

BOWiki - a collaborative annotation and ontology curation framework

Michael Backhaus
Max Plank Institute for
Evolutionary Anthropology
Deutscher Platz 6
04103 Leipzig, Germany
michael_backhaus@eva.mpg.de

Janet Kelso
Max Plank Institute for
Evolutionary Anthropology
Deutscher Platz 6
04103 Leipzig, Germany
kelso@eva.mpg.de

ABSTRACT

As the amount of data being generated in biology has increased, a major challenge has been how to store and represent this data in a way that makes it easily accessible to researchers from diverse domains. Understanding the relationship between genotype and phenotype is a major focus of biological research. Various approaches to providing the link between genes and their functions have been undertaken – most require significant and dedicated manual curation.

Advances in web technologies make possible an alternative route for the construction of such knowledge bases - large-scale community collaboration. We describe here a system, the BOWiki, for the collaborative annotation of gene information. We argue that a semantic wiki provides the functionality required for this project since this can capitalize on the existing representations in biological ontologies. We describe our implementation and show how formal ontologies could be used to increase the usability of the software through type-checking and automatic reasoning.

Categories and Subject Descriptors

H.3.5 [On-line Information Services]: Web-based services; H.5.3 [Group and Organization Interfaces]: Collaborative computing; K.4.3 [Organizational Impacts]: Computer-supported collaborative work; I.2.4 [Knowledge Representation Formalisms and Methods]: Representations

General Terms

Ontology of Functions, GFO-Bio

Keywords

semantic wiki, ontology curation

1. INTRODUCTION

There has been recent interest in the development of a public knowledge base for information about genes and gene functions [11]. Biomedical ontologies such as the *Gene Ontology* (GO) [1] provide terms that have been used to annotate genes with information about their function and cellular location. The *Ontology of Functions* (OF) [2] provides an additional function description layer that can be used

Copyright is held by the author/owner(s).
WWW2007, May 8–12, 2007, Banff, Canada.

to enrich the description of functions in biomedical ontologies. Applying OF introduces new, more specific relations such as *hasFunction* and *isRealization* for the annotation of gene products. This framework allows for the preservation of the original ontological structure while providing a more detailed description of the function and the genes involved.

To date the majority of gene annotation to biomedical ontologies has been performed by expert curators who use the public literature to determine the appropriate ontological terms for the annotation of genes and gene products. This method provides a highly accurate annotation, but is manually intensive and time-consuming. To overcome this bottleneck, automatic electronic annotation was also implemented, but this is generally less reliable than manual curation. The ability to tap the expertise of the large biological community potentially provides a stable, long-term strategy for maintaining and extending the information captured about gene functions. To this end we have developed the BOWiki, which provides a vehicle for the collaborative annotation of genes with concepts in the biomedical ontologies, as well as a suitable framework for the detailed description of functions as specified by OF.

The aim of the BOWiki is to capitalize on community knowledge to build an accurate and easy-to-use knowledge base about genes and gene functions [5]. Since the target audience of the BOWiki is largely made up of molecular biologists with little experience of ontologies or formal logics, the primary design principle was useability. To this end both data entry and retrieval make extensive use of a simple syntax, graphical representations, and web forms where appropriate. Additionally, BOWiki will provide automated consistency checking, which uses a type system based on a biological core ontology and an upper level ontology in order to ensure the consistency of the knowledge base during editing.

We discuss here in some detail how information capture, storage, searching and quality assurance are managed inside the BOWiki.

2. INFORMATION CAPTURE

2.1 Wiki

We evaluated the suitability of various technologies to serve as a framework for curating biomedical knowledge bases. Due to our need for facilitating intuitive collaboration editing, we chose a wiki as the base system for BOWiki [6]. The *MediaWiki* software is used by *Wikipedia*, probably making

Function Name:	Sugar_transporter_activity	Load
Functional Item:	[[Sugar_transporter]]	
Goal:	Sugar located_in B	
Requirements:	Sugar present AND Sugar located_in A	
	Store	Update

Figure 1: The Function Edit Form together with an example function, "Sugar transporter activity" and the corresponding function determinants, is shown.

it the most well-known wiki software. MediaWiki further supports a well-documented system for creating extensions, making it easy for developers to customize a wiki to meet their specific needs. The BOWiki is based on the *Semantic MediaWiki* [10] v0.4, an extension of the MediaWiki that employs binary relations and attributes, together with a formal semantics for their use. We extended the Semantic MediaWiki to satisfy to our requirements, and describe these extensions in the following sections.

