC bioinformatics*, **6 Suppl 1**, S21.

Dinakarpandian, D., Lee, Y., Vishwanath, K. and Lingambhotla, R. (2006) MachineProse: an Ontological Framework for Scientific Assertions, *J Am Med Inform Assoc*, **13**, 220-232.

Doms, A., Fuche, T., Burger, A. and Schroeder, M. (2005) How to Query the GeneOntology. *Proceedings of Symposium on Knowledge Representation in Bioinformatics*. Finland.

Doms, A. and Schroeder, M. (2005) GoPubMed: exploring PubMed with the Gene Ontology *Nucleic Acids Res.*, W783-W786.

Gene Ontology Consortium. Gene Ontology Annotations. http://www.geneontology.org/

Johnson, H.L., Cohen, K.B., Baumgartner, J.W.A., Lu, Z., Bada, M., Kester, T., Kim, H. and Hunter, L. (2006) Evaluation of Lexical Methods for Detecting Relationships Between Concepts from Multiple Ontologies. *Pacific Symposium on Biocomputing*. 28-39.

Lee, J.B., Kim, J.J. and Park, J.C. (2006) Automatic extension of Gene Ontology with flexible identification of candidate terms, *Bioinformatics*, **22**, 665-67

Mungall, C.J. (2004) Obol: integrating language and meaning in bio-ontologies, *Comparative and Functional Genomics*, 509-520.

Ogren, P.V., Cohen, K.B., Acquaah-Mensah, G.K., Eberlein, J. and Hunter, L. (2004) The compositional structure of Gene Ontology terms, *Pacific Symposium on Biocomputing*, 214-225.

Shannon, C.E. (1950) Prediction and entropy of printed English. The Bell System Technical Journal, **30**, 50-64

Wroe, C.J., Stevens, R., Goble, C.A. and Ashburner, M. (2003) A methodology to migrate the gene ontology to a description logic environment using DAML+OIL, *Pacific Symposium on Biocomputing*, 624-635

Using ontologies to annotate large-scale mouse phenotype data

Tim Beck[*]*, Ann-Marie Mallon, Hugh Morgan, Andrew Blake and John M. Hancock*

MRC Harwell, Harwell, Oxfordshire, OX11 0RD, U.K.

ABSTRACT

The annotation of mouse phenotype data generated during a large-scale primary phenotyping project is underway. Utilising OBO ontologies, a framework has been developed which incorporates two existing annotation approaches to form coherent and precisely defined descriptions of phenotyping procedures, the parameters of procedures and the data derived for each parameter. We introduce the storage of combinatorial phenotype ontology annotations at the database level with the use of interface controlled vocabularies, incorporating compound phenotype ontology terms, to assist with phenotype capture at the point of data entry and subsequent database querying.

1 INTRODUCTION

Two differing approaches can be adopted to annotate phenotype data with bio-ontologies. Either a single dedicated ontology of compound terms can be employed or an annotation can be built using terms from a number of distinct ontologies to form a more complex expression to describe an aspect of an organism's phenotype [1]. The Mammalian Phenotype (MP) ontology [2] is an example of a single dedicated phenotype ontology and the PATO model [3] of defining phenotypes in terms of an entity (E) which has a quality (Q) to build E+Q annotations is an example of the combinatorial approach. PATO is an ontology of species neutral phenotypic qualities and as such lends itself to the formation of comparable cross-species and cross-database phenotypic statements. Using the mouse kinked tail dysmorphological phenotype as an example, MP defines this phenotype using the single term *kinked tail* (MP:0000585) and PATO is used to assign a quality to the mouse anatomical entity defined by the Mouse Anatomy (MA) ontology to form the annotation E: *tail* (MA:0000008) and Q: *kinked* (PATO:0001798).

MP has been widely implemented within database resources with the Mouse Genome Informatics and the Rat Genome Database providing associations between genes and MP terms. However, although recently used for the description of phenotypes observed during zebrafish screens [4], there has, up to now, not been any such comprehensive implementation of the PATO combinatorial approach within mammalian phenotype related informatics resources.

The European Mouse Disease Clinic (EUMODIC, http://www.eumodic.org) is a major European project which is undertaking a primary phenotype assessment of up to 650 mouse mutant lines derived from embryonic stem (ES) cells developed in the European Mouse Mutagenesis (EUCOMM) project. The phenotype assessment consists of a selection of Standard Operating Procedures (SOPs) from the European Mouse Phenotyping Resource of Standardised Screens (EMPReSS, http://empress.har.mrc.ac.uk) [5] organised into two primary phenotyping pipelines. There is a wide range of screens collecting phenotype data from the mouse biological domains of morphology and metabolism; cardiovascular; bone; neurobehavioral and sensory; haematology and clinical chemistry and allergy and immunity. As a result of carrying out an individual SOP either quantitative data (e.g. blood pressure measurement), qualitative data (e.g. coat color description) or a combination of quantitative and qualitative data (e.g. cornea opacity description and the precise opacity level measurement) can be returned. The date derived from carrying out the phenotyping pipelines are stored in the EuroPhenome mouse phenotyping resource (http://www.europhenome.eu) [6].

In order to unify the reporting of results from unrelated mouse experimental procedures from different EUMODIC research institutions, the data is converted to XML format from the local Laboratory Information Management System (LIMS) before submission to EuroPhenome. Each XML file complies with the Phenotype Data XML (PDML) schema. PDML builds upon the Minimal Information for Mouse Phenotyping Procedures (MIMPP), a minimum information checklist which is under development to cover all mouse phenotyping domains. MIMPP is a member of the Minimal Information for Biological and Biomedical Investigations (MIBBI) project whose goal it is to ensure descriptions of methods, data and analyses support the unambiguous interpretation and reuse of data [7].

Each SOP has a number of parameters which define the type of data to be recorded for a specific component of the SOP. Either the parameter value will contain quantitative data, for which the SI unit is specified, or it will be qualitative data.

[*] To whom correspondence should be addressed.

Qualitative data has traditionally been recorded using free-text or from a limited set of options. The application of phenotype ontologies in the capture of qualitative data will lead to more coherent and descriptive datasets. The structure of the PATO quality hierarchy lends itself favorably to the annotation of parameters and the subsequently derived data. For example, a parameter could be coat hair texture and as a result of carrying out the SOP it is found to be greasy. The quality *greasy* [is a] *texture* quality within PATO, as are other potential mouse coat hair textures.

2 METHODS

The ontological annotation of mammalian phenotype data was undertaken on three levels: the annotation of SOPs; the annotation of individual SOP parameters and the annotation of the data derived for each parameter. A distinction is drawn between qualitative and quantitative phenotype data as the annotation of these two classes of data is handled differently.

2.1 SOP and parameter annotation

The SOPs were annotated using high-level MP terms to give a general description of the procedure and provide a global summary of all parameters within the SOP. The individual parameters for each SOP were defined using the E+Q combinatorial approach in collaboration with scientists with expert knowledge in each domain. European institutions participating in EUMODIC record their primary phenotype data using their in-house Laboratory Information Management System (LIMS), however the list of parameters for each SOP is standardised. An overriding factor in the process of parameter definition, especially while defining parameters for qualitative data, was making the parameters intuitive to the phenotyping scientists who would be interacting with the local LIMS. A desirable situation with respect to data accuracy and consistency would be one where original LIMS entries could be imported directly into the EuroPhenome data schema, therefore requiring that non-informaticians should be able to work seamlessly with ontologies. Given the large number of entries into the local LIMS which would be required for a single SOP during the lifetime of EUMODIC and the associated time cost, it was essential that the practical implementation of ontology terms to define parameters was accessible to phenotyping scientists. As a result of this process it was discovered that ontology classes and metadata did not exist to define anatomical entities using terminology that was understandable to phenotyping scientists. These omissions were dealt with by either proposing new terms for the MA ontology, submission of synonyms of existing terms or requests for term definitions.

2.2 Data annotation

Qualitative data, for example dysmorphology data, requires the objective analysis of data at the point of data entry. Qualitative phenotypes, for example variations in coat colours, are compared to wild-type mice and the researcher responsible for making the comparison must first make the decision as to whether a mouse is different and if it is, how it is different. The use of ontologies in capturing qualitative data at the point of data entry is desirable, since it would reduce the ambiguity associated with interpreting free-text and the subsequent mapping to an ontological structure. For this reason the allowed values that could be assigned to a qualitative parameter were restricted to PATO qualities, specifically qualities that were child terms to the parameter defining quality. This process, in unison with the definition of parameters, was carried out in collaboration with phenotyping scientists.

2.3 Interface parameter annotation using compound terms

The coherent and precisely defined E+Q structure of parameters and values lead to a marked increase in the number of parameters to be evaluated for each qualitative SOP. Additionally the decomposition of some community standard phenotypic terms to E+Q phenotype statements proved problematic for phenotyping scientists to relate to during data entry (see belly spot example below). For these reasons interface parameters were defined for qualitative SOPs whereby intuitive compound MP terms defined parameters and also could be assigned as values for parameters, instead of simply assigning a child PATO quality to a parameter PATO quality. The interface lists, while ensuring all value options were restricted and so eliminating the need for free-text, also ensured that all interface parameters values were mappable to the original formal E+Q parameters.

3 RESULTS

The need to develop interface controlled vocabularies that contain compound phenotype terms which are intuitive for use by mouse phenotyping scientists at the point of qualitative data entry presents an important data capture question. If the phenotype data is to be accessible how should it be contained within the database and subsequently queried and presented to users? All EMPReSS SOP parameters are currently in the process of being defined using E+Q terms. In addition to anatomical entities the ontological domains of biological chemicals (CHEBI) and behaviors are also associated with qualities. Where qualitative data is concerned the E+Q annotation is stored directly in the database.

Discussions with scientists during this practical ontology annotation process has revealed that there is a preference for

interacting with the database, either at the points of data entry or querying, via community standard compound phenotype ontology terms where complex qualitative phenotypes are concerned. It is recognised that for some compound terms, when deconstructed into E+Q format, they may lose their biological meaning. For example the term *belly spot* (MP:0000373) is deconstructed to *spotted* (PATO:0000333) [has quality] *white* (PATO:0000323) [inheres in] *coat hair* (MA:0000155) [part of] *abdomen* (MA:0000029) (G.V. Gkoutos, personal communication). A solution, as will be implemented within EuroPhenome, is to store the phenotype in the database in the deconstructed format but allow entry of the data and subsequent querying via the compound term, so in this example *belly spot*. The process of term mapping allows the interface parameter lists to be implemented within local LIMS and then the phenotype to be imported into the EuroPhenome database in a format which complies with the E+Q parameter lists.

As the use of bio-ontologies to define mouse phenotype observations becomes increasingly commonplace it is essential that the ontologies are accessible and understandable by those scientists who will make use of them and benefit from their implementation the most. This demographic is no longer restricted to ontologists or bioinformaticians, who will continue to play an essential role in developing and maintaining ontologies, but includes the "wet-science" researchers who will want to query large data sets using meaningful ontological terms and relationships in order to find phenotypes of interest. A specific example taken from the EUMODIC project would be scientists from secondary phenotyping clinics who will want to identify individual mice exhibiting relevant mutant phenotypes from EuroPhenome which will then undergo secondary phenotyping procedures. These researchers will also become increasingly responsible for entering their data into databases, albeit with appropriate quality control mechanisms in place, so the descriptive power of ontologies must be exploited to ensure they are as scientist friendly as possible. We have identified a number of omissions of terms from MA, for example *nose skin*, which were regarded as essential for the precise categorisation of phenotypes. In other cases existing terms were not intuitive to scientists and synonyms were suggested, for example *hind paw* as a synonym of *foot* (MA:0000044) and *skull* as a synonym of *head bone* (MA:0000576). Terms were also identified which required definitions in order to convey any useful meaning, for example *foot digit 1* (MA:0000465) and *hand digit 4* (MA:0000457). Our suggestions were passed onto MA curators. It is only through the practical application of phenotype ontologies that omissions and potential improvements such as these will be identified.

4 DISCUSSION

We have described the ongoing efforts within the EuroPhenome mouse phenotyping resource to implement both the MP and the E+Q combinatorial approach to systematically annotate real mouse phenotypes, derived from community approved SOPs, on a large scale. The three levels of annotations sees the marrying together of the two different phenotype annotation approaches into a framework that facilitates both data accessibility to mouse scientists using familiar terminology and also cross-database and cross-species phenotype statement comparisons through the storage of phenotypes in the E+Q format at the database level. Future interfaces for the querying of EuroPhenome data will exploit mappings between MP and E+Q terms to accommodate the direct retrieval of E+Q annotations in addition to querying via MP.

Currently only qualitative phenotypes are annotated with ontologies. Quantitative data for baseline and mutant strains across all phenotyping centers are entered into EuroPhenome. Comparative analyses between these two sets of data allude to statistically significantly differences. Those mutant mice, which display significantly different values, are then objectively annotated with MP and E+Q terms. The annotation of quantitative data is therefore dynamic depending on the statistically significant characteristics queried. It therefore follows that as sample sizes increase with the amount of data in the EuroPhenome database (as the result of future mouse strains undergoing phenotyping), the confidence in statistically deduced phenotype ontology annotations will in turn increase.

The work described here involving ontologies has focused on the use of high-level definitions of SOPs, the definition of parameters recorded as a result of carrying out the SOP and the description of the data derived for each parameter. Current research is focused on the development of an assay ontology which will provide coherent definitions of each individual procedural component contained within a SOP. The context of a specific phenotype E+Q annotation would be defined with the inclusion of this procedural data into a phenotype data capture schema as illustrated in Fig 1. Where qualitative data is available this would be stored within the database and qualifies the phenotype quality value, which is a PATO child term of the parameter quality.

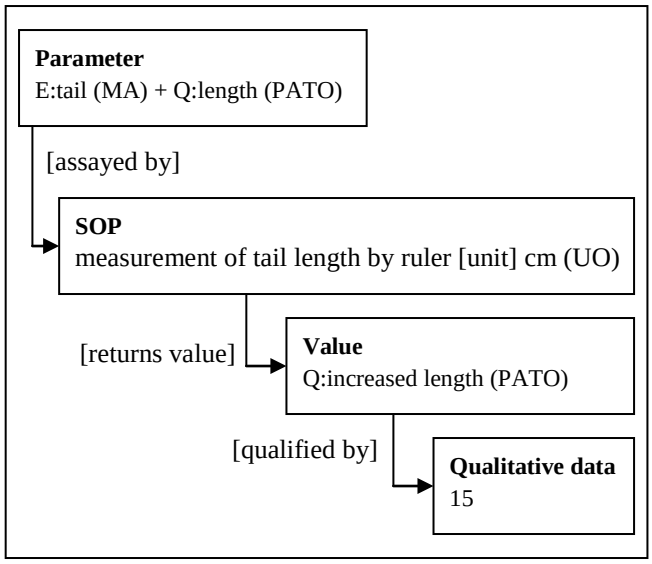


Fig 1. Phenotype and procedural data capture schema to describe an instance of tail length. An assay ontology will define individual SOP procedural components.

ACKNOWLEDGEMENTS

This research was funded as part of the EUMODIC project (funded by the European Commission under contract number LSHG-CT-2006-037188). The authors also thank the UK Medical Research Council for support.

REFERENCES

1. Hancock, J.M. and A.M. Mallon, *Phenobabelomics--mouse phenotype data resources.* Brief Funct Genomic Proteomic, 2007. **6**(4): p. 292-301.
2. Smith, C.L., C.A. Goldsmith, and J.T. Eppig, *The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information.* Genome Biol, 2005. **6**(1): p. R7.
3. Gkoutos, G.V., et al., *Using ontologies to describe mouse phenotypes.* Genome Biol, 2005. **6**(1): p. R8.
4. Knowlton, M.N., et al., *A PATO-compliant zebrafish screening database (MODB): management of morpholino knockdown screen information.* BMC Bioinformatics, 2008. **9**: p. 7.
5. Green, E.C.J., et al., *EMPReSS: European Mouse Phenotyping Resource for Standardized Screens.* Bioinformatics, 2005. **21**(12): p. 2930-2931.
6. Mallon, A.M., A. Blake, and J.M. Hancock, *Euro-Phenome and EMPReSS: online mouse phenotyping resource.* Nucleic Acids Res, 2008. **36**(Database issue): p. D715-8.
7. Taylor, C., et al., *Promoting Coherent Minimum Reporting Requirements for Biological and Biomedical Investigations: The MIBBI Project.* Nat. Biotechnol., 2008.

Developing an application focused experimental factor ontology: embracing the OBO Community

James Malone*, Tim F. Rayner, Xiangqun Zheng Bradley and Helen Parkinson

EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

ABSTRACT

Motivation: The recent crop of bio-medical standards has promoted the use of ontologies for describing data and for use in database applications. The standards compliant ArrayExpress database contains data from >200 species and >110,000 samples used in genotyping, gene expression and other functional genomics experiments. We considered two possible approaches in employing ontologies in ArrayExpress: select as many ontologies as cover the species, technology and sample diversity, choosing where there are non-orthogonal resources and attempt to make them interoperable; or build an extensible interoperable application ontology. Here we describe the development of an application focused Experimental Factor Ontology and describe its use at ArrayExpress.

www.ebi.ac.uk/ontology-lookup/browse.do?ontName=EFO

1 INTRODUCTION

The value of having explicit and rich semantic representations of data is becoming increasingly clear in bioinformatics. This is apparent in the emergence of the OBO foundry (Smith *et al*., 2007) and numerous metadata standards (http://www.mibbi.sf.net). The OBO foundry promotes the development of orthogonal ontologies that are expressed in a common shared syntax, use unique namespace identifiers and explicit textual definitions for all ontology terms. These ontologies give us the terminology to describe the level of detail that content standards such as MIAME require. Underpinning this increased focus on the use of ontologies is that richer and explicit representations enhance interoperability and facilitate machine readability. As the numbers of ontologies and standards increase, the complexity of supporting standards using ontologies also increases.

