

OBOPedia: An Encyclopaedia of Biology Using OBO Ontologies

Robert Stevens and Adam Nogradi

School of Computer Science, University of Manchester, UK,
robert.stevens@manchester.ac.uk

Abstract. Ontologies contain knowledge about a domain for use by tools or humans. A source of knowledge should be usable by a human to ‘find out about’ or learn about the entities of a domain and their relationship to each other. The corpus of biomedical ontologies now contains ‘encyclopaedic’ knowledge about biology and should be capable of being used by humans to learn about entities in molecular biology. Yet multiple separate ontologies and the typical style of presentation of the knowledge in the ontologies mean that their use as a learning resource is sub-optimal. To address this issue we have created OBOPedia, a web based encyclopaedia of biology as seen by the Open Biomedical Ontologies (OBO) Consortium. OBOPedia exploits the OBO’s use of standard representations and meaningful human readable terms and natural language definitions to create a basic OBOPedia encyclopaedia entry. An entry is supplemented with an ontology’s synonyms and uses the ontology’s taxonomic links to provide ‘see also’ cross-references within the alphabetical list of entries. Currently, OBOPedia has access to ten OBO ontologies, including all the OBO Foundry ontologies, which have a total of over 210,000 entries. Our evaluations indicate that an OBOPedia style of presentation has a role as an alternative way of presenting knowledge of a domain collected as an ontology or ontologies. OBOPedia offers another view into a field of interest that is based on a collection of ontologies as a reference resource, but one in which a user need not know it is an ontology. OBOPedia may be used via <http://www.obopedia.org.uk>. The source code and documentation for OBOPedia are available via <https://bitbucket.org/adam944/ontologyencyclopaedia>.

Keywords: Encyclopaedia, Open Biomedical Ontologies, Natural Language Definitions, User Interface

1 Introduction

We present OBOPedia, which offers the knowledge of biology captured in a corpus of ontologies, as an encyclopaedia whose intended audience are those seeking to learn about biological entities and are not experts in ontologies or wishing to use knowledge in the form of an ontology. An ontology captures what is universally true about a field of interest [15]. Ontologies contain representations of the entities in a domain, their logical definitions, descriptions and how they are related to one another [7]. An ontology may also be decorated with natural language information such as non-logical definitions and

the labels for entities used for display. As such, ontologies should be able to be used effectively as a source of knowledge about a field of interest by humans and machines, by ontology experts and people unaware of ontologies; an ontology talks about ‘things’ in a domain and should be usable by a range of users to find out about ‘things’ in that field of interest without them necessarily being *au fait* with ontologies.

The field of biology has captured a lot of its knowledge in the form of ontologies [4]. BioPortal [14], for example, now has over 400 ontologies covering many sub-domains of biology. The Open Biomedical Ontologies [16] (OBO) has attempted to co-ordinate the development of ontologies from genotype to phenotype and outwards to small molecules, scientific investigations, environments and so on; there are 133¹ ontologies available via the OBO Foundry [16] website².

This collection of biological knowledge has many uses, mainly for annotation of data that facilitates search and retrieval via query generalisation and analysis of data through techniques such as over-expression analysis [17, 9]. Most of this is facilitated by ontologies acting as a shared vocabulary for describing data and then exploiting the structure in which that vocabulary is held.

Despite an ontology being a mechanism for capturing and representing knowledge, ontologies do not appear to be used as a source of knowledge for learning about a domain by humans. The TAMBIS Ontology [3] offered itself as a tutorial for molecular biology and related bioinformatics, but did not realise this function. The widespread development of ontologies across biomedicine means there is a knowledge resource that spans broad parts of the subject; enabling this valuable resource to be used as knowledge to learn about biology should be a useful resource.

OBO ontologies are created in a common representation style, down to what relationships and annotation types are used. All OBO classes should have human readable labels, synonyms for those labels and natural language definitions.