2.2 Function Edit Form

The ability to edit biological functions in the BOWiki is realized through an implementation of the framework provided by OF. The OF treats functions as *goal-oriented* entities, and provides a formal way to represent functions in a *function structure* scheme. The scheme consists of a set of *labels*¹, a set of *requirements*², a *goal*³ and a *functional item*⁴, which together form the *function determinants*. Via the *function edit form* (see figure 1), the function determinants are presented for editing. When the information concerning the function determinants is stored, concepts with function-, requirement- and goal-name are created or modified, together with the appropriate semantic links and structure. The complete content of the function determinants, as well as a link to the function edit form for further editing, is presented inside an information box on the function page.

2.3 N-ary relations

The OF framework requires the use of *n*-ary relations. In order to use OF and to allow for intuitive knowledge acquisition, the BOWiki extends the Semantic MediaWiki to allow *n*-ary relations between wiki pages. In order to enable the use of *n*-ary relations, the BOWiki structures arguments of relations into *roles*, or named argument slots. This allows *n* arguments to be used in an arbitrary order. We believe that this provides an intuitive way to describe complex relations between entities in the BOWiki. This feature is not present in most semantic wikis. For the semantics of *n*-ary relations in the BOWiki, we use the proposal from the W3C Semantic Web Working Group[8].

¹labels are used as expressions which name a function

²requirements are necessary preconditions

³a goal specifies the part of the world directly affected by the function

⁴the functional item describes the role played by a certain entity in any realization of the function

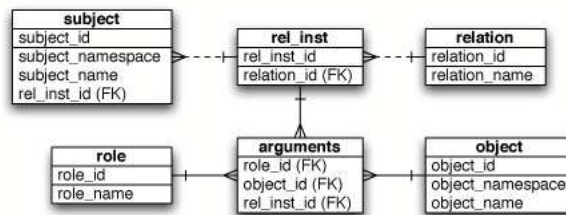


Figure 2: The figure above describes the general database layout used to store semantic data in BOWiki.

An example of a three place relation in BOWiki is the following function ascription (assumed to appear on the *MAL21 wiki page* describing the sugar transporting protein *MAL21*):

```

[[realizes::function=sugar transporter activity;
realization=sugar transport;
context=human body]]

```

This means that the protein MAL21 realizes the function "sugar transporter activity" (taken from the GO) by means of a process "sugar transport" (taken from the GO) in the context of a human body.

2.4 Database layout

In order to store the semantic content of the BOWiki, we extended the Semantic MediaWiki's database model. The database model must accommodate *n*-ary relations, roles and types from BOWiki's type system (see section 4.1). Therefore, a database layout that differs from the triple-representation used in the Semantic MediaWiki [10] is required. In the BOWiki, we introduce separate tables for storing subjects, relations, roles, types and objects.

The database model, as shown in figure 2, is subdivided into two parts, one focusing on classes (relations, roles and types), the other on instances (instances of relations, of roles and objects). Relations are instantiated each time they appear with a different set of arguments. These instances acquire an identifier from the *relation instance* table. The arguments table is comprised of the identifiers of the roles and objects used in a relation-instance. It is, therefore, a table of instances of roles. Using the relation-instance and argument table, the database layout allows storage of *n*-ary relations, and most operations can be completed by integer queries, which ensures rapid content retrieval from a large knowledge base.

For the incorporation of automated consistency checks, a table with available types will be added to the current database layout, and the types will be linked to the roles that may be used in a relation. This allows for the classification of objects occurring as arguments to relations in the wiki.

3. DATA ACCESS

3.1 Semantic Search

Building a structured knowledge base with semantic annotations allows for improving regular search techniques through content based queries. In particular, in the context of gene annotations, it is important to be able to find correlations between genes, i.e., to see which gene has a specific relation

Subject:	Relation:	Object:	Role:	
		sugar		
		activity	realization	Search Relations add row
		Attribute name:	Attribute value:	
				Search Attributes

Search results (relations)

1. MAL21 :: Realizes :: Function=Sugar transporter activity, Realization=Sugar transport

Figure 3: This figure shows a search using the semantic search tuple with the parameters **object.1=sugar**, **object.2=activity**, **role.2=realization**, and the result showing the facts about MAL21 fulfilling this search.