In this paper we describe development of the Experimental Factor Ontology (EFO), an application focused ontology. EFO models the experimental variables (e.g. disease state, anatomy) based on an analysis of such variables used in the ArrayExpress database. The ontology has been developed to increase the richness of the annotations that are currently made in the ArrayExpress 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 (Dyck and Chisholm, 2003) and Cell Type Ontology (Bard *et al*, 2005). This is achieved using a combination of automated and manual curation steps and the use of a phonetic matching algorithm. This mapping strategy allows us to support the needs of various ArrayExpress user groups who preferentially use different ontologies, to validate existing ontologies for coverage of real world high throughput data in public repositories and to provide feedback to the developers of existing ontologies. An additional reason to have a local application ontology – rather than simply create an enormous cross product ontology (i.e. classes created by combining multiple classes from other ontologies)– is that the structure of such an ontology may be challenging for many users and time consuming to produce (Bard and Rhee, 2004). Instead, data acquisition tools can employ one ontology rather than many external ontologies.

Brinkley *et al.* (2006) highlight the potential value in reference ontologies for performing mapping and integration for building application ontologies. However, at present these frameworks and all necessary reference ontologies do not exist. We therefore exploit the use of the several OBO Foundry ontologies as reference ontologies in contrast to the definition discussed by Brinkley *et al* by employing a softer and more cautious view of these ontologies. Specifically, we aim to map to the concept names and definitions provided by external ontologies without importing covering axioms, thereby reducing the potential for conflict and removing an obstacle for interoperability. Instead we use references in the same way many OBO Foundry ontologies reference external resources using a pointer to their identifier. This strategy avoids 'bedroom ontology development' wherein ontologies are developed *ab initio* without considering the reuse of existing ontologies. By re-using and mapping we leverage the user supplied annotations and existing ontologies.

The EFO is represented in the web ontology language (OWL) thereby conforming to an accepted common representation and we also implement a policy of unique

[*] **To whom correspondence should be addressed.**
Email: malone@ebi.ac.uk

namespace identifiers and definitions for all terms, as encouraged by OBO. Finally, we assess our ontology *post-hoc* using semi-automated methods to assess the coverage we have obtained in terms of our set of use cases (described in our web resource http://www.ebi.ac.uk/microarray-srv/efo/index.html) and, hence, assess the suitability of the ontology for the task at hand.

2 METHODOLOGY

Since the EFO is an application ontology, we developed a well defined set of requirements based on our needs for annotating experimental data. ArrayExpress typically has ~ five annotations per biological sample, and the most important annotations are those that contain information on the experimental variables. These are both biological i.e. properties of the experimental samples (e.g. sex or anatomy), and procedural; properties of protocols used to treat the samples (e.g. sampling time or treatment with compound). The initial focus in developing the EFO is on the former as they are more likely to be present in a reference ontology (i.e. non-numeric) and can be automatically discovered in unstructured data. This is an important use case for ArrayExpress as thousands of experiments are imported from the Gene Expression Omnibus where the sample annotation is essentially un-curated free text. Additionally from analysis of user queries, biological information is more commonly queried than procedural information.

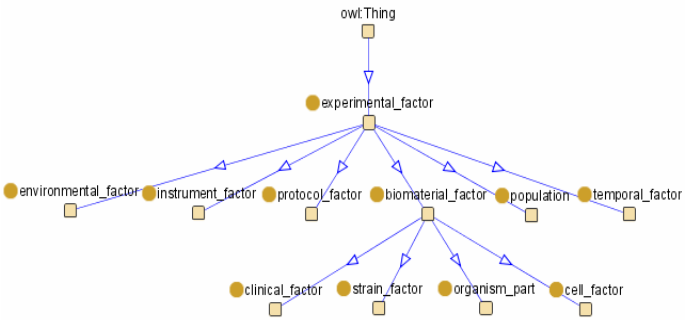


Figure 1 High level classes from the EFO

2.1 Mapping, curating and integrating

Our approach is a middle-out ontology methodology as described by Uschold and Gruninger (1996). In this method, we start with a core of basic terms identified from our use cases and specialize and generalize as required. Our set of initial core terms already provided some structure as the more specific concepts (called factor values) were grouped into factor categories. We then created generalized classes to give some additional structure to our ontology (shown in Figure 1). The structure at the highest level has been designed to be simple, and intuitive to biologists and the curators, who will be the primary users of the EFO in the short term, by constructing this as an abstraction of the existing structure in ArrayExpress.

EFO terms have no internal text definitions by design, instead we leverage the mapping strategy defined below to create links to text definitions created by domain experts.

The mapping strategy involves selecting likely reference ontologies and evaluating their coverage of terms present in ArrayExpress. This includes ontologies such as the Disease Ontology which also has many mapped terms, the Cell Type Ontology and Zebrafish Anatomy and Development ontology (Sprague *et al.*, 2006) and the NCI thesaurus (Fragoso *et al.*, 2004) which has human, mouse and rat terms related to cancer (Figure 2).

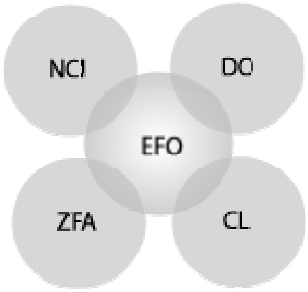


Figure 2 The intersection of the EFO and reference ontologies

To perform our mapping and add terms to EFO we used the following iterative methodology:

- Identify OBO Foundry ontologies relevant to an EFO category based on annotation use cases
- Create subset of classes of relevance to the ontology, e.g. classes under disease for disease ontology
- Perform mapping using text mining phonetic matching algorithm. This produces a list of candidate ontology class matches.
- Manually validate matched ontology classes and curate where necessary
- Manually map high quality annotations (identified as present in the ArrayExpress data warehouse) to multiple source ontologies
- Consider number of instances of terms used in ArrayExpress to determine depth and breadth
- Integrate into EFO, adding appropriate annotation values to definition and external ontology ID
- Structure EFO to provide an intuitive hierarchy with user friendly labels

2.2 Phonetic matching

Our matching approach uses the Metaphone (Phillips, 1990) and Double Metaphone algorithms (Phillips 2000) which were selected following an empirical study of commonly used matching algorithms and their utility in the biomedical

domain. We were particularly interested in algorithms yielding low false positive rates, as we wished to use the same algorithm for automatic annotation of incoming data.

We matched the user supplied cell type terms deposited in ArrayExpress with the Cell Type Ontology using Soundex (http://en.wikipedia.org/wiki/Soundex), Levenshtein edit distance (Levenshtein, 1966), Metaphone (Phillips, 1999) and Double Metaphone (Phillips, 2000) algorithms. Synonyms and term names were used during the matching process and matches were either single or multiple. For the purposes of automated annotation, single matches are obviously more desirable. The Metaphone algorithm yielded the lowest false positive rate, with 98% of the matches mapping to single ontology terms, and of these only 6% were deemed to be invalid following inspection by an expert curator. However, the overall coverage of the input term list was relatively low (17% of all terms matched). In comparison, the Double Metaphone algorithm provided higher list coverage (50% of terms) at the expense of generating a smaller proportion of single matches (48% of total matches) and a much higher false positive rate (34% of single matches). The Levenshtein and Soundex algorithms yielded results similar to the Metaphone and Double Metaphone algorithms, respectively, but both generated slightly higher levels of false positives. A combined strategy was therefore implemented, using Metaphone for a first pass and then falling back to Double Metaphone for those terms not matched by Metaphone. Using this strategy with curator supervision to select the correct term in the multiple-match cases yielded the highest overall number of matches with minimal human intervention. Verified matched terms identified by this strategy were included in the EFO and placed manually in the hierarchy.

2.3 Ontology conventions

Naming conventions described by Schober *et al.* (2007) were used. Specifically, class labels are intended to be meaningful to human readers, short and self-explanatory. They are singular and conform to the conventional linguistic and common usage of the term, for example, the term *Huntingdon's disease* has a capital H since it is a proper noun, whereas *cancer* would not. Identifiers have the format EFO:00000001, where a unique integer identifies a term and EFO identifies the ontology. We use an alternative term annotation property to capture synonyms for class labels, text definitions are not provided at present. The ontology is developed in Protégé and converted to OBO format for display in OLS.

3 THE EFO

Part of the hierarchy visualized in OLS is shown in Figure 3. The current version of EFO has ~800 child terms of the class experimental factor. The majority of these have been mapped to external reference ontologies and knowledge resources, as indicated by the definition citation annotation property.

As an early version, the ontology still has parts that are under review and is evolving. In particular, the hierarchy still contains classes that are likely to be moved and changed to add more structure as it is relatively flat at present. Furthermore, the additional group of use case covering cross species queries, e.g. disease and mouse model of disease, and the representation of anatomical parts in different species are required but are currently not supported by the EFO. However, as the iterative engineering process is ongoing, these will be addressed in the near future. Where possible we will use existing resources to address these use cases.

3.1 Validation

The ArrayExpress data flow doubles on a yearly basis. This allows us to constantly validate the ontology against fast changing annotation with a variety of granularities. It also allows us to develop the ontology against emerging use cases. We have implemented an iterative evaluation of the ontology against the data content of the ArrayExpress repository, against newly submitted data for curation purposes and also against the ArrayExpress data warehouse – a set of additionally annotated and curated data which represents the ArrayExpress 'gold standard'. As the ontology evolves it will be used daily by the ArrayExpress production team and incremental versions will be tested internally prior to public release. Early stage evaluation is performed semi-automatically by mapping between the ontology and very large meta-analyzed curated experiments and by comparison with reference ontologies. We were able

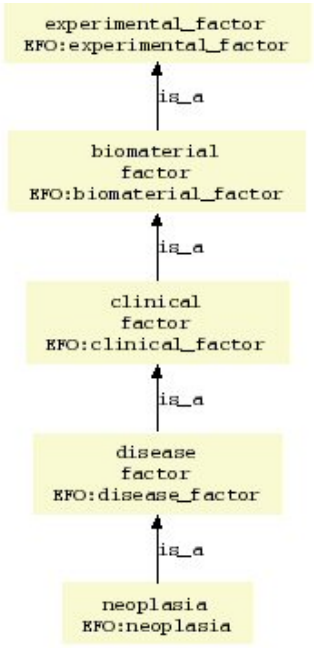


Figure 3 EFO term 'neoplasia' visualized in OLS

to assess granularity and overall coverage of the ontology and structure is manually evaluated by the curators who use the ontology.

Version 0.1 of EFO produces automated mappings comparable in coverage (~35%) for a 6000 sample test set between the EFO and the NCI thesaurus. Replacing the NCI thesaurus with the EFO reduced false positives and multiple matches by an order of magnitude (60% reduced to 8.6%). We believe by continuing an iterative process of mapping, curating and integrating EFO terms alongside an iterative evaluation strategy and restructuring the ontology we can continue to improve the quality and coverage of the ontology throughout its lifecycle.

4 DISCUSSION

It is our belief that application ontologies such as the EFO should be constructed with a principal to minimize redundancy and maximize information sharing. Wherever possible, mapping to external resources such as OBO Foundry ontologies increases interoperability through a common and shared understanding. Furthermore, this removes the temptation to 'reinvent the wheel' and allows the exploitation of the efforts currently underway to represent particular communities. It also permits updating when reference ontologies change.

A complication of this approach is the implication of mapping to external ontology concepts and their implicit hierarchy. In EFO our 'meaning' is limited to the textual definitions of the concepts externally mapped to EFO terms. Importing and accepting all axioms associated with concepts is a desirable long term goal. However the potential for conflicting logical definitions and lack of an intuitive standardized and easy to use upper ontology framework have caused us to initially defer this task. BFO (Grenon *et al*., 2004) was not considered as an upper level ontology for EFO in its earliest form as the primary focus of this project is the application of the ontology and rapid development. However, mapping to BFO (or some other upper level ontology) is something we are now beginning to look into.for future development and will appear in the forthcoming future releases.

The OBO Foundry has resolved issues, of orthogonal coverage and unique namespace identifiers and has made our task easier. In the future we will make bimonthly releases of EFO, continue the validation process, consider requests for new terms and map additional data resources to the EFO. GEO data imported into the ArrayExpress framework is already mapped during import, and any data resource with biological annotation could be mapped semi-automatically. Obvious candidates include Uniprot and other gene expression databases which are targets for integration with ArrayExpress. Version 0.2 of EFO is available from the EBI Ontology Lookup Service, comments and questions can be sent to exfactorontology@ebi.ac.uk

ACKNOWLEDGEMENTS

The authors are funded in part by EC grants FELICS (contract number 021902), EMERALD (project number LSHG-CT-2006-037686), Gen2Phen (contract number 200754) and by EMBL**.** Thanks to the ArrayExpress production team: Anna Farne, Ele Holloway, Margus Lukk, and Eleanor Williams for useful comments.

REFERENCES

Bard, J, Rhee, SY et al. (2005) An ontology for cell types. Genome Biol. 6(2): R21.

Bard, J, and Rhee, SY (2004) Ontologies in biology: design, applications and future challenges. Nature Rev Gen 5, 213-222.

Brazma, A, Hingamp, P et al. (2001) Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nature Genetics 29, 365-371.

Brinkley, JF, Suciu, D et al. (2006) A framework for using reference ontologies as a foundation for the semantic web. In Proceedings, American Medical Informatics Association Fall Symposium, 96-100, Bethesda, MD.

Dyck, P and Chisholm, R (2003) Disease Ontology: Structuring Medical Billing Codes for Medical Record Mining and Disease Gene Association. Proceedings of the Sixth Annual Bio-ontologies Meeting, Brisbane, 2003, 53-55.

Fragoso, G, de Coronado, S, et al (2004) Overview and Utilization of the NCI Thesaurus. Comp Func Gen 5:8:648-654.

Grenon, P., Smith, B. et al. (2004) Biodynamic ontology: applying BFO in the biomedical domain. In Ontologies in Medicine (ed. Pisanelli, D.M.) 20–38 (IOS, Amsterdam, 2004).

Levenshtein, VI (1966) Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10:707–710.

Parkinson H, Kapushesky M, et al. (2006). ArrayExpress - a public database of microarray experiments and gene expression profiles. Nucl Acids Res 35, D747-750.

Phillips L (1990) Hanging on the Metaphone. Comp Lan 7: 39-49.

Phillips L (2000) The Double Metaphone Search Algorithm. C/C++ Users Journal.

Schober, D, Kusnierczyk, W et al. (2007) Towards naming conventions for use in controlled vocabulary and ontology engineering, Proceedings of the Bio-Ontologies Workshop, ISMB/ECCB, Vienna, July 20, 2007, 87-90.

Smith B, Ashburner, M, et al. (2007). The OBO Foundry: coodinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 25, 1251.

Sprague, J, Bayraktaroglu L, (2006) The Zebrafish Information Network: the zebrafish model organism database. Nucleic Acids Res. 34, D581–D585.

Uschold, M, Grüninger, M (1996) Ontology: Principles, Methods and Applications. Knowledge Engineering Review 11(2), 93-155.

TGF-beta Signaling Proteins and the Protein Ontology

*Cecilia Arighi[#1], Hongfang Liu[#1], Darren Natale[1], Winona Barker[1], Harold Drabkin[2], Zhangzhi Hu[1], Judith Blake[2], Barry Smith[3] and Cathy Wu[1]**

[1]Georgetown University Medical Center, Washington DC, USA; [2]The Jackson Laboratory, Bar Harbor, USA; [3] State University of New York at Buffalo, Park Hall, Buffalo,USA.

ABSTRACT

Motivation: The Protein Ontology (PRO) addresses the need for a formal description of proteins and their evolutionary relationships. PRO is authored via manual curation on the basis of content derived automatically from various data sources. Curation is needed to ensure correct representations of relationships both internally (between PRO nodes) and externally (to other ontologies). Focusing specifically on the TGF-beta signaling proteins, we describe how this ontology can be used for multiple purposes, including annotation, representation of objects in pathways, data integration, and the representation of biological system dynamics and of disease etiology.

1 INTRODUCTION

The Open Biomedical Ontologies (OBO) Foundry is a collaborative effort to establish a set of principles for ontology development with the goal of creating a suite of orthogonal interoperable reference ontologies in the biomedical domain [Smith et al., 2007]. The Foundry ontologies are organized along two dimensions: (1) granularity (from molecule to population); and (2) relations to time (objects, qualities, processes). In terms of this scheme, PRO is a representation of entities on the level of granularity of single molecules. It treats the molecules themselves, and interoperates with other ontologies, like the Sequence Ontology (SO) and the Gene Ontology (GO), for protein qualities and processes. PRO encompasses (i) a sub-ontology of proteins based on evolutionary relatedness (ProEvo), and (ii) a sub-ontology of the multiple protein forms produced from a given gene locus (ProForm) [Natale et al., 2007]. Here we summarize the current PRO framework focusing on the representation of proteins from the TGF-beta signaling pathway since they provide a rich body of protein annotation relating to a wide spectrum of protein forms (derived from cleavage and/or post-translational modifications (PTMs), alternative splicing, and sequence variants that are related to disease).

* To whom correspondence should be addressed.

\# Equal contribution to this work

2 THE PRO FRAMEWORK

Fig.1A shows the current working model and a subset of the possible connections to other ontologies. The root in the ontology is the class *protein*, which is defined as a biological macromolecule that is composed of amino acids linked in a linear sequence (a polypeptide chain), and is genetically encoded. PRO terms are connected by the relationship *is_a* or *derives_from*, both defined in the OBO Relations Ontology [Smith et al., 2005].