An example of labels and natural language definition for the class GO:0005794 from the Gene Ontology (GO) [2] are shown in Table 1. The structure of the ontology gives more general and more specialised concepts (along with definitions and labels). Here the subsumption and other relationships say what is universally true about this class of objects; in GO each and every golgi apparatus is also a kind of ‘intracellular membrane-bounded organelle’ as well as being a ‘cytoplasmic part’ and each and every golgi apparatus is also part of an ‘endomembrane system’. The natural language definition fragment in Table 1 contains much more ‘background’ knowledge, sometimes knowledge that is not universally true about the entities in the ontology. This broader knowledge may still be useful for someone wishing to explore and understand a field of interest. In this sense, the universal, logical knowledge in an ontology, together with the wider ranging knowledge in the natural language definitions can be said to form ‘background’ knowledge [15].

An encyclopaedia is “a type of reference work holding a comprehensive summary of information from either all branches of knowledge or a particular branch of knowledge” [19]. An encyclopaedia consists of entries, containing concise, factual descrip-

¹ 11 OBO Foundry ontology and 122 candidate ontologies and others of interest (<http://www.obofoundry.org/> on 23 June 2015).

² <http://www.obofoundry.org/>

Table 1. Primary label, synonyms and natural language definitions for the Gene Ontology class GO:0005794

Metadata type	Content
Label	Golgi apparatus
Exact Synonym	Golgi complex
Broad Synonym	Golgi
Narrow Synonym(s)	Golgi ribbon
Related Synonym	
Definition	A compound membranous cytoplasmic organelle of eukaryotic cells, consisting of flattened, ribosome-free vesicles arranged in a more or less regular stack. The Golgi apparatus differs from the endoplasmic reticulum [. . .]

tions about entities. This description of an encyclopaedia entry looks remarkably like an OBO class' descriptive metadata (see Table 1); so a collection of OBO entities and their metadata could be presented as an encyclopaedia of biology. OBOPedia's attempt at such a presentation for the entry in Table 1 is shown in Figure 1.

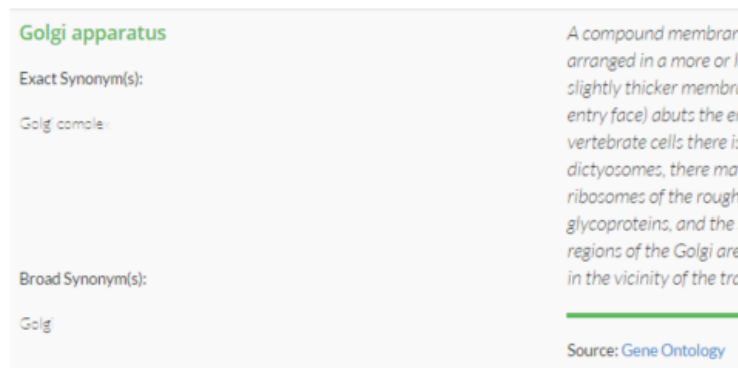


Fig. 1. The (partial) entry 'Golgi apparatus' GO:0005794 from the Gene Ontology, with its description, exact synonyms and source ontology link as shown in OBOPedia

The knowledge captured within the OBO is now encyclopaedic. To capitalise on this resource as a source of background knowledge about biology we present the content of some OBO in the form of an encyclopaedia as OBOPedia (see Figure 1). OBOPedia's value is in the ability to hide away everything that is unnecessary to show in a simple encyclopaedia. Users can browse a corpus of ontologies' knowledge about a field of interest by reading definitions, search for terms and their synonyms, or explore super- and subclasses of terms using the 'see-also' links next to them. OBOPedia is made by using each class' labels, its synonyms, its natural language definition as the encyclopaedia entry text, and the relationships held by the entity as the 'see also' links of an entry.

2 Method

The building of OBOPedia's content is straight forward:

- Select the OBO ontologies that may form OBOPedia's content.
- Remove all obsolete classes.
- Create an encyclopaedia object aggregating the concept's label, synonyms, natural language definition, origin and sub- and super-class relationships (see table 1).
- Index on the labels using Apache Solr³, normalising to the first alphabetic character of the label.
- Display using OBOPedia's user interface.