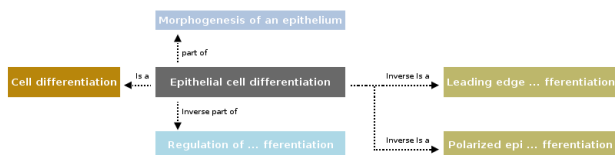


Figure 4: Shown above is the semantic graph of the concept *Epithelial cell differentiation* from the Gene Ontology, which is generated on the special pages for the external ontologies.

to another, or which gene plays a certain role in a specific relation. In order to allow for content-based searches within the BOWiki, the *Semantic Search Triple* from the Semantic MediaWiki v0.4 has been expanded into a *Search Tuple* (see figure 3). This facilitates searching for an arbitrary number of terms of the following types: *subject*, *relation*, *object*, *role* and *attributes*. The search function employs numerical identifiers in order to increase its efficiency. This is in particular significant when querying large databases, such as the GO, which contains more than 50,000 relations between concepts.

Inline Queries, as introduced by the Semantic MediaWiki, are adapted to the more complex representation schema of semantic information within BOWiki. Inline queries allow for the performance of semantic searches within wiki pages, representing the results as part of the rendered wiki page.

3.2 External Ontologies

For referencing, browsing and as a starting point for local articles, the BOWiki allows access to several ontologies like the Gene Ontology [1] and ontologies about cell types or anatomy. These ontologies are accessible via special pages, which prohibit changing the ontology’s concepts in the external ontology directly, but permit editing or updating the local article about this concept. In this way, it is not possible to change the external ontologies, but a user may modify the information contained in these ontologies for use in an extended description of these concepts within the BOWiki. The ontologies are maintained separately from the BOWiki’s semantic storage, and are updated regularly to current versions. It is possible to browse these ontologies graphically, represented as a directed acyclic graph. This graph is automatically generated from the relations that hold between the concepts in these ontologies, by extracting these informations and using the MediaWiki plugin Graph::Easy⁵ for visualization.

⁵<http://bloodgate.com/perl/graph/>

4. QUALITY & CONSISTENCY

4.1 Type System

Collaborative curation of a knowledge base may quickly lead to semantic inconsistencies. A simple example of such an inconsistency is the use of a *part of* relation between a process and an object, which is often prohibited[3]. Another example may be the classification of some entity as both an object and a process, leading to an inconsistency when objects and processes are considered disjoint. In order to ensure the quality of the wiki’s content, at least some of these possible inconsistencies must be identified and prevented. Since the wiki contains descriptions of biological concepts and entities, we use an upper biological ontology as a type system.

4.1.1 Core ontology

A core ontology provides definitions for the most basic concepts within a domain. We developed GFO-Bio [4] as a core ontology for biology. GFO-Bio extends the General Formal Ontology (GFO) [3] by adding definitions for biological concepts and relations. GFO-Bio is formalized in the decidable fragment of the Web Ontology Language (OWL) [7]. It contains definitions for concepts such as *Cell*, *Biological process* and *Protein*.

GFO-Bio incorporates higher order categories and employs an explicit instantiation relation, which are particularly relevant for its efficient use in a collaborative knowledge acquisition environment, where both concepts and individuals are described formally. Higher order categories allow biological concepts to be treated as instances instead of classes in OWL. We believe that this will also result in a more rapid and efficient classification.

4.1.2 Reasoning

All BOWiki pages, whether concepts or individuals, are instances of a class in GFO-Bio; we use instantiation instead of subsumption to increase the efficiency in reasoning. In some sense, reasoning is restricted to the core ontology because all wiki pages appear as OWL individuals for the reasoner. Consequently, subsumption links among concepts in the wiki are currently not exploited, which is appropriate for basic type checking and a tribute to the performance of current reasoners. Two types of inconsistencies may arise that can be avoided.

First, consider the BOWiki concept *Hydrogen* and the GFO-Bio concepts *Molecule* and *Atom*. In the BOWiki, *Hydrogen* may become an instance of *Molecule*. After instantiation, we want to prevent a user from making the *Hydrogen* also an instance of GFO-Bio’s *Atom* concept, because those concepts are disjoint in GFO-Bio. A molecule consist of atoms, but it is not an atom.

Second, concepts taken from GFO-Bio may serve as type-restrictions for roles in n-ary relations. Consider the realizes relation and its roles: the implicit subject role, the function, realization and context roles. These are restricted to the GFO-Bio concepts: Biological object, Biological function, Biological process and Entity. The use of a particular concept "A" to fill the realization role of the realized relation causes this concept to be classified as a Biological process. If the same concept "A" is used in the same relation, but in the subject role, it is then classified as both Biological Process and Biological Object, resulting in an inconsistency because

in GFO-Bio, Biological Processes and Biological Objects are disjoint⁶.

Both types of inconsistencies are automatically detected within the BOWiki using the description logic reasoner Pellet⁷. The description logic reasoner Pellet⁷ is capable of performing these operations, and will be used for automated consistency checking in the next version of the BOWiki.