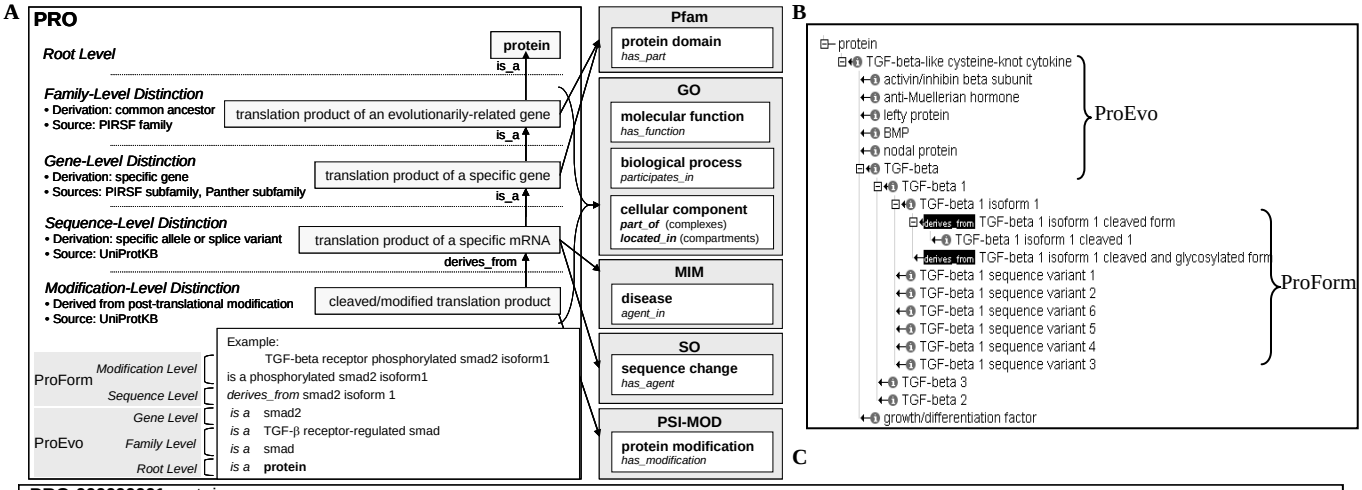
ProEvo: Proteins with similar domain architecture (that is, the same combination of domains in the same order) and full-length sequence similarity are said to be homeomorphic; they share a common ancestor and, usually, a specific biological function. Also, within any given homeomorphic group, there may be monophyletic subgroups of proteins that have distinct functions [Wu et al., 2004a] [Mi et al., 2006]. ProEvo defines protein classes based on these concepts, and captures the relationship between these classes. An illustrative example in PRO is depicted in **Fig.1B-C.** The PRO term PRO:000000008 *TGF-beta-like cysteine-knot cytokine* is defined as a protein with a signal peptide, a variable propeptide region and a cysteine-knot domain (definition in **Fig.1C**). The class represented by this term has seven children (**Fig.1B**), each of which can be defined as a separate group on the basis of a distinctive functional feature. PRO represents the proteins and not the individual domains. Thus, domain information is included in the ontology as part of the annotation of ProEvo terms with a link to the Pfam domain database [Finn et al., 2006] to indicate that a given protein class *has_part* some domain (see Pfam annotation in PRO:000000008 in **Fig.1C**). Therefore, tracing the relation between different ProEvo nodes would involve reasoning over the presence of a given domain. The gene product class, which is the leaf node of ProEvo, defines all protein products of strictly orthologous genes.

ProForm: This part of the ontology (**Fig1A-B**) describes the subset of the translational products that is experimentally characterized, and includes definition of sequence forms arising from allelic, splice, and translational variation, and from PTM and cleavage. Moreover, it allows representation of proteins that are products of a gene fusion due to chromosomal translocation, such as

PRO:000000091 *creb-binding protein/zinc finger protein HRX* that is encoded by part of the CREBBP gene at the N-terminus and part of the MYST4 gene at the C-terminus. This form is observed in some cases of acute myelogenous leukemia. In ProForm, equivalent protein forms in different species are represented as a single node. We use the *derives_from* relationship to describe the relation between a modified form and the parent protein. In addition, each form is connected to other ontologies, which provide the annotation (**Fig.1A, C**).

3 BUILDING THE ONTOLOGY

For the current release (1.0) we focus on the set of proteins in the TGF-beta signaling pathway from the KEGG pathway database [Kanehisa et al., 2002], which includes the TGF-beta, the bone morphogenetic protein and activin-mediated signaling pathways. The ontology consists of a total of 667 PRO terms, including 111 ProEvo and 544 ProForm terms. It covers 79 human/mouse orthologous proteins that mapped to 34 PIRSF homeomorphic



C

PRO:000000001 protein
is_a **PRO:000000008** TGF-beta-like cysteine-knot cytokine
def: A protein with a core domain composition consisting of a signal peptide, a variable propeptide region and a transforming growth factor beta like domain, that is a cysteine-knot domain containing four conserved beta strands, S1-S4, which form two antiparallel beta sheets (S1-S2 and S3-S4) interconnected by three disulfide bridges in a knot-like topology.
Pfam:PF00019 "has_part Transforming growth factor beta like domain".
is_a **PRO:000000046** TGF-beta { PIRSF001787 }
def: A TGFB-like cysteine-knot cytokine whose propeptide region is a latent associated peptide (LAP) that remains associated to the mature TGF-beta after cleavage and secretion, rendering TGF-beta inactive. This is the founding member of the cysteine-knot cytokine family, and is related to the activin/inhibin, anti-Muellerian hormone, and bone morphogenic protein families, which are all involved in the regulation of cell growth and differentiation.
is_a **PRO:000000182** TGF-beta 1 (TGFB1 gene products) { PANTHER:PTHR11848:SF32 }
def: A TGF-beta that is a translation product of the TGFB1 gene.
is_a **PRO:000000397** TGF-beta 1 isoform 1 (precursor) { UniProtKB: P01137, P04202 }
def: A TGF-beta 1 that is a translation product of a processed transcript of the TGFB1 gene, and includes all core domains (signal peptide, latent associated peptide and a transforming growth factor beta like domain), as in the human sequence UniProtKB:P01137. This form is a precursor.
GO:0005796 "located_in Golgi lumen [PMID:1544940, TaxID:9606]".
derives_from **PRO:000000512** TGF-beta 1 isoform 1 cleaved and glycosylated
MOD:00693 "has_modification glycosylated residue".
is_a **PRO:000000617** TGF-beta 1 isoform 1 cleaved and glycosylated 1 (TGF-beta 1 latent peptide)
is_a **PRO:000000618** TGF-beta 1 isoform 1 cleaved and glycosylated 2 (Latent complex) { P01137 (30-279, N-glycosylated Asn-82):(279-390) }
GO:0005160 "NOT has_function transforming growth factor beta receptor binding [PMID:3162913, TaxID:9606]".
GO:0005578 "located_in proteinaceous extracellular matrix [PMID:9008713, TaxID:10090]".
derives_from **PRO:000000513** TGF-beta 1 isoform 1 cleaved form
is_a **PRO:000000616** TGF-beta 1 isoform 1 cleaved 1 (TGF-beta 1 active peptide) { P01137-1 (279-390) }
def: A TGF-beta 1 isoform 1 cleaved form that has been processed by proteolytic cleavage and released from the latent complex. This form is the active mature peptide.
GO:0005160 "has_function transforming growth factor beta receptor binding [PMID:14764882, TaxID:10090]".
GO:0005578 "located_in proteinaceous extracellular matrix [PMID:9008713, TaxID:10090]".
is_a **PRO:000000401** TGF-beta 1 sequence variant 4 (precursor) { His-218 in human sequence, VAR_017607 }
def: A TGF-beta 1 that is a translation product of the a polymorphic sequence variant of TGFB1 gene that has a His residue at the position equivalent to Arg-218 in the human sequence UniProtKB:P01137.
MIM:131300 "agent_in CAMURATI-ENGELMANN DISEASE [PMID:12493741, TaxID:9606]".
SO:1000093 "has_agent mutation_causing_amino_acid_substitution".
derives_from **PRO:000000516** TGF-beta 1 sequence variant 4 cleaved form
is_a **PRO:000000627** TGF-beta 1 sequence variant 4 cleaved 1 (TGF-beta 1 active peptide) { VAR_017607 (279-390) }
GO:0008083 "has_function (**increased**) growth factor activity [PMID:12843182, TaxID:9606]".
is_a **PRO:000000628** TGF-beta 1 sequence variant 4 cleaved 1 (TGF-beta 1 latent peptide) { VAR_017607 (30-279) }
SO:1000093 "has_agent mutation_causing_amino_acid_substitution".
SO:1000124 "has_agent mutation_causing_partial_loss_of_function_of_polypeptide [PMID:11278244, TaxID:9606]".

Fig.1- PRO framework and DAG view of the ontology. A) Current working model and a subset of the possible connections to other ontologies. B) Snapshot of the ontology (partial view) in OBO Edit 1.1 including terms representing ProEvo and ProForm. C) A PRO example illustrated by the TGF-beta 1 protein. The above is a partial view, not all forms are listed, and only key annotations are shown.

families and 36 Pfam domains. An automated process has been developed to generate PRO nodes from PIRSF [Wu et al., 2004a] and iProClass [Wu et al., 2004b] databases, UniProtKB [UniProt Consortium, 2008], as well as MGI, Pfam, and PANTHER [Mi et al, 2005]. The computationally-generated file is in OBO format and we use OBO Edit 1.1 [Day-Richter et al., 2007] as the curation platform. Manual curation includes (i) merging of nodes, for example whenever PIRSF and PANTHER families represent the same homeomorphic protein class (same membership), (ii) reviewing the literature and sequence analysis to verify or create the protein forms, e.g., analyzing what combination of modifications occurs in a specific form, and determining what forms are equivalent in mouse and human. Furthermore, new ProForm nodes can be created for newly characterized isoforms or sequence variants not yet represented in UniProtKB (e.g., PRO:000000478 *smad5 isoform 2*, and PRO:000000483 *smad9 isoform 2*). Names of ProEvo nodes are adapted from the underlying data sources or from the literature. The names of ProForm nodes are based on their parent node (**Fig.1B**, under the ProForm bracket). All PRO terms have a definition that conforms to OBO foundry standards with examples delineated in **Fig.1C**. A reference to conserved motifs and domain regions is used whenever possible and, in some cases, examples are supplied. Each PRO definition has source attribution to PubMed ID, PRO curator, or other resource ID. In addition, the annotations are introduced via cross-references to other ontologies. Currently, the majority of model organism databases supply GO annotation to a gene object rather than to a specific protein form. PRO assigns the GO terms to the specific forms (**Fig.1C**). In the example above, although TGF-beta 1 is annotated in databases with the GO:0005160 *transforming growth factor beta receptor binding*, this term is not appropriate to annotate the precursor (PRO:000000397), but rather the active peptide (PRO:000000616). So the advantage of the PRO framework is that it can provide a basis for more accurate annotation. To further illustrate the importance of this statement, **Fig.1C** shows some of the nodes and relationships for the TGF-beta 1 protein, thereby demonstrating the complexity and variety of sequence forms that can be derived from a given parent sequence. TGF-beta 1 precursor is a dimer and undergoes two cleavages–by a signal peptidase and by furin in the Golgi–to generate two functionally important chains: the TGF-beta 1 mature and the latent peptide (PRO:000000617). These two chains remain associated (as a latent complex) until proteases in the extracellular space degrade the latent peptide. The latent complex is represented by PRO:000000618, whereas the active mature protein is represented by PRO:000000616. Note that only the latter is associated with the GO term corresponding to receptor binding activity, and that the TGF-beta 1 isoform 1 precursor and any of its derived forms differ in cellular localization. In addition, an arginine to histidine variant (R218H) in the human protein is responsible for the Camurati-Engelmann disease (*agent_in* the disease). This mutation affects the stability and conformation of the latent peptide, elevating the levels of free (active) mature peptide. This situation is formally represented in PRO by associating the corresponding SO terms to the corresponding products (see PRO:000000401 and its children nodes) and, also, by adding a modifier to the *has_function* relationship to reflect the constitutively active mature peptide (PRO:000000627). Also note that there is no term corresponding to the latent complex derived form for this variant. A total of 1667 annotations have been added to PRO nodes in release 1.0. **Table 1** shows the statistics for GO terms. The examples illustrate how appropriate annotation can be assigned to appropriate protein forms.

Table 1: Statistics on GO terms in PRO release 1.0

GO term	OBO Relation	# terms	Example	
Molecular Function	has_function	181	PRO:000000650 smad 5 isoform 1 phosphorylated 1	GO:0046332 SMAD binding
	NOT has_function	43	PRO:000000478 smad 5 isoform 2	GO:0046332 SMAD binding
Cellular Component Complex	part_of	38	PRO:000000178 RING-box protein 2 isoform 1	GO:0000151 ubiquitin ligase complex
	NOT part_of	5	PRO:000000179 RING-box protein 2 isoform 2	GO:0000151 ubiquitin ligase complex
Cellular Component	located_in	171	PRO:000000457 noggin isoform 1 cleaved 1	GO:0005615 extracellular space
Biological Process	participates_in	235	PRO:000000086 chordin isoform 1	GO:0001501 skeletal development

4 DISSEMINATION OF PRO

The results of the PRO project are disseminated through several mechanisms: the entire ontology and associated wiki are both accessible through the PRO public website (http://pir.georgetown.edu/pro/), as well as through the OBO Foundry (http://www.obofoundry.org/), and the National Center for Biomedical Ontology (NCBO) BioPortal (http://www.bioontology.org/bioportal.html).

5 PRO AND ITS USER COMMUNITY

Any project that needs to specify protein objects of the type described in the ontology can benefit from PRO. The TGF-beta signaling pathway described above shows how the protein ontology can assist in the explicit annotation of states of a molecule. These states are natural components of pathway ontologies or databases such as INOH Event Ontology [Kushida et al., 2006] or Reactome [Vastrik et al., 2007] (the latter does contain the relevant entities, but as accessions only; they are not as yet formed

into an ontology structure which supports reasoning). As biomedical data expand, it will be increasingly important to explicitly represent these protein forms so that representations of attributes can be attached to the appropriate entities. **Fig.2** shows part of the TGF-signaling pathway as described by Reactome with PRO terms mapped to the associated Reactome events. This improves the mapping of the entities involved in the pathway, and gives a more accurate and complete framework for researchers to analyze their data. In addition, PRO allows modeling of the specific objects involved in a given disease, as the *TGF-beta 1 sequence variant 4* case described above (**Fig.1C**). Other examples include the representation of (1) a cleaved form of rho-associated protein kinase 1 (PRO:000000563), which is constitutively active in the mouse myopathy model and in human heart failure patients, and of (2) smad4 and BMP receptor type-1A sequence variants associated with a common disease allowing the inference of a specific pathway failure. PRO terms could also be adopted by GO to accurately define protein complexes in the cellular component ontology. PRO could potentially be used for cross-species comparison of protein forms, since only the forms with experimental evidence are included (with the associated literature and taxon IDs). Finally, PRO could be adopted where data integration at the molecular level of proteins is needed, as in systems biology or in translational medicine.

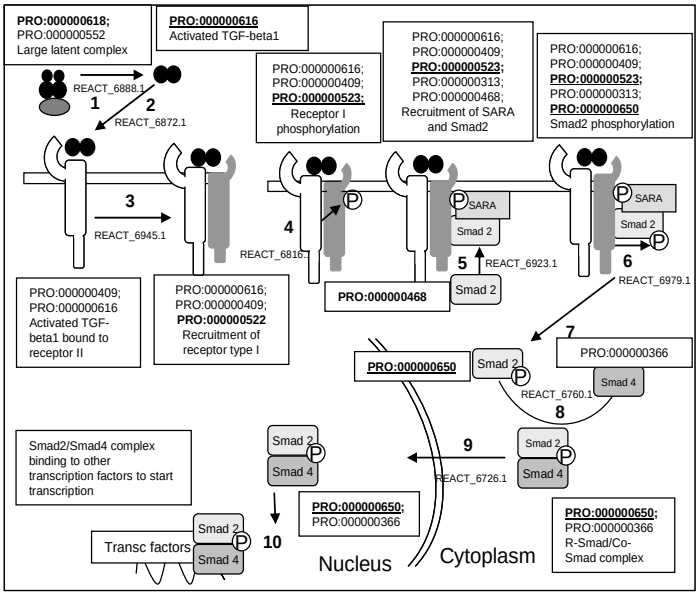


Fig.2- PRO and the Reactome TGF-beta signaling pathway (REACT_6844). Each step in the pathway is described by a Reactome event ID. Bold PRO IDs indicate objects that undergo some modification that is relevant for function (the modified form is underlined).

6 CONCLUSION

We illustrated key aspects of the PRO framework through reference to proteins involved in the TGF-beta signaling pathway. The significance of the framework can be summarized as follows: (1) it provides a structure to support formal, computer-based inferences based on data pertaining to shared attributes among homologous proteins; (2) it helps us to delineate the multiple protein forms of a gene locus; (3) it provides important interconnections between existing OBO Foundry ontologies; (4) it provides a framework that can be adopted by other ontologies and/or databases, as for example, to better define objects in pathways, or complexes or in disease modeling; (5) it allows the community to annotate their proteins of interest. Finally, it offers a comprehensive picture of the protein realm by connecting protein evolution, function, modification, variants, gene ontology and disease.

ACKNOWLEDGEMENTS

Funding for PRO development is provided by NIH grant 1 R01 GM080646-01.

REFERENCES

Day-Richter J., Harris M.A., Haendel M.; Gene Ontology OBO-Edit Working Group, Lewis S. (2007) OBO-Edit--an ontology editor for biologists.*Bioinformatics*, **23**:2198-2200.

Finn R.D., Mistry J., Schuster-Bockler B., Griffiths-Jones S., Hollich V., et al. (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**:D247-251.

Vastrik I., D'Eustachio P., Schmidt E., Joshi-Tope G., Gopinath G., et al. (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biology*, **8**:R39.

Kanehisa M, Goto S, Kawashima S, Nakaya A. (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res*. **30**:42-46.

Kushida T., Takagi T., Fukuda K.I. (2006) Event Ontology: A pathway-centric ontology for biological processes. *Pacific Symposium on Biocomputing 2006*, **11**:152-163.

Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, et al. (2005) The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res., **33**:D284-288.

Natale D., Arighi C., Barker W.C., Blake J., Chang T., et al. (2007) Framework for a Protein Ontology. *BMC Bioinformatics*, **8**(Suppl 9):S1.

Smith B., Ceuster W., Klagger B., Kohler J., Kumar A., et. al. (2005) Relations in Biomedical Ontologies. *Genome Biol*., **6**:R46.

Smith B., Ashburner M., Rosse C., Bard J., Bug W., et al. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*., **25**: 1251-1255.

UniProt Consortium (2008) The Universal Protein Resource (UniProt). *Nucleic Acids Res*., 36(Database issue):D190-195.

Wu C.H., Nikolskaya A., Huang H., Yeh L-S., Natale D.A., et al. (2004a) PIRSF family classification system at the Protein Information Resource. *Nucleic Acids Res.*, **32**:D112-114.

Wu C.H., Huang H., Nikolskaya A., Hu Z., Yeh L.S., et al. (2004b) The iProClass Integrated database for protein functional analysis. *Computational Biol. and Chem,* **28**: 87-96.