WE took the OBO Foundry ontologies in their OWL representation from <http://purl.obolibrary.org/obo/xxx.owl> (where 'xxx' is the ontology's short name—GO for the Gene Ontology and ChEBI for the Chemicals of Biological Interest ontology) in May 2015. Processed using the OWL API version 3.5.0 [8]. The list of ontologies used in this version of OBOPedia is:

1. The Gene Ontology (GO)—the Molecular Function, Biological Process and Cellular Component ontologies.
2. The Protein Ontology (PrO).
3. Chemical Entities of Biological Interest ontology (ChEBI).
4. The Human Disease Ontology (HDO).
5. The Human Phenotype Ontology (HPO).
6. The Ontology of Biomedical Investigations (OBI).
7. the Plant Ontology (PO).
8. The Phenotypic quality Ontology (PATO).
9. The Xenopus Anatomy and development Ontology (XAO).
10. The Zebrafish Anatomy and Development Ontology (ZFA).

OBOPedia follows the classic structure of a Web application, implemented in a Model-View-Controller style. It processes data independently from the UI layer and the two ends are connected via servlets acting as controllers. Figure 2 shows an architecture for OBOPedia.

The created encyclopaedia entries can then be read by the front end's JSP mechanisms using getter methods, thus abstracting all the logic away from the UI layer. The application uses Bootstrap⁴ components to make the menus, buttons and text more attractive. Most of the design components are based on Bootstrap templates or adjustments of them. The letter browser on the left hand side of the main entries page uses a part of the MooTools⁵ JavaScript framework to make the menu more interactive.

OBOPedia's evaluation proceeded as follows with evaluator and participant face to face using a think aloud protocol [13] followed by a questionnaire:

1. Introduction and demonstration of OBOPedia.

³ <http://lucene.apache.org/solr/> accessed on 15 July 2015

⁴ <http://getbootstrap.com> (visited on 25 April 2015)

⁵ <http://mootools.net/> (visited on 25 April 2015)

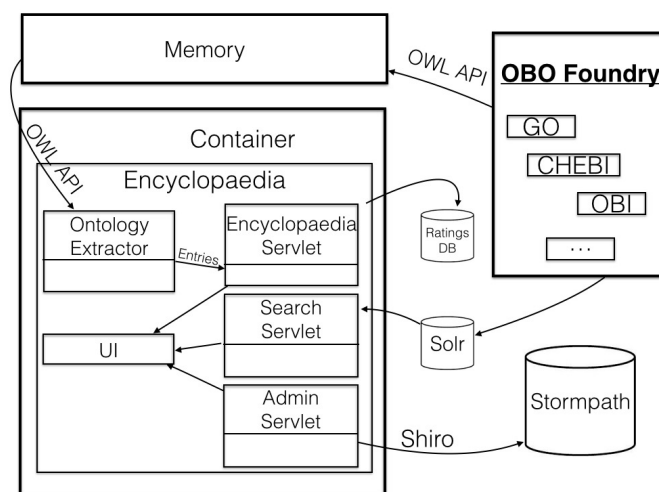


Fig. 2. The basic architecture for OBOPedia

2. Participant then uses OBOPedia, exploring topics of his or her own choosing.
3. Finally, a short questionnaire was filled out, in which the participant rated the application, stated their background and provided any type of feedback.

The evaluation had ethical approval and the Personal Information sheet and consent form are available on the project's repository.

3 Results

Figure 1 shows an entry in OBOPedia from the Gene Ontology. A typical encyclopaedia presentation is used, with the ontology supplying the main entry wording, synonyms, the natural language definition and the cross-links. The source of the ontology entry is also shown with the ability to link through to the ontology's home pages.