4.2 Quality rating

To further improve the quality of the articles, we are considering the inclusion of a content evaluation system, which may take the form of a direct evaluation of an article, or a cumulative score for an author's body of work. Further research is needed in order to represent these ratings in the semantics of BOWiki's content.

5. OUTLOOK

The BOWiki is still under development. The main task for the future will therefore be to enhance the user interface in order to make the BOWiki more user-friendly and intuitive.

We are considering integrating the BOWiki with a grid platform such as the MediGRID⁸, a portal for applications and web-services for biomedical research and the life sciences. Using web-services, the BOWiki may access biomedical ontologies directly rather than relying on daily updates. Web-services also provide the ability for users to perform queries of the BOWiki content within an integrated web-service platform.

For more restricted domains, a more expressive syntax may be required. For example, logical conjunction, negation or existential and universal quantification may be allowed. This may make it possible to extend the type system itself, instead of classifying every article as an instance of a concept in the type system. This will facilitate the collaborative construction of expressive ontologies. However, the main limitation will be the efficiency of the available reasoners, as well as the integration of a syntax for these constructs in the BOWiki.

To provide a graphical representation of the structured content, we are considering conceptual graphs [9]. Conceptual graphs may provide a more intuitive representation schema for logical formulae than their representation in text format.

6. CONCLUSION

The BOWiki can be used to collaboratively create a knowledge base about biological domains. It allows for the specification and description of functions according to the framework laid out by the Ontology of Functions [2]. By employing the core ontology GFO-Bio as a type system, the BOWiki facilitates the creation of a structured knowledge base while preserving consistency. It further extends another semantic wiki, the Semantic MediaWiki, for a customized handling of n -ary relations, making knowledge acquisition more intuitive. Within the biological and biomedical community, the BOWiki provides a framework for constructing a high-quality, large-scale knowledge base to be

⁶It is possible to discuss articles and the restrictions of relations in the BOWiki on a discussion page available on each concept page.

⁷<http://pellet.owldl.com/>

⁸<http://www.medigrid.de/>

used for annotating genomic data, as well as for describing and defining both biological functions and other biological concepts⁹.

7. ADDITIONAL AUTHORS

Joshua Bacher, email: bacher@eva.mpg.de, Heinrich Herre, email: heinrich.herre@imise.uni-leipzig.de, Robert Hoehndorf, email: hoehndorf@eva.mpg.de, Frank Loebe, email: frank.loebe@imise.uni-leipzig.de, Johann Visagie, email: visagie@eva.mpg.de

8. REFERENCES

- [1] M. Ashburner et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29, May 2000.
- [2] P. Burek, R. Hoehndorf, F. Loebe, J. Visagie, H. Herre, and J. Kelso. A top-level ontology of functions and its application in the Open Biomedical Ontologies. *Bioinformatics*, 22(14):e66–e73, July 2006.
- [3] H. Herre, B. Heller, P. Burek, R. Hoehndorf, F. Loebe, and H. Michalek. General Formal Ontology (GFO) – A foundational ontology integrating objects and processes [Version 1.0]. Onto-Med Report 8, Research Group Ontologies in Medicine, Institute of Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, 2006.
- [4] R. Hoehndorf, F. Loebe, R. Poli, J. Kelso, and H. Herre. GFO-Bio: A biological core ontology. a new methodology for integrating ontologies. *Applied Ontology*, 2007. under review.
- [5] 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. H. et al., editors, *OTM Workshops 2006*, number 4277 in LNCS, pages 669–678. Springer-Verlag, 2006.
- [6] B. Leuf and W. Cunningham. *The wiki way: Quick collaboration on the web*. Addison-Wesley, 2001.
- [7] D. L. McGuinness and F. van Harmelen. OWL Web Ontology Language overview. W3C recommendation, World Wide Web Consortium (W3C), February 2004. <http://www.w3.org/TR/2004/REC-owl-features-20040210/>.
- [8] N. Noy and A. Rector. Defining n -ary relations on the semantic web. <http://www.w3.org/TR/swbp-n-aryRelations/>, 2006.
- [9] J. Sowa. *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley, Reading, MA, 1984.
- [10] M. Völkel, M. Krötzsch, D. Vrandečić, H. Haller, and R. Studer. Semantic wikipedia. In *Proceedings of the 15th international conference on World Wide Web*, pages 585–594, 2006.
- [11] K. Wang. Gene-function wiki would let biologists pool worldwide resources. *Nature*, 439(7076):534, Feb 2006.

⁹The interested reader can find a running system using the current stable release of BOWiki at <http://onto.eva.mpg.de/bowiki>