Tamoxifen to Systemic Lupus Erythematosus: Constructing a Semantic Infrastructure to Enable Mechanism-based Reasoning and Inference from Drugs to Diseases

Angela Xiaoyan Qu[1,3,4], Ranga C. Gudivada[1,3], Eric K. Neumann[5] , Bruce J Aronow[1,2,3*]*

University Departments of Biomedical Engineering[1] and Pediatrics[2], University of Cincinnati and Division of Biomedical Informatics[3], Cincinnati Childrens Hospital Medical Center; Procter & Gamble Pharmaceutical[4], Cincinnati OH; Clinical Semantics Group[5], Lexington, MA

ABSTRACT

Discovering new clinical applications for existing drug products and predicting novel drug combinations for improved efficacy represent promising opportunities for both pharmaceutical development and personalized medicine. To enable these efforts, we have sought to develop a systematic framework for representation and inference of drug/disease relationships based on mechanistic knowledge. To do this we have developed a *Disease-Drug Correlation Ontology (DDCO)* that provides a framework for asserting entity and relationship type characteristics of heterogeneous data from pharmacological, medical, and genetic, and other biological domains. The DDCO, formalized in OWL, allows for the representation of multiple ontologies, controlled vocabularies, and data schemas and normalized mappings of relationships between elements of each source. In the present study we used the DDCO framework to form relationships across a collection of data sources including DrugBank, EntrezGene, OMIM, Gene Ontology, SNOMED, MeSH Anatomy, and other sources in UMLS, to construct an extensible Pharmacome-Genome-Diseasome network. As an example, we illustrate the utility of this approach to simultaneously model biological processes associated with disease processes, phenotypic attributes, and mechanisms of drug action to predict a new indication for Tamoxifen could be to treat the therapeutically challenging disease entity *Systemic Lupus Erythematosus*.

1. INTRODUCTION

Drug repositioning, i.e. the use of established drugs to treat diseases that are not established as indications for its use, represents a promising avenue based on its lower development cost and availability of extensive data and knowledge from prior research[1]. Despite impressive successes shown by repositioned drugs, most of these are the result of "serendipity" –ie unexpected findings made during or after late phases of clinical study. Thus, a forecasting model that could improve data capture, integration, analysis, and prediction of potential new therapeutic indications for drugs based on integrated biomedical knowledge around drug and disease mechanisms is highly desirable. Currently, most drug-oriented databases e.g. PharmGKB[2], KEGG[3] and DrugBank[4] tend to support limited dimensionality of mechanism-associated relationships and lack the multi-disease gene and phenotype relationships that are likely to be necessary to infer between disparate diseases.

The advancements of Semantic Web (SW)[5] and related knowledge representation technologies provide a promising platform for semantic integration of heterogeneous data and knowledge interoperability. Hypothesizing that associating comprehensive biomedical information and prior knowledge around pharmacological entities (i.e. biological, chemical, and clinical processes) and using SW principles and technologies can facilitate reveal new knowledge such as novel indications for known or unknown drugs, we devised a knowledge framework, Disease-Drug Correlation Ontology (DDCO), using Web Ontology Language (OWL) representation formalism to facilitate mapping and assertion of these relationships across multiple ontologies and hierarchically organized data sources.

The working ontology, DDCO, is thus an aggregation of manual curation and integration of relevant components from multiple existing ontologies, vocabularies, and database schemas. We used the DDCO to link DrugBank, OMIM, EntrezGene, KEGG, BioCarta, Reactome, and UMLS; and the data was semantically integrated into a Resource Description Framework (RDF) network. Using this, we present as an example scenario the implication of Tamoxifen, an established drug product, as potential therapeutic for systemic lupus erythematosus (SLE). We propose that further populating knowledge bases with similar structure will enable both new indications and the identification of synergistic drug combinations.

2. DRUG-DISEASE CORRLEATION ONTOLOGY CONSTRUCTION

2.1 Ontology Development

Our goal is to devise a drug- and disease-centric knowledge framework that serves both data integration and knowledge exploitation needs. The ontology was designed with high-level of granularity and aims to reuse knowledge components whenever possible. Therefore, the first step for our ontology development effort was to examine and select from previously existing resources that allow efficient knowledge mapping and sharing among independent data sources. We used UMLS Semantic Network[7] to construct the scaffold of the DDCO. Though containing a set of broad semantic types and relationships defining biomedical concepts, UMLS Semantic Type has knowledge "gaps" and is insufficiently organized. For example, for pharmacological domain, it only contains one Semantic Type, "*Pharmacologic_Substance*" with only one child term,

* Corresponding authors: qu.ax@pg.com, bruce.aronow@cchmc.org.

"*Antibiotic*", which is far from the full representation of drug-centric entities. To fill in such gaps, we proposed to also use below ontology or vocabulary sources (Fig.1):

- MeSH (medical subject headings): the controlled vocabulary thesaurus in biomedical fields.
- NCI Thesaurus[8]: an ontology-like vocabulary in cancer-centric disease areas
- The Anatomical Therapeutic Chemical Classification[9]: a WHO recommended classification system for internationally applicable methods for drug utilization research
- Common Terminology Criteria for Adverse Event (CTCAE) [10]: A descriptive terminology and grade scales for drug adverse event reporting
- Gene Ontology[11]
- SNOMED CT[12]: clinical health care terminology and infrastructure

Ontology editor *Protégé* was used as the primary tool for implementing the OWL framework[13]. To enhance the editing and visualization flexibility, several plug-ins were also used including PROMPT for ontology comparison and merging and OwlViz for visualization. Ontology mapping and aligning techniques were applied for concept and relationship integration. In addition, manual modifications, such as pruning irrelevant or duplicate branches or adding new concepts/relationships, were performed to maximize integration and minimize non-connectivity. In addition, concept restrictions and property constraints were also manually curated to support the inferential capability enabled by description logic. We used RACER[14], a description logic reasoning system with support for T-Box and A-box reasoning, to pose DL queries for the ontology evaluation. On average, the subsumption computations were completed within ten seconds and we sought to solve any inconsistencies to assure the integrity of the DDCO.

2.2 Ontology Model Metrics

While efforts to expand and refine the conceptualization are continuing, the current DDCO contains 2046 classes (excluding GO which was imported directly), with average sibling number of 17 (maximum 35 and minimum 1) per class. There are total of 221 properties, with 99 properties domain-specified, 69 range-specified, and 36 inverse-specified. These properties include 135 selected UMLS Semantic Network relations, 40 SNOMED attributes, and 46 custom-defined properties for constraint development. To further refine the entities, we created 67 restrictions: 7 existential, 36 universal, and 25 cardinality. Fig.2 presents a top-level view of the ontology concepts as well as properties connecting them. Fig.3 shows the semantic model for the "Drug" entity including our curation of concept restrictions including necessary and sufficient restrictions. For example, one of the criteria to define a drug is it needs to have at least one "active" ingredient (Fig.3).

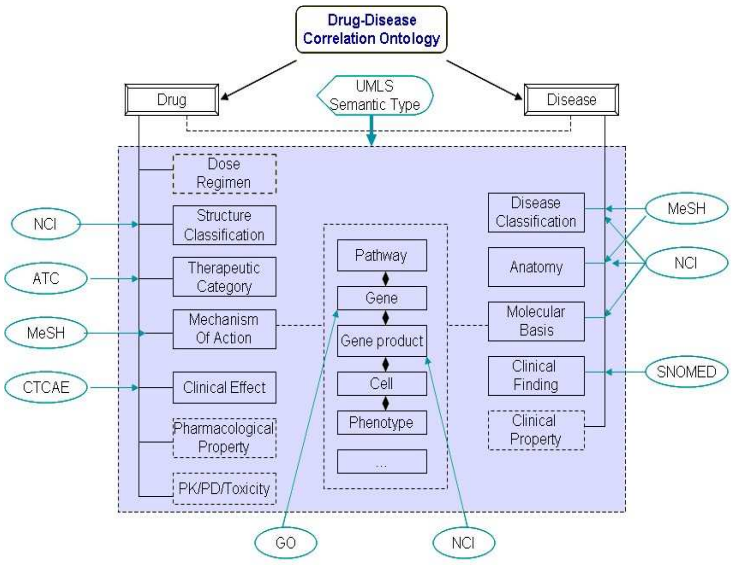


Fig 1. Schematic view of Drug-Disease Correlation Ontology model and the major resources (in oval) included

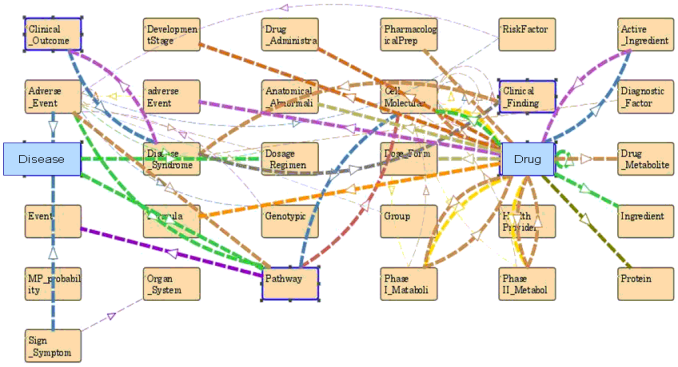


Fig 2. Top-level view of conceptual frames and domain/range relationships in DDCO.

3. APPLICATIONS

3.1. An Integrated Pharmacome-Genome-Diseasome RDF Network

A key benefit of the Semantic Web is its ability to integrate relevant data from different origins and in incompatible formats. We used the DDCO as the knowledge framework to integrate a diverse collection of data sources across pharmacome- , genome-, and diseasome- domains. Specifically, following data sources were used for extracting and integrating relevant data:

- *Pharmacome Data Source:*

Drug-associated information was compiled from DrugBank. The resultant data set contains 4,763 drug entries, including over 1,400 FDA-approved drugs.

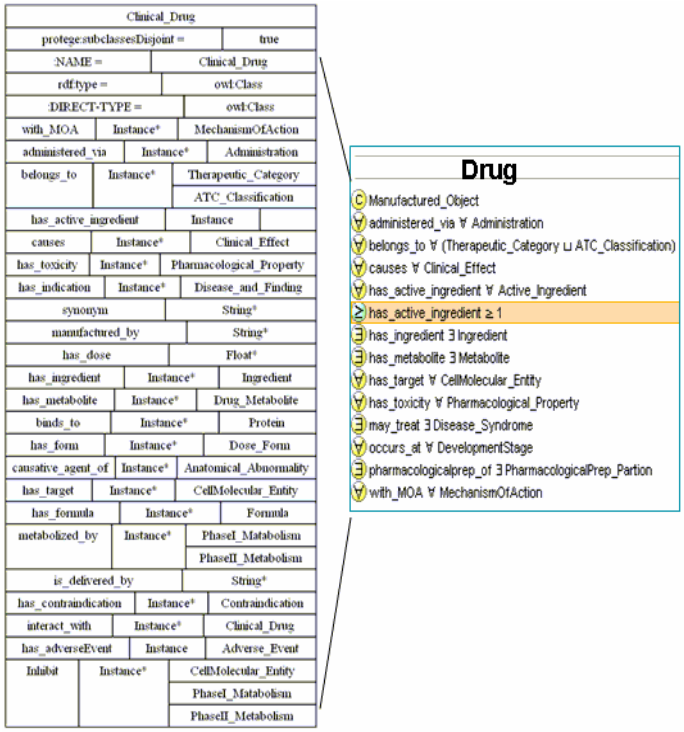


Fig 3. Partial view of properties and restrictions for Drug entity modeling in DDCO

- *Genome Data Source:*

The annotation of human genes and interactome data including BIND, BioGRID and HPRD data were downloaded was from NCBI ftp site. Gene-pathway annotations were compiled from KEGG, BioCarta, BioCyc and Reactome databases. The total data set contains 15068 human genes annotated with 7124 unique GO terms, and 14899 gene-pathway associations.

- *Diseasome Data Source:*

OMIM[15] records were downloaded in XML format. OMIM ID and the corresponding gene associations were downloaded from NCBI EntrezGene ftp site.

To explore the implicit associations between drug and disease, we need to understand the "explicit" relationships between them too. Thus, we have extracted the known drug-disease associations (i.e. indication for FDA-approved drugs) using UMLS 2007AC files from UMLS Knowledge Server. We used the table MRCONSO.RRF to map the FDA-approved drugs to the UMLS unique concept identifier (CUI). Next, the table MRREL.RRF was used to extract the associated indications for these CUI concepts. The semantic relationships of "*may_treat*" and "*may_be_treated_by*" were used to restrict the relationship mapping. To further refine the extraction and eliminate false positive mapping, the semantic type "Chemicals & Drugs" and "Disorders" were used to constrain the association concepts. As a result, a total of 230,114 drug-disease associations were extracted.

As the next step to build the integrated RDF data graph, we created models based on the logic and semantic relationships defined in the DDCO for each of our pharmacome, genome, and diseasome domains. These models provide the required mapping mechanism from the instance data to DDCO in order to semantically annotate and relate different –omic entities. The data parsed and extracted from above sources was converted from various formats (i.e. RDF/OWL, XML, txt) into RDF triples using different RDF converters in compliance with the definitions by the DDCO model. The converted RDF triples were further converted into N-Triple format using Oracle RDF loaders before loading to the Oracle 10g release 2 RDF store[16]. With the assigned unique name space and the shared identifiers, the data loaded in the RDF model are thereafter integrated automatically in a seamless manner.

3.2 Exploiting Drug-Disease Association: A Scenario from *Tamoxifen to SLE*

To find novel applications for established therapeutics, we chose to investigate if evidence could be accrued to indicate if Tamoxifen, a selective estrogen receptor modulator approved for breast cancer, might have additional uses. Some beneficial effects of Tamoxifen on SLE (a chronic autoimmune disease that may affect multiple organ systems) have been observed in animal tests[17] as well as some preliminary clinical studies, supporting the hypothesis that selective estrogen receptor modulators such as Tamoxifen may have therapeutic potential in SLE patient management[18, 19].

While the pathogenesis of SLE is complex and poorly understood, we sought to identify connectivities offered via representation of gene networks associated with perturbations of its implicated cell, pathway, ontology and phenotype correlates with those of Tamoxifen. First, we issued Oracle RDF queries to retrieve Tamoxifen and SLE RDF subgraph respectively:

- For Tamoxifen, we developed RDF queries to ask the complex question "*retrieve all genes and their annotation (interacting gene, pathway, and gene ontology) that associated with Tamoxifen by acting as its drug target(s) or indication(s)"*
- Similarly, for SLE, we developed RDF queries to "*retrieve disease genes, or genes interacting with or sharing pathways with SLE disease gene as well as their annotation*"

Each query returns a set of variable bindings matching to the query parameters and each unique result produces a graph formed from the triples matching the criteria. The components of the resultant RDF subgraph were summarized in Table 1. As expected, since the connection between "Tamoxifen" and "SLE" is non-trivial, no association was detected in each individual RDF subgraph. However, by combining the extracted subgraphs and applying inference rules using GO and Disease subsumption relationships, we were able to extract the implicit connections between the two entities of interest. For example, one of the shortest associations extracted is via a common biological process "apoptosis" (*GO_0006915*) that are both traversed by PDCD1 and CDH1, two genes that are found to be associated with known

Tamoxifen indication and SLE. Fig.4 shows all the embedded associations extracted from the combined SLE-Tamoxifen RDF graph, consisting 45 entity nodes with the minimal geodesics of 6 traversing between Tamoxifen and SLE. In addition, we also borrowed the centrality analysis algorithm and approach[20] to compute the key biological entities for the extracted RDF graphs. The RDF triples were used as input for generating nodes and edges. As a result, two critical genes were identified with high ranking scores: ESR1 (*estrogen receptor 1*), AR (*androgen receptor*). Based on literature mining, both genes are found to be differentially expressed in SLE patients with an indicated role in SLE pathogenesis or patient management[18, 21, 22].

Table 1: Statistics of RDF Graph associated with Tamoxifen, SLE, and Combined

RDF Graph	SLE	Tamoxifen	Combined
Entities	114	695	768
Associations	121	947	1050

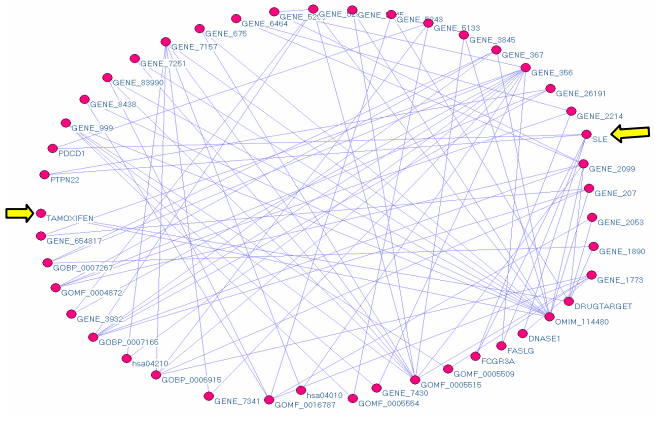


Fig 4**:** Implicit associations between "Tamoxifen" and "SLE" (entities pointed by yellow arrows) consisting of 45 vertices, with minimal geodesics of 6

2. CONCLUSION AND FUTURE WORK

We have presented a novel OWL-formalized ontology framework for use in biomedical and pharmacological domain applications. Our work to implement an integrated pharmacome-genome-diseasome RDF network based on this framework suggests that the DDCO is effective and robust in knowledge acquisition, integration, and inconsistency resolution. The application scenarios we presented in this paper illustrates that the DDCO framework and its supported RDF graph data model, in combination with graph traversal and mining methods, can be used in an exploratory context to formulate either initiating or validating hypotheses. The scenarios can also be generalized to other research questions in drug development area (see our prior work[23, 24]) to support identifying new target or therapeutics. Our current and planned work seeks to deepen knowledge capture and mechanism modeling to further refine the reasoning capability of the OWL/RDF model and include additional dimensions such as genetic polymorphisms, mutations, deeper clinical features, and diverse pharmacological properties and principles of drug action. Doing so should greatly extend sensitivity and specificity for individual patients. We also plan to continuously evaluate and improve the framework in conjunction with future expansion of the semantic infrastructure by enabling expert review for a specific disease to model its mechanisms and variations using entities and relations from the DDCO.

REFERENCES

1. Ashburn TT, Thor KB. Drug repositioning: Identifying and developing new uses for existing drugs. *Nat Rev Drug Discov*. 2004;3:673-683.
2. Hewett M, Oliver DE, Rubin DL, et al. PharmGKB: The pharmacogenetics knowledge base. *Nucleic Acids Res*. 2002;30:163-165.
3. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000;28:27-30.
4. Wishart DS, Guo AC, et al. DrugBank: A comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res*. 2006;34:D668-72.
5. J. H. Tim Berners-Lee The semantic web. *Sci Ameri Mag*.;(284): p. 29-37.
6. Bodenreider O. The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Res*. 2004;32:D267-70.
7. Kashyap V. The UMLS semantic network and the semantic web. *AMIA Annu Symp Proc*. 2003:351-355.
8. de Coronado S, Haber MW, Sioutos N, Tuttle MS, Wright LW. NCI thesaurus: Using science-based terminology to integrate cancer research results. *Medinfo*. 2004;11:33-37.
9. ATC::http://www.whocc.no/atcddd/.
10. CACTE: http://ctep.cancer.gov/reporting/ctc.html.
11. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: Tool for the unification of biology. the gene ontology consortium. *Nat Genet*. 2000;25:25-29.
12. Stearns MQ, Wang AY.et al SNOMED clinical terms: Overview of the development process and project status. *Proc AMIA Symp*. 2001:662-666.
13. Noy NF, Crubezy M, Fergerson RW, et al. Protege-2000: An open-source ontology-development and knowledge-acquisition environment. *AMIA Annu Symp Proc*. 2003:953.
14. Haarslev V, Möller R. Racer: An OWL reasoning agent for the semantic web. *Proceedings of the International Workshop on Applications, Products, and Sevices of Web-based Support Systems.* 2003;18:91.
15. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*. 2005;33:D514-7.
16. ORACLE: http://www.oracle.com/technology/tech/semantic_technologies.
17. Sthoeger ZM, Zinger H, Mozes E. Beneficial effects of the anti-oestrogen tamoxifen on systemic lupus erythematosus of (NZBxNZW)F1 female mice are associated with specific reduction of IgG3 autoantibodies. *Ann Rheum Dis*. 2003;62:341-346.
18. Cohen-Solal Diamond B. et al Sex hormones and SLE: Influencing the fate of autoreactive B cells. *Curr Top Microbiol Immunol*. 2006;305:67-88.
19. Peeva E, Venkatesh J, Michael D, Diamond B. Prolactin as a modulator of B cell function: Implications for SLE. *Biomed Pharmacother*. 2004;58:310-319.
20. Koschützki D, Richter S, et al Tenfelde-Podehl D, Zlotowski O, ed. *Centrality Indices. In Network Analysis: MethodologicalFoundations.* Springer; 2005; No. Volume 3418 of LNCS Tutorial.
21. Strand V. Approaches to the management of systemic lupus erythematosus. *Curr Opin Rheumatol*. 1997;9:410-420.
22. Inui A, Ogasawara H, Naito T, et al. Estrogen receptor expression by peripheral blood mononuclear cells of patients with systemic lupus erythematosus. *Clin Rheumatol*. 2007;26:1675-1678.
23. Gudivada R, Qu AX, Jegga AG, Neumann EK, Aronow BJ. A genome – phenome integrated approach for mining disease genes using semantic web. *WWW2007Workshop on Healthcare and Life Sciences. 2007.*
24. Qu AX, Aronow B. et al Semantic web-based data representation and reasoning applied to disease mechanism and pharmacology. *Proceedings of 2007 IEEE Bioinformatics and Biomedicine.* 2007:131-143.

BIOINFORMATICS

Vol. 00 no. 00 2008
Pages 1–4

BOWiki: An ontology-based wiki for annotation of data and integration of knowledge in biology

Robert Hoehndorf,[a,c,d] Joshua Bacher,[b,c] Michael Backhaus,[a,c,d] Sergio E. Gregorio, Jr.,[e] Frank Loebe,[a,d] Kay Prüfer,[c] Alexandr Uciteli,[c] Johann Visagie,[c] Heinrich Herre [a,d] and Janet Kelso [c]

[a]Department of Computer Science, Faculty of Mathematics and Computer Science, University of Leipzig, Johannisgasse 26, 04103 Leipzig, Germany; [b]Institute for Logics and Philosophy of Science, Faculty of Social Science and Philosophy, University of Leipzig, Beethovenstrasse 15, 04107 Leipzig, Germany; [c]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany; [d]Research Group Ontologies in Medicine (Onto-Med), Institute of Medical Informatics, Statistics and Epidemiology (IMISE), University of Leipzig, Härtelstrasse 16–18, 04107 Leipzig, Germany; [e]Communications and Publications Services, International Rice Research Institute, College, 4030 Los Baños, Laguna, Philippines

ABSTRACT

Ontology development and the annotation of biological data using ontologies are time-consuming exercises that currently requires input from expert curators. Open, collaborative platforms for biological data annotation enable the wider scientific community to become involved in developing and maintaining such resources. However, this openness raises concerns regarding the quality and correctness of the information added to these knowledge bases. The combination of a collaborative web-based platform with logic-based approaches and Semantic Web technology can be used to address some of these challenges and concerns.

We have developed the BOWiki, a web-based system that includes a biological core ontology. The core ontology provides background knowledge about biological types and relations. Against this background, an automated reasoner assesses the consistency of new information added to the knowledge base. The system provides a platform for research communities to collaboratively integrate information and annotate data.

The BOWiki and supplementary material is available at `http://www.bowiki.net/`. The source code is available under the GNU GPL from `http://onto.eva.mpg.de/trac/BoWiki`.

Contact: bowiki-users@lists.informatik.uni-leipzig.de

1 INTRODUCTION

Biological ontologies have been developed for a number of domains, including cell structure, organisms, biological sequences, biological processes, functions and relationships. These ontologies are increasingly being applied to describe biological knowledge. Annotating biological data with ontological categories provides an explicit description of specific features of the data, which enables users to integrate, query and reuse the data in ways previously not possible, thereby significantly increasing the data's value.

Developing and maintaining these ontologies requires manual creation, deletion and correction of concepts and their definitions within the ontology, as well as annotating biological data to concepts from the ontology. In order to overcome the arising acquisition bottleneck, several authors suggest using community-based tools such as wikis for the description, discussion and annotation of the functions of genes and gene products [Wang, 2006, Hoehndorf et al., 2006, Giles, 2007].

However, an open approach like wikis frequently raises concerns regarding the quality of the information captured. The information represented in the wiki should adhere to particular quality criteria such as internal consistency (the wiki content does not contain contradictory information) and consistency with biological background knowledge (the wiki content should be semantically correct). To address some of these concerns, logic-based tools can be employed.

We have developed the BOWiki, a wiki system that uses a core ontology together with an automated reasoner to maintain a consistent knowledge base. It is specifically targeted at small- to medium-sized communities.

2 SYSTEM DESCRIPTION

The BOWiki is a semantic wiki based on the MediaWiki[1] software. In addition to the text-centered collaborative environment common to all wikis, a semantic wiki provides the user with an interface for entering structured data [Krötzsch et al., 2007]. This structured data can be used subsequently to query the data collection. For instance, *inline queries* [Krötzsch et al., 2007] can be added to the source code of a wikipage, which will always produce an up-to-date list of results on a wikipage.

The BOWiki significantly extends the MediaWiki's capabilities. It allows users to characterize the entities specified by wikipages as *instances* of ontological categories, to *define new relations* within the wiki, to *interrelate* wikipages, and to *query* for wikipages satisfying some criteria. In particular, the BOWiki provides features beyond those offered by common wiki systems (for details see the Implementation section and table 1): typing wikipages (table 1), *n*-ary semantic relations among wikipages (table 1), semantic

[1] `http://www.mediawiki.org`

search (special page, inline queries), reasoner support for content verification, adaptability to an application domain, import of bio-ontologies for local accessibility and simple reuse, graphical ontology browsing and OWL [McGuinness and van Harmelen, 2004] export of the wiki content.

We consider both adaptability to the application domain and content verification as the BOWiki's two most outstanding novel features. Adaptability means that during setup, the software reads an OWL ontology selected by the user that provides a type system for the wikipages and the relations that are available to connect them. New relations can be introduced using specific wiki syntax, while the types remain fixed after setup.

While semantic wikis allow for the structured representation of information, they often provide little or no quality control, and do not verify the consistency of captured knowledge. Using the imported ontology as a type system in the BOWiki enforces the use of a common conceptualization and provides additional background knowledge about the selected domain. This background knowledge is used to check user-entered, semantic content by means of an OWL reasoner. For example, the ontology can prevent typing an instance of p45 both with *Protein* and *DNA molecule* at the same time. Currently, the performance of automated reasoners remains a limiting factor. Nevertheless, the reasoner delivers a form of quality control for the BOWiki content that should be adopted wherever possible.

The BOWiki was primarily designed to describe biological data using ontologies. In conjunction with a biological core ontology [Valente and Breuker, 1996] like GFO-Bio [Hoehndorf et al., 2007] or BioTop [Schulz et al., 2006], the BOWiki can be used to describe biological data. For this purpose, we developed a module that allows OBO flatfiles[2] to be imported into the BOWiki. By default, these ontologies are only accessible for reading; they are neither editable nor considered in the BOWiki's reasoning. Users can then create wikipages containing information about biological entities, and describe the entities both in natural language text and in a formally structured way. For the latter, they can relate the described entities to categories from the OBO ontologies, and these categories are then made available for use by the BOWiki reasoning.

In contrast to annotating data with ontological categories, i.e., asserting an undefined association relation between a biological datum and an ontological category, it is possible in the BOWiki to define precisely the relation between a biological entity (e.g. a class of proteins) and another category: a protein may not only be *annotated* to *transcription factor activity*, *nucleus*, *sugar transport* and *glucose*. In the BOWiki, it may stand in the *has_function* relation to transcription factor activity; it can be *located_at* a nucleus; it can *participate_in* a *sugar transport* process; it can *bind* glucose. The ability to make these relations explicit renders annotations in a semantic wiki both exceptionally powerful and precise.

The BOWiki can be used to describe not only data, but also biological categories, or to create relations between biological categories. As such, the BOWiki could be used to create so-called cross-products [Smith et al., 2007] between different ontologies.

[2] `http://www.cs.man.ac.uk/~horrocks/obo/`

3 IMPLEMENTATION

Within our MediaWiki extension, users can specify the type of entity described by a wikipage (see table 1). One of the central ideas of the BOWiki is to provide a pre-defined set of types and relations (and corresponding restrictions among them). We deliver the BOWiki with the biological core ontology GFO-Bio, but any *consistent* OWL [McGuinness and van Harmelen, 2004] file can be imported as the type system. Types are modelled as OWL classes and binary relations as OWL properties. Relations of higher arity are modeled according to use case 3 in [Noy and Rector, 2006], i.e., as classes whose individuals model relation instances. Wikipages as (descriptions of) instances of types give rise to OWL individuals, which may be members of OWL classes (their types).

An OWL ontology can provide background knowledge about a domain in the form of axioms that restrict the basic types and relations within the domain. This allows for automatic verification of parts of the semantic content created in the BOWiki: users may introduce a new page in the wiki and describe some entity; they may then add type information about the described entity; and this added type information is then automatically verified. The verification checks the logical consistency of the BOWiki's content – as OWL individuals and relations among them – with the restrictions of the OWL ontology's types and relations, like those in GFO-Bio. The BOWiki uses a description logic [Baader et al., 2003] reasoner to perform these consistency checks. We implemented the BOWikiServer, a stand-alone server that provides access to a description logic reasoner using the Jena 2 Semantic Web Framework [Carroll et al., 2003] and a custom-developed protocol. A schema of the BOWiki's architecture is illustrated in figure 1.

Whenever a user edits a wikipage in the BOWiki, the consistency of the changes with respect to the core ontology is verified using the BOWikiServer. Only consistent changes are permitted. In the event of an inconsistency, an explanation for the inconsistency is given, and no change is made until the user resolves the inconsistency.

In addition to verifying the consistency of newly added knowledge, the BOWikiServer can perform complex queries over the data contained within the wiki. Queries are performed as retrieval operations for description logic concepts [Baader et al., 2003], i.e., as queries for all individuals that satisfy a description logic concept description.

A performance evaluation of our implementation using the Pellet description logic reasoner [Sirin and Parsia, 2004] for ontology classification showed, that presently, only small- to medium-sized wiki installations can be supported. The time needed for consistency checks increases as the number of wiki pages increases[3].

4 DISCUSSION

Using different reasoners

The BOWikiServer provides a layer of abstraction between the description logic reasoner and the BOWiki. Depending on the description logic reasoner used, different features can be supported. Currently, the BOWikiServer uses the Pellet reasoner [Sirin and Parsia, 2004]. Pellet supports the explanation of inconsistencies,

[3] The results of our performance tests can be found on the wiki at `http://bowiki.net`.

BOWiki syntax	OWL abstract syntax
Generic	
1 `[[OType:C]]`	Individual(**page** type(C))
2 `[[R::page2]]`	Individual(**page** value(R page2))
3 `[[R::role1=page1;...;roleN=pageN]]`	Individual(R-id type(R)) Individual(R-id value(subject **page**)) Individual(R-id value(R-role1 page1)) ... Individual(R-id value(R-roleN pageN))
4 `[[has-argument::` `name=roleName;type=OType:C]]`	SubClassOf(**page** gfo:Relator) ObjectProperty(R-roleName domain(**page**) range(C))
Examples	
1 on page Apoptosis: `[[OType:Category]]`	Individual(Apoptosis, type(Category))
2 on page Apoptosis: `[[CC-isa::Biological_process]]`	Individual(Apoptosis value(CC-isa Biological_process))

3 on page HvSUT2:
`[[Realizes::`
`function=Sugar_transporter_activity;`
`process=Glucose_transport]]`
	Individual(Realizes-0 type(Realizes))
	Individual(Realizes-0 value(Realizes-subject HvSUT2))
	Individual(Realizes-0 value(Realizes-function Sugar_transporter_activity))
	Individual(Realizes-0 value(Realizes-process Glucose_transport))

4 on page Realizes:
`[[has-argument::`
`name=function;`
`type=OType:Function_category]]`
	SubClassOf(Realizes gfo:Relator))
	ObjectProperty(Realizes-function domain(Function_category))

Table 1. Syntax and semantics of the BOWiki extensions. The table shows the syntax constructs used in the BOWiki for semantic markup. The second column provides a translation into OWL. (**page** refers to the wikipage in which the statement appears; "R-id" is a name for an individual whose "id" part is unique and generated automatically for the occurrence of the statement). Because OWL has a model-theoretic semantics, this translation yields a semantics for the BOWiki syntax. In the lower half of the table we illustrate each construct with an example and present its particular translation to OWL.

which can be shown to users to help them in correcting inconsistent statements submitted to the BOWiki. It also supports the nonmonotonic description logic ALCK with the auto-epistemic **K** operator [Donini et al., 1997]. This permits both open- and closed-world reasoning [Reiter, 1980] to be combined, which has several practical applications in the Semantic Web [Grimm and Motik, 2005] and the integration of ontologies in biology [Hoehndorf et al., 2007]. On the other hand, reasoning in the OWL description logic fragment [McGuinness and van Harmelen, 2004] is highly complex. It is possible to use reasoners for weaker logics to overcome the performance limitations encountered with Pellet.

Comparison with other approaches

WikiProteins [Giles, 2007] is a software project based also on the MediaWiki software, focused on annotating Swissprot [Boeckmann et al., 2003]. Similar to the BOWiki, it utilizes ontologies like the Gene Ontology [Ashburner et al., 2000] and the Unified Medical Language System [Humphreys et al., 1998] as a foundation for the annotation. It is generally more targeted at creating and collecting definitions for terms than on formalizing knowledge in a logic-based and ontologically founded framework. As a result, it contains a mashup of lexical, terminological and ontological information. In addition, WikiProteins neither supports *n*-ary relations nor provides a description logic reasoner to retrieve or verify information. It therefore lacks the quality control and retrieval features that are central to the BOWiki. On the other hand, because of the different use-cases that WikiProteins supports, it is designed to handle much larger quantities of data than the BOWiki, and it is better suited for creating and managing terminological data.

The Semantic Mediawiki [Krötzsch et al., 2007] is another semantic wiki based on the Mediawiki software. It is designed to be applicable within the online encyclopedia Wikipedia. Because of the large number of Wikipedia users, performance and scalability requirements are much more important for the Semantic Mediawiki than for the BOWiki. Therefore, it also provides neither a description logic reasoner nor ontologies for content verification.

The IkeWiki [Schaffert et al., 2006], like the BOWiki, includes the Pellet description logic reasoner for classification and verification of consistency. In contrast to the BOWiki, parts of the IkeWiki's functionality require users to be experts in either Semantic Web technology or knowledge engineering. As a consequence, the BOWiki lacks some of the functionality that the IkeWiki provides (such as creating and modifying OWL classes) as it targets biologist users, most of whom are not trained in knowledge engineering.

Conclusion

We developed the BOWiki as a semantic wiki specifically designed to capture knowledge within the biological and medical domains. It has several features that distinguish it from other semantic wikis and from similarly targeted projects in biomedicine, most notably its ability to verify its semantic content for consistency with respect to background knowledge and its ability to access external OBO ontologies.

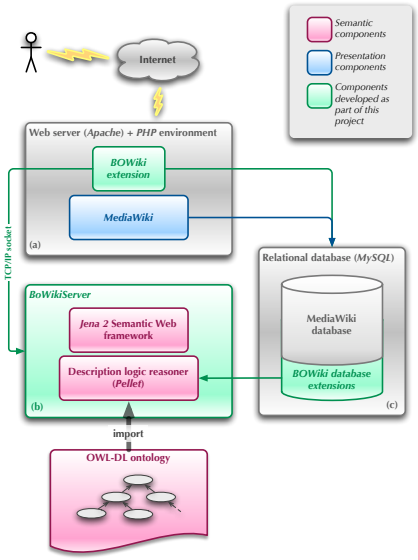


Fig. 1: BOWiki Architecture. (a) The BOWiki extension to the MediaWiki software processes the semantic data added to wiki pages. The semantic data is subsequently transferred to the BOWikiServer using a TCP/IP connection. (b) To evaluate newly entered data or semantic queries, the BOWikiServer requires an ontology in OWL-DL format (provided during installation of the BOWiki). Consistent semantic data will be stored. If an inconsistency is detected, the edited page is rejected with an explanation of the inconsistency. The BOWikiServer currently uses the Jena 2 Semantic Web framework together with the Pellet reasoner. (c) After successful verification the semantic data is stored in a separate part of the SQL database.

The BOWiki allows a scientific community to annotate biological data rapidly. This annotation can be performed using biomedical ontologies. In addition to data annotation, the specific type of relations between entities can be made explicit. It is also possible to integrate different biological knowledge bases by creating partial definitions for the relations and categories used in the knowledge bases.

The BOWiki employs a type system to verify the consistency of the knowledge represented in the wiki. The type system is provided in the form of an OWL knowledge base. If the type system is a core ontology for a domain (i.e., it provides background knowledge and restrictions about the categories and relations for the domain), its use contributes to maintaining the ontological adequacy of the BOWiki's content, and thereby the content's quality.

ACKNOWLEDGEMENTS

We thank Christine Green for her help in preparing the English manuscript.

REFERENCES

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, and *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29, 2000.

F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, Cambridge, UK, 2003.

B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, and *et al.*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*, 31 (1):365–370, January 2003.

J. J. Carroll, I. Dickinson, C. Dollin, D. Reynolds, A. Seaborne, and K. Wilkinson. Jena: Implementing the Semantic Web recommendations. Technical Report HPL-2003-146, Hewlett Packard, Bristol, UK, 2003.

F. M. Donini, D. Nardi, and R. Rosati. Autoepistemic description logics. In M. E. Pollack, editor, *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, IJCAI 1997, Nagoya, Japan, Aug 23-29*, volume 1, pages 136–141, San Francisco, 1997. Morgan Kaufmann.

J. Giles. Key biology databases go wiki. *Nature*, 445(7129):691, 2007.

S. Grimm and B. Motik. Closed world reasoning in the Semantic Web through epistemic operators. In B. Cuenca Grau, I. Horrocks, B. Parsia, and P. Patel-Schneider, editors, *Proceedings of the OWLED'05 Workshop on OWL: Experiences and Directions, Galway, Ireland, Nov 11-12*, volume 188 of *CEUR Workshop Proceedings*, Aachen, Germany, 2005. CEUR-WS.org.

R. Hoehndorf, K. Prüfer, M. Backhaus, H. Herre, J. Kelso, F. Loebe, and J. Visagie. A proposal for a gene functions wiki. In R. Meersman, Z. Tari, and P. Herrero, editors, *Proceedings of OTM 2006 Workshops, Montpellier, France, Oct 29 - Nov 3, Part I, Workshop Knowledge Systems in Bioinformatics, KSinBIT 2006*, volume 4277 of *Lecture Notes in Computer Science*, pages 669–678, Berlin, 2006. Springer.

R. Hoehndorf, F. Loebe, J. Kelso, and H. Herre. Representing default knowledge in biomedical ontologies: Application to the integration of anatomy and phenotype ontologies. *BMC Bioinformatics*, 8(1):377, 2007.

B. L. Humphreys, D. A. B. Lindberg, H. M. Schoolman, and G. O. Barnett. The Unified Medical Language System: an informatics research collaboration. *J Am Med Inform Assoc*, 5(1):1–11, 1998.

M. Krötzsch, D. Vrandečić, M. Völkel, H. Haller, and R. Studer. Semantic wikipedia. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(4):251–261, 2007.

D. L. McGuinness and F. van Harmelen. OWL Web Ontology Language overview. W3C recommendation, World Wide Web Consortium (W3C), 2004.

N. Noy and A. Rector. Defining N-ary relations on the Semantic Web. W3C working group note, World Wide Web Consortium (W3C), 2006.

R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13(1-2):81–132, 1980.

S. Schaffert, R. Westenthaler, and A. Gruber. IkeWiki: A user-friendly semantic wiki. In H. Wache, editor, *Demos and Posters of the 3rd European Semantic Web Conference, ESWC 2006, Budva, Montenegro, Jun 11-14*, 2006.

S. Schulz, E. Beisswanger, J. Wermter, and U. Hahn. Towards an upper-level ontology for molecular biology. *AMIA Annu Symp Proc*, 2006:694–698, 2006.

E. Sirin and B. Parsia. Pellet: An OWL DL reasoner. In V. Haarslev and R. Möller, editors, *Proceedings of the 2004 International Workshop on Description Logics, DL2004, Whistler, British Columbia, Canada, Jun 6-8*, volume 104 of *CEUR Workshop Proceedings*, pages 212–213, Aachen, Germany, 2004. CEUR-WS.org.

B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotech*, 25(11):1251–1255, 2007.

A. Valente and J. Breuker. Towards principled core ontologies. In B. R. Gaines and M. A. Musen, editors, *Proceedings of the 10th Knowledge Acquisition Workshop, KAW'96, Banff, Alberta, Canada, Nov 9-14*, pages 301–320, 1996.

K. Wang. Gene-function wiki would let biologists pool worldwide resources. *Nature*, 439(7076):534, 2006.

PubOnto: Open Biomedical Ontology-Based Medline Exploration

Weijian Xuan, Manhong Dai, Brian Athey, Stanley J. Watson and Fan Meng[*]

Psychiatry Department and Molecular and Behavioral Neuroscience Institute, University of Michigan, US

ABSTRACT

Motivation: Effective Medline database exploration is critical for the understanding of high throughput experimental results and the development of biologically relevant hypotheses. While existing solutions enhance Medline exploration through different approaches such as document clustering, network presentation of underlying conceptual relationships and the mapping of search results to MeSH and Gene Ontology trees, we believe the use of multiple ontologies from the Open Biomedical Ontology can greatly help researchers to explore literature from different perspectives as well as quickly locate the most relevant Medline records.

Availability: The PubOnto prototype is freely accessible at: http://brainarray.mbni.med.umich.edu/brainarray/prototype/pubonto

1 INTRODUCTION

The popularity of data driven biomedical research leads to large volumes of data such as gene expression profiles, MRI images and SNPs related to various pathophysiological processes. As a result, understanding the biological implications of the high throughput data has become a major challenge (Boguski and McIntosh, 2003). It requires time-consuming literature and database mining and is the main goal of the "literature-based discovery", "conceptual biology", or more broadly, "electronic biology", through which biologically important hypotheses are derived from existing literature and data using various approaches (Jensen, et al., 2006; Srinivasan, 2004; Swanson, 1990; Wren, et al., 2004). The effectiveness of such knowledge mining also relies heavily on researchers' background knowledge about novel genes or SNPs, and this knowledge, at present, is sparse.

The Medline database is without doubt the foremost biomedical knowledge database that plays a critical role in the understanding of high throughput data. Unfortunately, prevailing Medline search engines such as PubMed and Google Scholar have been designed largely for the efficient retrieval of a small number of records rather than an in-depth exploration of a large body of literature for discovery and proof purposes. They rely heavily on a step-wise narrowing of search scope but such an approach does not work well for the exploration of uncharted territories. This is because background knowledge is needed for defining sensible filtering criteria and guessing what are potentially relevant topics for additional exploration. For example, in microarray gene expression analysis, researchers frequently have to deal with lists of genes that are not known to be associated with the targeted biological processes. Researchers have to utilize other intermediate concepts to establish indirect links between gene lists and specific biological processes. However, identifying such intermediate concepts is very difficult in existing solutions and it is not easy even in systems devoted for this purpose such as ArrowSmith (Smalheiser, et al., 2007; Swanson, 1986). Frequently researchers have to go through large number of retrieved records one-by-one and examine external databases to find interesting new relationships.

Another major shortcoming for prevailing search solutions is that they do not present results in the contexts that a user maybe interested in. For example, Google Scholar/PubMed basically present search results as a linear list of papers. Users do not know the context of each paper nor the relationship among these papers. Besides the original ranking provided by the search engine, there is little additional cues and sorting/filtering methods that can facilitate the exploration of search results.

We believe the projection of search results to existing knowledge structure is very important for hypothesis development. This is because researchers often need to explore unfamiliar fields in the age of high throughput experiments and such projections can provide much needed guidance in new areas. In fact, even if a researcher wants to examine related facts in his/her own field, there are many details related to the search topic that require additional efforts to retrieve. Mapping search results to knowledge structures will also be very useful for revealing hidden relationships not easily identified by prevailing approaches. For example, if Medline search results show several genes in a brain region are related to a disease in a statistically significant manner, it will be worthwhile to explore the relationship of other genes expressed in this brain region with the disease. Naturally, exploration of multiple knowledge structures is often needed to facilitate the formation of new insights. The projection of search results to multiple dynamically-linked knowledge structures is thus necessary for such context-assisted data and literature exploration.

Newer Medline search solutions such as GoPubMed (Doms and Schroeder, 2005) and Vivisimo (Taylor, 2007)

[*] To whom correspondence should be addressed.

attempts to organize search results in the context of either predefined ontology such as Gene Ontology or dynamically generated ontology structures based on clustering results. In such solutions, users can rely on the tree-like organization of search results to easily navigate to topics of interest. The neighborhood of a given tree branch automatically suggests related topics for additional exploration. Here predefined ontologies have an advantage over clustering results for exploring unfamiliar territories due to their systematic listing of related concepts and their relationships.

However, given the huge number of biomedical concepts (e.g., over 1 million in the Unified Medical Language System) and the complexity of relationships among them, it is not possible to rely on one or two ontologies for effective exploration. Researchers must have the capability to examine their search results from different perspectives. We believe an ontology-based Medline exploration solution must allow the use of different orthogonal ontologies, i.e., ontologies that addressing different aspects of biomedical research. In addition, it is critical to enable interactive filtering of search results using terms from different ontologies for more efficient Medline exploration.

The main goal of this work is to develop a flexible ontology-based Medline exploration solution to facilitate the understanding of high throughput data and the discovery of potentially interesting conceptual relationships. Our solution enables interactive exploration of search results through the use of multiple ontologies from OBO foundry. It also has an open architecture that allows flexible selection of Medline retrieval algorithms through different web services.

2 METHODS

Selection of Ontologies: The Open Biomedical Ontologies (OBO) foundry is a comprehensive collaborative effort to create controlled vocabularies for shared use across different biological and medical domains (NCBO, 2008) (Rubin, et al., 2006; Smith, et al., 2007). It already includes around 50 ontologies from various biomedical domains. We selected Gene Ontology, Foundational Model of Anatomy, Mammalian Phenotype Ontology and Environment Ontology for inclusion in our prototype since they provide key perspectives for topics of great interest for biomedical research and they are almost orthogonal to each other conceptually.

Mapping of Ontology to Medline: We developed a very efficient general purpose ontology to free-text mapping solution in collaboration with researchers in the National Center for Biomedical Ontology. In brief, our solution relies on the pre-generation of lexical variations, word order permutations for ontology terms, their synonyms together with a highly efficient implementation of a suffix-tree based string match algorithm. Our solution is able to map all concepts in UMLS to the full Medline database in 15 hours on a mainstream Opteron server. It achieves over 95% recall rate when compared to the results from the MMTx program, which is about 500 times slower and does not support the use of non-UMLS ontologies. The details of our ontology mapping solution will be presented in a separate paper.

PubOnto Architecture: In order to provide a web-based Medline exploration tool with rich interactivity, we developed PubOnto on Adobe's latest Flex 3.0 platform. It allows us to build highly interactive user interface that is compatible in virtually all major browsers. We developed an innovative technique that dynamically updates the XML-based ontology tree structure by building a web service for each ontology for feeding expanded nodes with ontological information and literature searching results. As a result, only a minimum amount of data is transferred asynchronously and PubOnto can thus handle very large ontologies. Fig. 1 shows the architecture of PubOnto. Since the web service layer separates the user interface from ontologies, search services and other databases, the back end changes do not affect the client side user interface.

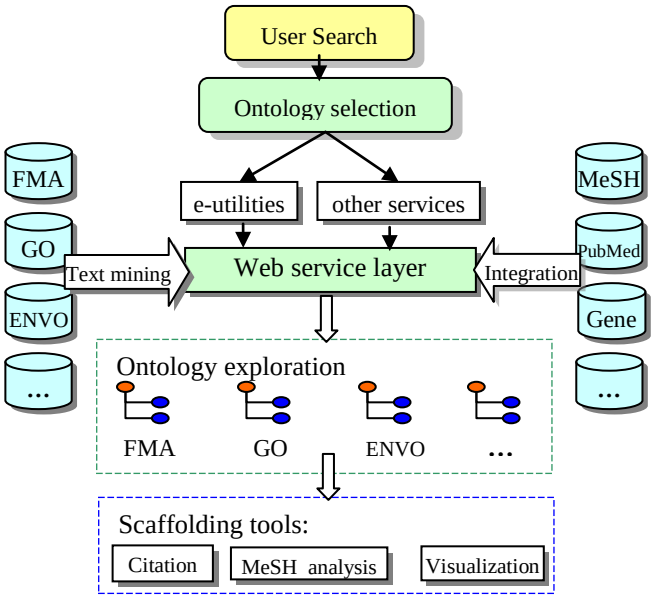


Fig. 1. PubOnto architecture

3 RESULTS

PubOnto is a FLEX application providing high level of interactivity for efficient Medline search result exploration. We illustrate a number of key features in this section.

Ontology-based exploration of search results: Simply displaying search results for each individual node is often not satisfactory. Typically users want to know quickly how many literatures are retrieved for all children under a branch so that they can decide if something is interesting that needs to be explored further. Rolling up such mapping data in a large ontology such as FMA or GO on-the-fly is not an easy task. Traditional tree traversal algorithms are very CPU intensive and usually require large in-memory tree structures

on the server. To provide real time interactivity, we pre-traverse the entire ontology and generate a parent-child table that matches all nodes in the subtree to their parent nodes. We also save the literature retrieval results to a session-based table. When a user expands a node, our service will perform an efficient table join to obtain the aggregated information.

Ontology Selection: PubOnto support a series of OBO ontologies. However, we also understand that users may not need to examine all of them. Therefore, we present a flexible way for users to choose which of the supported ontologies they want to use (Fig. 2). Once a user selects certain ontologies, PubOnto will dynamically create a new ontology tab with consistent display and interactive functions. We are also developing functions that can use multiple ontologies as combined filters to better navigate through citations.

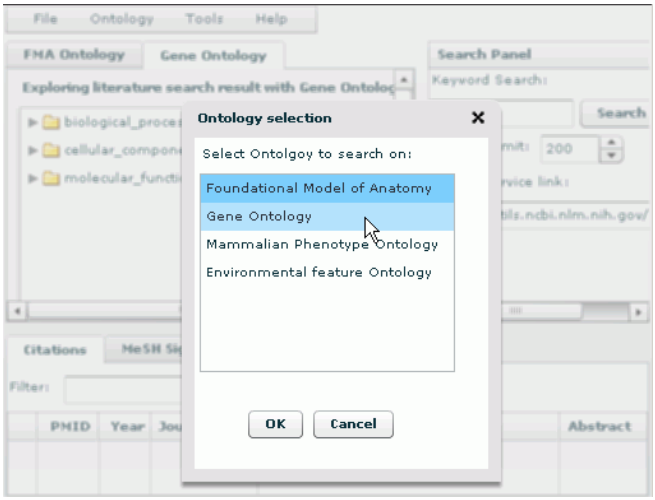


Fig. 2. Ontology selection

Search Result Exploration: When a user submits a keyword search request, web services will return retrieval results for each selected ontology. The user can expand tree nodes to explore results, as show in Fig. 3.

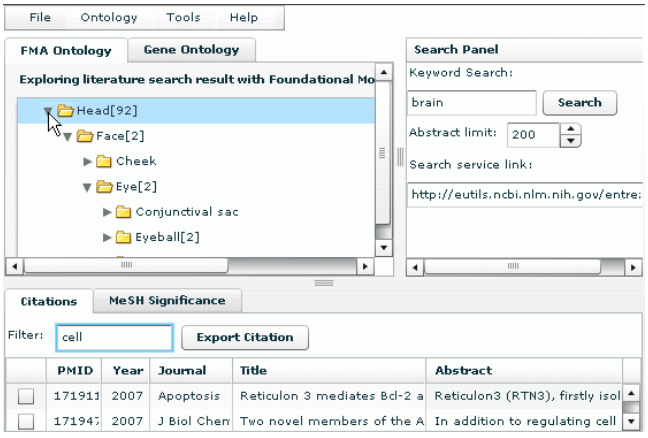


Fig. 3. Ontology-based exploration

Scaffolding Tools: PubOnto provides a number of tools for easy exploration. When a user clicks an ontology node, corresponding citations will show up in the bottom panel. Clicking on each citation will bring up a dialog for detailed citation information (Fig. 4). PubOnto also provide aggregated MeSH information to highlight concepts that are significant in this particular citation set versus the whole Medline corpus. In addition, PubOnto provides charting function to visualize MeSH concept distribution in search results.

Fig. 4. Citation exploration

While the most important feature of PubOnto is the ability to use multiple OBO ontologies for Medline exploration, it also offers a number of unique features summarized in Table 1. The PubOnto prototype currently does not include several functions in GoPubMed that are not directly related to ontology but similar functions will be added in future upon users' request.

Table 1. Comparison between PubOnto and GoPubMed

	PubOnto	GoPubMed
More ontologies besides MeSH, GO	Yes	No
Interaction among ontology	Yes	No
Customizable search service	Yes	No
Client side filter	Yes	No
Customizable ontology search	Yes	No
Customizable interaction function	Yes	No
Rich interactions	Yes	No
Search history maintenance	Yes	Yes
Sorting citation by various criteria	Yes	No
Export to Citation managers	Direct	Indirect
Citation linkouts	Yes	Yes
Where/Where/When analysis	No	Yes
Keyword highlight	No	Yes
Hot topics	No	Yes
Wikipedia mapping	No	Yes

4 DISCUSSIONS

Systematic ontology development efforts such as those related to the Open Biomedical Ontologies are likely to generate expansive conceptual framework for the integration, analysis and understanding of data generated in different areas of biomedical research. PubOnto aims to capitalize on the impressive progresses in ontology development for the exploration and mining of biomedical literature. The ability to utilize multiple orthogonal ontologies during Medline exploration can significantly increase the efficiency of locating interesting search results in areas that researchers are not familiar with. Mapping Medline results to multiple ontologies also enables researchers to explore search results from different angles for new hypothesis development.

While the PubOnto prototype provides a conceptual demo for the power of using multiple ontologies for Medline exploration, there are a number of improvements we hope to incorporate in the coming months. For example, although the ability to select different ontologies for organizing search results is quite powerful, it is based on the assumption that users know which ontologies they want to use. It should be possible to rank ontologies for their usefulness to the topic based on distribution of returned Medline records on different concepts under a given ontology. For example, an ontology is not very useful for Medline search result exploration if only a small fraction of returned records can be mapped to this ontology. On the contrary, an ontology will be very effective if many records can be mapped to it and those records are relatively evenly distributed across many terms in that ontology. Of course, an ontology is still not useful if most of the search results can be mapped to only a few terms in an ontology. Consequently, it should be possible to develop an ontology scoring system based on the number of records that can be mapped to an ontology and the distribution of Medline records in an ontology for the automatic selection of default ontology for a given Medline search result. Conceivably, once the first ontology is selected, it is possible to select the second best ontology based on the "orthogonality" with the first ontology. Of course, such automated ontology ranking procedures are only based on the statistical properties of the Medline records to ontology mapping. Users' biomedical knowledge and their understanding of different ontologies will be essential for effective exploration of Medline literature.

Similarly, the exploration of a given ontology tree currently is also dependent on users background knowledge since only the number of Medline records hits for a given term can be used as external cues for ontology exploration now. If there are many different ontologies for a user to select from or the user is not familiar with the corresponding ontology at all, it is desirable to have additional information to help users to use such ontology guided exploration more effectively. It is conceivable that we can weigh the specificity of each ontology term based on their inverse frequency of showing up in Medline corpus so that users can focus on more specific terms rather than exploring generic terms.

In summary, we believe the use of multiple ontologies in OBO for Medline exploration can significantly increase the efficiency of Medline exploration and facilitate the examination of the same search result from different perspectives. We will continue to improve PubOnto to make it an effective tool for novel biomedical hypotheses development, and ultimately incorporate it into PubViz, our more comprehensive biomedical literature exploration engine.

ACKNOWLEDGEMENTS

W. Xuan, M. Dai, S. J. Watson and F. Meng are members of the Pritzker Neuropsychiatric Disorders Research Consortium, which is supported by the Pritzker Neuropsychiatric Disorders Research Fund L.L.C. This work is also partly supported by the National Center for Integrated Biomedical Informatics through NIH grant 1U54DA021519-01A1 to the University of Michigan

REFERENCES

Boguski, M.S. and McIntosh, M.W. (2003) Biomedical informatics for proteomics, *Nature*, **422**, 233-237.

Doms, A. and Schroeder, M. (2005) GoPubMed: exploring PubMed with the Gene Ontology, *Nucleic Acids Res*, **33**, W783-786.

Jensen, L.J., Saric, J. and Bork, P. (2006) Literature mining for the biologist: from information retrieval to biological discovery, *Nature Reviews Genetics*, **7**, 119-129.

NCBO (2008) The Open Biomedical Ontologies, http://www.obofoundry.org/.

Rubin, D.L., Lewis, S.E., , et al (2006) National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge, *Omics*, **10**, 185-198.

Smalheiser, N.R., Torvik, V.I., et al. (2007) Arrowsmith, http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html.

Smith, B., Ashburner, M., Rosse, C., et al (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration, *Nat Biotechnol*, **25**, 1251-1255.

Srinivasan, P. (2004) Text mining: Generating hypotheses from MEDLINE, *JASIST*, **55**, 396-413.

Swanson, D.R. (1986) Fish oil, Raynaud's syndrome, and undiscovered public knowledge, *Perspect Biol Med*, **30**, 7-18.

Swanson, D.R. (1990) Medical literature as a potential source of new knowledge, *Bull Med Libr Assoc*, **78**, 29-37.

Taylor, D.P. (2007) An integrated biomedical knowledge extraction and analysis platform: using federated search and document clustering technology, *Methods Mol Biol*, **356**, 293-300.

Wren, J.D., Bekeredjian, R., et al (2004) Knowledge discovery by automated identification and ranking of implicit relationships, *Bioinformatics*, **20**, 389-398.

Minimal Anatomy Terminology (MAT): a species-independent terminology for anatomical mapping and retrieval

Jonathan B.L. Bard[1], James Malone[2], Tim F. Rayner[2] and Helen Parkinson[2]

1. Computational Biology Research Group, Weatherall Institute for Molecular Medicine, John Radcliffe Hospital, Oxford OX3 9DS, UK
2. EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

ABSTRACT

Motivation: The problem of integrating a multiplicity of non-orthogonal anatomy ontologies is well known in ontology development. There are now major public ontology repositories (e.g. the Ontology for Biomedical Ontologies) that require a multi-species anatomy ontology. We present MAT (Minimal Anatomy Terminology) an OBO format terminology (~400 terms) using SKOS *broader-than* relationships designed for annotating and searching tissue-associated data and timelines for any organism. Identifiers from >20 anatomy ontologies are mapped to each MAT term to facilitate access to and interoperability across tissue-associated data resources

Availability: www.ebi.ac.uk/microarray-srv/mat/

1 INTRODUCTION

Data in public biomedical databases typically has various classes of metadata has associated with it that enable searching and analysis, and standards for different data types and domains are now becoming available (e.g. a series of *Minimum Information* protocols for this purpose, mibbi.sourceforge.net/resources.shtml). There is no such minimal standard for annotating anatomy because tissues are much harder than other (e.g. experimental) data types to formalize simply. This is partly because organisms have so many diverse tissues and partly because tissue organization is so complex. Nevertheless, because of the need to handle tissue-associated data in databases, user communities for all of the main model organisms have produced formalized and fairly complete anatomical hierarchies (ontologies) that are largely based on *part_of* and *is_a* relationships (Bard, 2005, 2007; Smith et al., 2007; Burger et al., 2007). These high-granularity ontologies are complex and their use presupposes considerable anatomical knowledge of the organism whose anatomy is represented, as well as some understanding of the representation format of the ontology. They are therefore mainly used by specialist curators annotating data for the main model organism databases using rich annotation tools (e.g. Phenote, www.phenote.org).

Elsewhere, anatomical annotation is essentially free text, or at best loosely controlled. Databases such as those from the NCBI (www.ncbi.nlm.nih.gov) typically do not control anatomical annotation, and text mining is needed to extract any anatomical information. It is unrealistic for these multi-species databases archiving high throughput data to develop annotation tools that provide intuitive access to all (anatomy) ontologies and expect biomedical users to use them consistently. ArrayExpress (www.ebi.ac.uk/microarray-as/aer/entry), for example, uses a text-mining strategy and string-matching methodology that adds no burden at the point of submission but does require representative ontologies for automated annotation (Parkinson et al, 2006).

For query purposes, a simple anatomical ontology is needed that allows searching and tree browsing, with its complexity limited to that which is comprehensible to a bench biologist. The simplest format for accessing annotation terms is a controlled vocabulary or terminology where informal relationships connect the terms (unlike an ontology whose formal relationships carry inheritance implications). Two such terminologies are currently available: the eVOC terminology set (Kelso et al., 2003) whose scope is limited to human and mouse, and the very short SAEL terminology (Parkinson et al., 2004) mainly intended for core mammalian anatomical annotation and which has no relations at all. Neither resource includes identifiers for other anatomy-based resources that can be used for cross-mapping and interoperability purposes.

This paper reports the development and validation of a terminology entitled MAT (*Minimal Anatomy Terminology*). It is similar in format to eVOC but expanded to include high-level tissues and timelines appropriate for the great majority of taxa rather than just mammals. Data associated with these tissue terms include synonyms and ontology identifiers for tissues from other anatomical ontologies currently downloadable from the Open Biomedical Ontologies (OBO) website (obofoundry.org/), and is thus compatible with them.

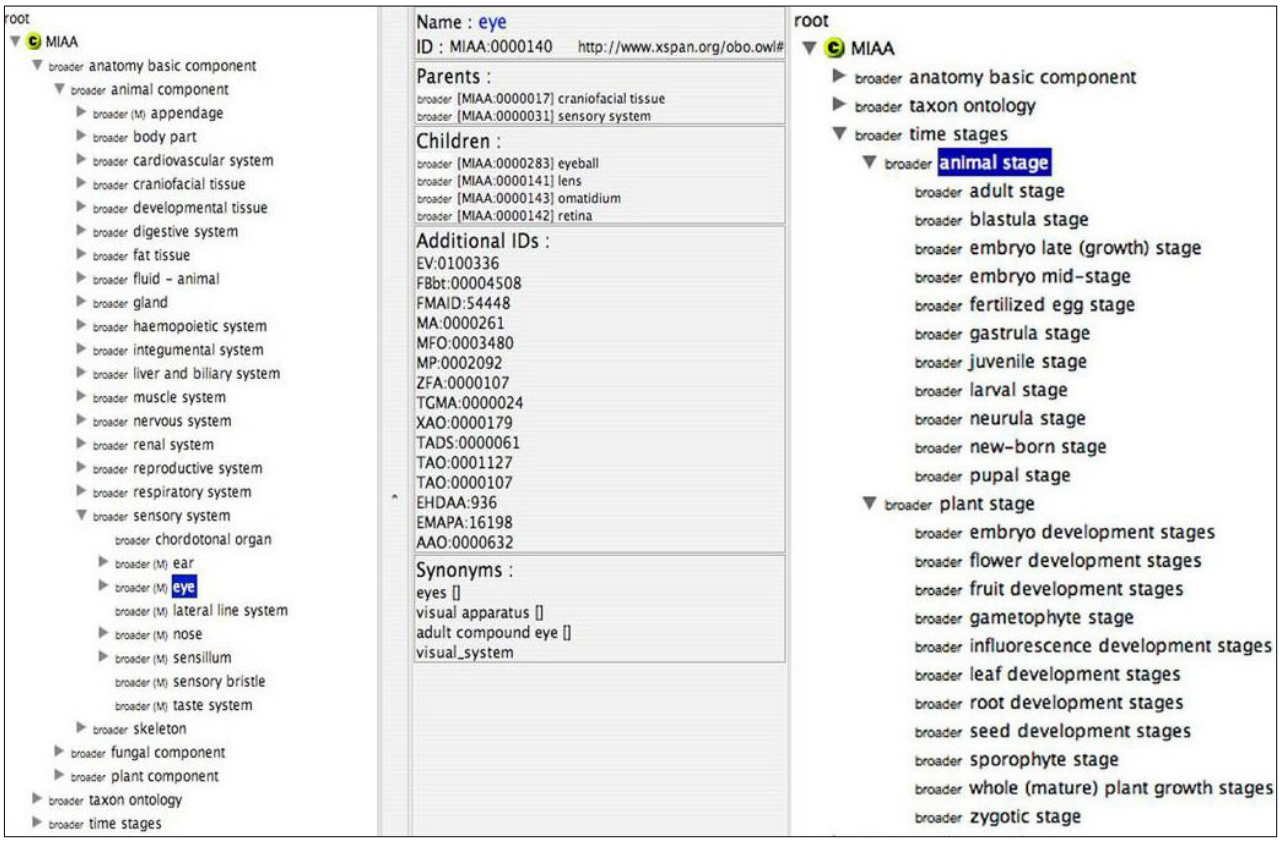


Figure 1 MAT terminology displayed in the CoBrA editor (www.xspan.org/cobra). The left panel shows the four top categories with 'anatomy basic component' expanded. The 'eye' has been expanded in the middle panel to show parent and child terms, synonyms and identifiers. The right panel shows the organizing classes 'taxon ontology' and 'time stages'

The MAT terminology is designed to facilitate the easy annotation, curation and searching of tissue-associated data while the ready availability of the various ontology identifiers will facilitate tissue-associated interoperability across databases

2 METHODS & RESULTS

Scope

The MAT terminology is intended to cover the basic anatomy for all common taxa from fungi to plants and animals to support anatomical information and mapping of data contained in existing public resources. MAT is not intended to represent formal knowledge about all these organisms with its inherent implications for inheritance.

Identifiers

Each term has an identifier of the form MAT:0000001, and is mapped to one or more identifiers from the anatomy ontologies currently available in the OBO foundry (Smith et al., 2007). It also contains identifiers from the Mammalian Phenotype Ontology (Smith, 2004) as these typically include anatomical information and may be useful in the context of mapping abnormal phenotypes within an anatomical context.

Granularity

Determining the appropriate level of granularity for MAT is critical: too light and its archiving and searching uses would be inadequate; too heavy a complexity would prohibit use by non-anatomists. The main indicator for tissue selection is that formal species-specific ontologies (*Drosophila*, mouse etc) include these terms at a high level in their respective representations. A second indicator is that the selected tissues should be accessible for molecular analysis. A third was that their meaning was obvious and unambiguous to a biological user.

The current version of MAT has ~400 anatomical child terms of the class *anatomy basic component* (Fig. 1). The majority of these are used in their stage-independent form. This is possible as most of the external ontologies to which MAT is mapped have either restricted their scope to adults or are structured so time and tissue are handled independently (Burger et al., 2003).

Organizing principles

The terminology is intended to be intuitively navigable by a biologist, and obvious choices for high level terms in the hierarchy were *organ* and *major tissue systems* as these both underpin anatomical organization in an intui-

tive way and are used by most anatomical ontologies. MAT includes ~300 animal, ~75 pant and ~20 fungal systems and tissues, Where tissues naturally fall into more than one system (e.g. the mouth is both a craniofacial tissue and part of the alimentary system), multiple inheritance has been used.

MAT includes two high level nodes in addition to *anatomy basic component*: *taxon ontology,* and *time stage* (Fig. 1). Detailed staging for each organism is outside the scope of MAT, but a generic set of 11 stages each for animals and plants that extend from the zygote to adult are used. This allows distinctions to be made between, for example, the embryonic, the juvenile and the adult testes.

It should be noted that a few ontologies (e.g. adult human and adult mouse) only handle adult tissues, and their identifiers should not be used for developmental tissues.

As the MAT terminology is designed for mapping and annotation rather than for logical inference, the use of formal relationships such as *is-a* and *part-of* were replaced by the single *broader-than* relationship used as defined by SKOS, the Simple Knowledge Organization System, (www.w3.org/2004/02/skos/, a part of the Semantic Web (www.w3.org/2001/sw/). This allows us to represent the terminology as a tree with a single, informal relationship carrying no inheritance implications.

The MAT terms are intended to be species-independent, *trachea* in the *respiratory system* has the associated identifiers from the *Drosophila*, human and mouse anatomy ontologies even though the insect and vertebrate tracheae are very different – they are analogues and not homologues. A *sensu* tag is used in only twice: the vertebrate and invertebrate limbs are so different in structure and development that it seemed unreasonable to include them under the same term, while the insect and amphibian fat bodies are neither homologues nor analogues. MAT also contains some transitional development-specific tissues with no timing details (e.g. somite). These terms were included as they would thus not be present in adult organism lists.

Different anatomy ontologies use different terms and spellings for what are essentially equivalent terms (e.g. oesophagus and esophagus, digestive system and alimentary system). We have made a subjective decision to use the most common term as the standard (e.g. *eye* rather than *visual system*, see Fig. 1), but synonyms are included in the file and can be searched. In assigning identifiers from other anatomy ontology to MAT terms (~1600 in all), there was sometimes a choice as to which term to map to. In the *Drosophila* ontology, for example, there is a term for the digestive system and sub-terms for the embryonic/larval digestive systems and the pupal/adult digestive systems. Where alternatives exist the broadest term is used preferentially. A very few tissues have been included that are not present in other anatomical ontologies as they may be interesting in a wider evolutionary context (e.g. *phyllid*, the gametophyte leaf).

The MAT terminology has very few text definitions as almost all the terms are in common use by biologists. Indeed, it was often impossible to provide anything but a very loose definition for tissues from different taxa with the same name (e.g. mammalian and invertebrate *trachea* are both involved in the respiratory system, and this is explicit in the terminology). The definitions that are provided cover tissues that may be unfamiliar (e.g. phyllid) or whose meaning is slightly technical (e.g. mesonephros – adult).

We explicitly decided not to use the CARO upper level anatomy ontology (Haendel et al., 2007) as it is not intuitive to the biologist and is therefore not useful for use in annotation tools or browsing data, and is actually intended for use as a template in developing anatomy ontologies rather than for representing multi-species mappings. We also decided not to adopt the view that multiple parentage of terms is undesirable as we are not trying to represent full anatomical knowledge, rather to produce a resource to aid data integration pragmatically, and biologists intuitively comprehend multiple parentage as is, for example, present in the Gene Ontology (Ashburner, et al., 2000).

Validation

Prior to the construction of the MAT terminology, the ArrayExpress user supplied annotation of 'OrganismPart' comprising 817 unique terms used in the annotation of >60,000 samples obtained from >200 species was mapped to multiple anatomy ontologies using a Perl implementation of the MetaPhone 'sound-alike' algorithm (Phillips, 1990). The FMA was found to provide the highest coverage of all existing anatomy ontologies, but still covered only 38% of ArrayExpress anatomical annotations. MAT was mapped to the ArrayExpress annotations three times during the development of the MAT terminology and the results curated to identify categories of coverage. The final coverage is ~39%.This figure is comparable with the FMA, and uses only 400 terms to achieve the same coverage. The FMA in contrast contains 25,000 terms and is far less tractable in the context of annotation tools and usability for the general biomedical scientist.

Format

The MAT terminology uses the OBO (oboedit.org) flat file format which allows SKOS relationships, and was constructed using the COBrA editor which has good annotation capabilities that facilitate the mapping of properties such as identifiers and synonyms to MAT terms (www.xspan.org/cobra).

3: DISCUSSION

The aim of the MAT controlled vocabulary is not only to produce a standard terminology which can be assigned to any anatomical parts from any organism, but to provide primary search terms for those interested in accessing tissue-associated data. It is intended as a way of integrating data and allowing interoperation between many ontologies.

MAT is also intended to help with the strategy of determining the molecular basis of some process in one organism by using information relevant to its development and function gleaned from other organisms. Here, MAT provides candidate tissues and identifiers, although MAT tissue groupings may or may not be viewed as equivalent in any particular context, and the onus is on the user to choose which tissues may be relevant to their own and, in turn, which associated data is helpful.

MAT may also be useful in the wider context: as more data is being generated and funders and journals require data to be archived, it becomes impossible for database curators to keep up with the annotation needed for archiving the files, and indeed, harder for the funding agencies to be able to provide the necessary financial support. A practical solution to this problem is that people who deposit material in databases annotate their own data in at least in part. This has always been difficult to achieve formally for tissue-associated data, and we hope the use of terminologies such as MAT will be helpful here.

We expect that the majority of users will be interested in a limited number of taxa, and an editing tool (e.g. COBrA or OBO-edit) can be used to select only the tissues for particular organisms. Note that MAT does not seek to replace the existing ontologies. A more common problem may be that MAT's granularity may be too coarse, and terms may need to be added. This could be solved by using a species-specific ontology, free text, or evolving the MAT for a specific groups needs. Suggestions, criticisms and requests should be emailed to j.bard@ed.ac.uk.

The MAT terminology does not address the issue of developing an all-encompassing multi-species ontology that precisely describes orthologous anatomical parts across evolutionary time. This is a much larger task and has been attempted in the development of the Bilateria ontology used in the 4DExpress database of developmental gene expression data (Haudry et al., 2008). We and others are participating in discussions to make this a more general effort. We applaud these efforts and hope that MAT will be useful in the interim.

ACKNOWLEDGEMENTS

Thanks to Dawn Field for comments on the manuscript, and to Dawn, Stuart Aitken, Nick Kruger, Robert Stevens and Steve Taylor for discussions.

FUNDING

JB thanks the Leverhulme Trust, HP, JM, TFR are funded in part by EC grants FELICS (contract number 021902), EMERALD (project number LSHG-CT-2006-037686), Gen2Phen (contract number 200754) and by EMBL.

REFERENCES

Ashburner, M, Ball, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, Nat Genet, 25, 25-29.

Bard JBL (2005) Anatomics: the intersection of anatomy and bioinformatics. J Anat 206: 1-16.

Bard J (2007) Anatomy ontologies for model organisms: the animals and fungi. In: Burger A, Davidson D, Baldock RA editors. Anatomy Ontologies for Bioinformatics. Springer. pp 3-26.

Burger A, Davidson D, et al. (2003) Formalization of mouse embryo anatomy, Bioinformatics 19: 1-9.

Burger A, Davidson D, et al. (2007) (editors) Anatomy Ontologies for Bioinformatics. Springer. pp 356.

Haendel, MA., Neuhaus, F, et al (2007). CARO – the common anatomy reference ontology. In: Burger A, Davidson D, Baldock RA (editors). Anatomy Ontologies for Bioinformatics. Springer. pp 327-350.

Haudry Y, Berube H, et al. (2008). 4DXpress: a database for cross-species expression pattern comparisons. Nuc Acids Res 36: D847-53.

Kelso J, Visagie J, et al (2003). eVOC: a controlled vocabulary for unifying gene expression data. Genome Res 6A: 1222-30.

Parkinson H, Kapushesky M, et al. (2006). ArrayExpress -a public database of microarray experiments and gene expression profiles. Nucl Acids Res 35: D747-750.

Parkinson H, Aitken S, et al. (2004) The SOFG Anatomy Entry List (SAEL): An Annotation Tool for Functional Genomics Data. Comp Funct Gen 5: 521-527.

Phillips L (1990) Hanging on the Metaphone. Comput Lang 7: 39-49

Smith B, Ashburner M, et al. (2007). The OBO Foundry: coodinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 25: 1251

Smith, C, Goldsmith, C-A and Eppig, J (2004) The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information, Genome Biology, 6.R9

Function, Role, and Disposition in Basic Formal Ontology

Robert Arp[*] *and Barry Smith*

National Center for Biomedical Ontology (NCBO) and New York State Center of Excellence in Bioinformatics and Life Sciences, University at Buffalo, rarp@buffalo.edu, phismith@buffalo.edu

ABSTRACT

Numerous research groups are now utilizing Basic Formal Ontology (BFO) as an upper-level framework to assist in the organization and integration of biomedical information. This paper provides elucidation of the three BFO categories of *function*, *role*, and *disposition*, and considers two proposed sub-categories of *artifactual function* and *biological function*. The motivation is to help advance the coherent treatment of functions, roles, and dispositions, to help provide the potential for more detailed classification, and to shed light on BFO's general structure and use.

1 INTRODUCTION

Many of the members of the Open Biomedical Ontologies (OBO) Foundry initiative, including the Gene Ontology, the Foundational Model of Anatomy, the Protein Ontology, and the Ontology for Biomedical Investigations (http://www.obofoundry.org/) are utilizing Basic Formal Ontology (BFO) to assist in the categorization of entities and relationships in their respective domains of research.

Fig. 1. The continuant categories of BFO.

- BFO:entity
  - continuant
    - independent continuant
      - object
      - object boundary
      - object aggregate
      - fiat object part
      - site
    - dependent continuant
      - generically dependent continuant
      - specifically dependent continuant
        - quality
        - realizable entity
          - function
          - role
          - disposition
    - spatial region
      - zero-dimensional region
      - one-dimensional region
      - two-dimensional region
      - three-dimensional region

Many individuals and groups involved in organizations such as BioPAX, Science Commons, Ontology Works, AstraZeneca, and the Computer Task Group utilize BFO as well.

Fig. 2. The occurrent categories of BFO.

- BFO:entity
  - occurrent
    - processual entity
      - process
      - process boundary
      - process aggregate
      - fiat process part
      - processual context
    - spatiotemporal region
      - scattered spatiotemporal region
      - connected spatiotemporal region
        - spatiotemporal instant
        - spatiotemporal interval
    - temporal region
      - scattered temporal region
      - connected temporal region
        - temporal instant
        - temporal interval

Versions of BFO in OBO, OWL and first-order logic formats are maintained by Holger Stenzhorn at http://www.ifomis.org/bfo. Definitions and other content taken from there have been modified to provide additional clarity of exposition.

BFO is an upper-level ontology developed to support integration of data obtained through scientific research. It is deliberately designed to be very small, in order that is should be able to represent in consistent fashion the upper-level categories common to domain ontologies developed by scientists in different domains and at different levels of granularity. BFO adopts a view of reality as comprising (1) *continuants*, entities that continue or persist through time, such as objects, qualities, and functions, and (2) *occurrents*, the events or happenings in which continuants participate. The subtypes of continuant and occurrent represented in BFO are presented in Figures 1 and 2 (Grenon and Smith, 2004; Smith and Grenon, 2004; http://www.ifomis.uni-saarland.de/bfo/).

[*] To whom correspondence should be addressed.

2 FUNCTION, ROLE, AND DISPOSITION

Use of the term 'function' is common in descriptions of molecular and cellular processes, as in assertions such as:

- the function of the kidney of *Mus musculus* is to filter out waste and water which become urine,
- *Arabidopsis thaliana* has a multifunctional protein
- *there are several fo*lD bifunctional proteins in *Campylobacter jejuni*.

Functions thus play a central role in the Gene Ontology (http://www.geneontology.org/).

What, however, of the non-biological functions of artifacts such as screwdrivers, microplates, or pycnometers? Are there both designed (artifactual) and natural (biological) functions, representing distinct subtypes of the more general category of BFO:function?

A related issue is that of the use of the terms 'function' and 'role'. These are distinguished by BFO as representing two distinct categories (Figure 1), but outside BFO circles they are often used interchangeably, as when function is defined as 'the *role* that a structure plays in the processes of a living thing'. Analogous difficulties arise with regard to the terms 'disposition' and 'tendency', as in: 'blood has the tendency or disposition to coagulate', 'a hemophiliac has the disposition or tendency to bleed an abnormally large amount of blood', and 'that patient has suicidal dispositions or tendencies'.

In this paper, we attempt to elucidate the categories of *function*, *role*, and *disposition* in BFO. We also describe two sub-type categories of function, the *artifactual* and the *biological*, and provide definitions for each.

Within the context of BFO, one should correctly state:

- the (or a) *function* of the heart is to pump blood
- the *role* of the surrogate is to stand in for the patient
- blood has the *disposition* to coagulate
- that patient has suicidal *tendencies*

To see why this is so, we need first to consider BFO's more general approach to classification.

In BFO, all entities are divided into *continuants* and *occurrents*; continuants in turn are divided into *independent* and *dependent*. Independent continuants are things (the objects we see around us every day) in which dependent continuants—such as qualities, functions, roles, dispositions—can inhere.

Dependent continuants stand to their bearers in the relation of existential dependence: in order for them to exist, some other (independent) entity must exist. For example, instances of qualities such as *round* and *red* are dependent continuants in that they cannot exist without being qualities of some independent continuant such as a ball or a clown's nose. So too, functions, roles, and dispositions exist only insofar as they are functions, roles, and dispositions of some (one or more) independent continuant. The function of my heart is an instance of the BFO type *function*, and so also is the function of your heart.

One major subcategory of dependent continuants in BFO is that of *realizable entity*. Realizable entities are defined by the fact that they can be realized (manifested, actualized, executed) in occurrents of corresponding sorts. Examples of realizable entity types include: the function of the liver to store glycogen, the role of being a doctor, the disposition of metal to conduct electricity.

Realizable entities are entities of a type whose instances are typically such that in the course of their existence they contain periods of actualization, when they are manifested through processes in which their bearers participate. They may also exhibit periods of dormancy where they exist by inhering in their bearers, but are not manifested, as for example, in the case of certain diseases. Some realizables, such as the function of a sperm to penetrate an ovum, may be such that they can be manifested only once in their lifetime; or, as again in the case of sperm, they are realized only in very rare cases.

We are now in a position where we can define *function*, *role*, and *disposition*.

2.1 Function

A *function* f is

(1) a realizable dependent continuant.

Thus,

(2) it has a bearer, which is an independent continuant, and

(3) it is of a type instances of which typically have realizations; each realization is
   a. a process in which the bearer is participant
   b. that occurs in virtue of the bearer's physical make-up,
   c. and this physical make-up in something which that bearer possesses because of how it came into being.

Examples include: the function of a birth canal to enable transport and the function of a hammer to drive in nails. The process under a. may be specified further as an end-directed activity, by which we mean in the biological case something like: an activity that helps to realize the characteristic physiology and life pattern for an organism of the relevant type. Each function has a bearer with a physical structure which, in the biological case, the bearer has naturally evolved to have (as in a hypothalamus secreting hormones) or, in the artifact case, the bearer has been constructed to have (as in an Erlenmeyer flask designed to hold liquid) (Ariew and Perlman, 2002).

It is not accidental or arbitrary that a given eye has the function to see or that a given screwdriver have been designed and constructed with the function: to fasten screws. Rather, these functions are integral to these entities in virtue

of the fact that the latter have evolved or been constructed to have a corresponding physical structure.

If a continuant has a function, then it is built to exercise this function reliably on the basis of this physical structure. But again: a function is not in every case exercised or manifested. Its bearer may be broken; it may never be in the right kind of context. Hence, when we say that a given structure is designed in such a way as to bring about a certain end reliably, then this reliability presupposes the fulfillment of certain conditions, for example of an environmental sort.

On the level of instances, this can be stated as: if f is the function of c, then (in normal circumstances), c exercises f.

On the level of universals, as: if F is the function universal exemplified by instances of the independent continuant universal C, then (in normal circumstances) instances of C participate in process instances which are realizations of F. The implications of this analysis for the treatment of functions in the Gene Ontology are outlined in Hill, Smith, McAndrews-Hill, and Blake (2008).

2.2 Role

In contrast to function, *role* is a realizable entity whose manifestation brings about some result or end that is not typical of its bearer in virtue of the latter's physical structure. Rather, the role is *played* by an instance of the corresponding kind of continuant entity because this entity is in some special natural, social, or institutional set of circumstances (http://www.ifomis.org/bfo).

Examples include: the role of a chemical compound to serve as analyte in an experiment, the role of penicillin in the treatment of a disease, the role of bacteria in causing infection, the role of a person as student or surgeon.

What is crucial for understanding a role—as distinct from a function—is that it is a realizable entity that an independent continuant can take on, but that it is not a reflection of the in-built physical structure of that independent continuant. Certain strains of *Escherichia coli* bacteria have the role of pathogen when introduced into the gut of an animal, but they do not have this role when merely floating around in a pool of water. A heart has the function of pumping blood; but in certain circumstances that same heart can play the role of dinner for the lion.

Roles are optional, and they often involve social ascription. This is why a person can play the role of being a lawyer or a surrogate to a patient, but it is not necessary for persons that they be lawyers or surrogates.

So, when researchers are considering whether some realizable entity is a function or a role, the question to ask is this: Is the realizable entity such that its typical manifestations are based upon its physical structure? If so, then it is a function. Or, is the realizable entity such that its typical manifestation is a reflection of surrounding circumstances, especially those involving social ascription, which are optional? If so, then it is a role.

From this perspective, it is incorrect to make assertions such as:

- the *role* of the heart is to pump blood;
- driving nails is a *role* that this hammer fulfills;
- the *function* of the surrogate is to stand in for the patient;
- the *function* of James is to serve as my servant.

2.3 Disposition versus Tendency

It is common to find researchers making claims like: 'water has the disposition to rise in a tube', 'Carbon-10 has a disposition to decay to Boron-10', and 'the cell wall is disposed to filter chemicals in endocitosis and exocitosis.' A *disposition* is a realizable dependent continuant that typically causes a specific process in the object in which it inheres when the object is introduced into certain specific circumstances. In addition, these processes occur as a result of the object's physical structure (Jansen, 2007).

A disposition invariably leads to a certain result given certain circumstances. Consider: the disposition of a car windshield to break if struck with a sledgehammer moving at 100 feet per second; the disposition of a cell to become diploid following mitosis; the disposition of a magnet to produce an electrical field.

Contrasted with a disposition is a *tendency*, which is a realizable dependent continuant that potentially (not invariably or definitely) causes a specific process in the object in which it inheres when the object is introduced into certain specific circumstances as a result of the object's physical structure property.

Examples include: the tendency on the part of a hemophiliac to bleed an abnormally large amounts of blood and the tendency on the part of a person who smokes two packs of cigarettes a day throughout adulthood to die of a disease at a below average age. A patient may have a tendency, and not a disposition, to commit suicide; while a crystal vase has a disposition, and not a tendency, to break when it hits the ground after being dropped from a tall building. We are referring to tendencies when we refer to genetic and other *risk factors* for specific diseases.

3 TWO SUB-CATEGORIES OF FUNCTION

It is possible that BFO has failed to recognize categories or sub-categories of entities existing in reality. The ontology is, however, developed on the basis of a principle of scientific fallibilism (Grenon and Smith, 2004). Thus, it is possible that future research in ontology or in the natural sciences will reveal the need for an expansion or restructuring of the categories that BFO recognizes.

In its present form, BFO categories are those included in the taxonomic hierarchy illustrated in Figures 1 and 2 above. However, we are exploring the possibility of introducing two sub-categories under function, namely *artifac-*

tual function and *biological function*, as illustrated in Figure 3.

We are also exploring the question of whether to include tendency as a further sub-category within the ontology.

Fig. 3. Two proposed sub-categories of function in BFO.

BFO:realizable entity
- function
  - artifactual function
  - biological function
- role
- disposition

3.1 Artifactual Function

An *artifactual function* is a function which inheres in an independent continuant that exists, and has the physical structure which it has, because it has been designed and made intentionally (typically by one or more human beings) to function in a certain way and does indeed reliably function in this way (Lind, 1994; Dipert, 1993).

Examples include: the function of a pycnometer to hold liquid, the function of a fan to circulate air, and the function of a Bunsen burner to produce a flame.

3.2 Biological Function

A *biological function* is a function which inheres in an independent continuant that is (i) part of an organism and (ii) exists and has the physical structure it has as a result of the coordinated expression of that organism's structural genes (Rosse and Mejino, 2003). The manifestations of a function of this sort form part of the life of the organism.

Examples include: the function of a mitochondrion in the production of ATP and the function of the wax-producing mirror gland of the worker honey bee to produce beeswax.

The manifestations of biological functions are not in every case beneficial to the survival of the corresponding organism. (Consider the case of organisms that die when they reproduce, like *Arabis laevigata* and *Octopus lutens*.) Rather, they are (in typical environments) such as to contribute to the realization by an organism of a life that is typical or characteristic for an organism of its kind.

It is an open question whether the dichotomy between biological and artifactual function should or should not be included as an addition to BFO, or reflected rather in the creation of two new domain ontologies of artifactual and of biological functions. The latter has already been proposed as a complement to the GO's molecular function and biological process ontologies.

ACKNOWLEDGEMENTS

We wish to thank Andrew Spear for helpful comments. This work is funded by the United States National Institutes of Health (NIH) through the NIH Roadmap for Medical Research, Grant 1 U54 HG004028.

REFERENCES

Allen, C., Bekoff, M., and Lauder, G., eds. (1998) *Nature's Purposes: Analyses of Function and Design in Biology*. MIT Press, Cambridge.

Ariew, A. and Perlman, M. (2002) Introduction. In A. Ariew, R. Cummins, and M. Perlman, eds., *Functions: New Essays in the Philosophy of Psychology and Biology* (pp. 1–7). Oxford University Press, New York.

Arp, R. (2006) Evolution and two popular proposals for the definition of function. *Journal for General Philosophy of Science*, **37**, 2–12.

Dipert, R. (1993) *Artifacts, Art Works, and Agency*. Temple University Press, Philadelphia.

Grenon, P. and Smith, B. (2004) SNAP and SPAN: Towards dynamic spatial ontology. *Spatial Cognition and Computation*, **4**, 99–104.

Hill, D., Smith, B., McAndrews-Hill, M., and Blake, J. (2008) Gene Ontology annotations: What they mean and where they come from. *BMC Bioinformatics*, **9**, S2.

Hunter, L. (2009) *An Introduction to Molecular Biology for Computer Scientists*. MIT Press, Cambridge, MA. In preparation.

Jansen, L. (2007) Tendencies and other realizables in medical information sciences. Available at: http://ontology.buffalo.edu/bfo/Tendencies.pdf.

Lind, M. (1994) Modeling goals and functions of complex industrial plants. *Applied Artificial Intelligence*, **8**, 259–283.

Perlman, M. (2004) The modern philosophical resurrection of teleology. *The Monist*, **87**, 3–51.

Rosse, C. and Mejino, J. (2003) A reference ontology for bioinformatics: The Foundational Model of Anatomy. *Journal of Biomedical Informatics*, **36**, 478–500.

Smith, B. and Grenon, P. (2004) The cornucopia of formal-ontological relations. *Dialectica*, **58**, 279–296.

Smith, B., Kusnierczyk, W., Schober, D., and Ceusters, W. (2006) Towards a reference terminology for ontology research and development in the biomedical domain. *Proceedings of KR-MED*, **1**, 1–10.

Smith et al. (2007) The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, **25**, 1251–1255.