OBOPedia offers the following functionality:

- It displays an alphabetically ordered list of encyclopaedia style entries for OBO showing their labels, synonyms and natural language definitions.
- 'See also' cross-references are provided by sub- and super-class relationships, enabling navigation to more general and more specific entries.
- Entries are searchable by their primary labels.
- An administration panel allows ontologies in OBOPedia's library of ontologies to be made accessible to the user interface.
- Each OBOPedia entry has an associated 'rate this definition' widget. This allows readers to give a simple numeric score for how much the definition is 'liked'. In the fullness of time this is intended to provide feedback to ontology authors. As yet, insufficient data has been collected for anything to be concluded.

The main page of the OBOPedia user interface has an alphabetic index and a search box to give access to the content of OBOPedia. Many entries are put on the screen for context and to promote ‘serendipitous discovery’ of interesting topics, as is sometimes the case when looking up topics in a traditional paper based encyclopaedia or dictionary [6].

The current version of OBOPedia has available the ontologies listed in Section 2, giving OBOPedia over 210,000 entries from ten ontologies (counting GO as one ontology). At present OBOPedia’s administration section is rudimentary. An administrator can choose whether or not to make available to the user any number of ontologies in OBOPedia’s library of ontologies. The ability to add new ontologies and to update the versions of the ontologies in OBOPedia’s library remain to be done. The administration panel also shows some basic metrics for OBOPedia, such as the number of ontologies available at the front end, the number of entries, synonyms per entry, and so on.

3.1 OBOPedia evaluation

OBOPedia was evaluated as described in Section 2. The ‘test drive’ of OBOPedia took approximately 30 minutes. There were 13 participants, 4 biologists and 9 computer scientists. The quantitative question ‘Rate the application overall 1–10’, gave the mean score of 8.9; median of 9 and the mode of 9.

Some quotes from evaluators of OBOPedia are:

- ‘The encyclopaedia was very straightforward to use. The search function was simple and presented the results in a way which made choosing the most appropriate entry very easy.’
- ‘As someone with a biology background, having all the biology information sources in one place instead of searching many different databases is very useful.’
- ‘A very useful tool for searching ontological terms. I wanted to have this before; when I was searching for the terms for my annotation scheme.’
- ‘Very well organised and easy to navigate.’
- ‘The search bar could have an autocomplete feature to ease finding entities.’
- ‘The application could use the inferred class hierarchy rather than the asserted one, thus making the set of contents broader.’
- ‘I’d prefer the letter browser bar on the right hand side.’

These demonstrate an appreciation of what OBOPedia is attempting and the benefit of gathering the breadth of knowledge in the OBO into one view. Most of the quotes here are about technical and usability enhancements. Most of them are relatively minor, indicating a reasonably well polished, if early, version of the software. The addition of reasoning support to broaden the querying capabilities, and thus the retrievals of views across the content, is a more major change.

Users appeared to appreciate the look and feel of the application, ensuring that no important changes have to be made on the user interface layout. They found it straightforward to understand, browse and search, with minimal amount of training provided prior to the test.

Users with a biology background particularly liked the idea of having data from multiple areas of biology present under one site, thus saving them the effort of searching for a term in Google then ending up on multiple different sites for the different subdomains.

A list of enhancements to OBOPedia to come out of the evaluation are:

- To provide some simple graphic of the ‘see also’ context to show what is at the other end of the cross-references.
- To use the other ontological relationships in the ‘see also’ cross-references.
- Using an automated reasoner behind OBOPedia would afford the ability to construct query driven views of content beyond that of the alphabetical or search based views.
- To enable ‘roll over’ from one letter’s entries to another, so when coming to the end of the ‘A’ entries, the content just scrolls straight to the start of the ‘B’ entries.
- Enable searching over not just the primary label, but also the synonyms and natural language definition. Searching would also be enhanced by auto-complete within the search field to avoid having to type in sometimes lengthy and difficult to spell names.
- Adding hyperlinks to OBOPedia entries from within the natural language definitions.
- Add a list of the currently available ontologies to the front page so users know what is available to explore.
- Add an ‘entry of the day’—a pseudo-randomly selected entry is put on the front page as a ‘taster’ and potential serendipitous start point.

Overall, the evaluation was positive, both from an exploration of OBOPedia’s usability and of its aim as a reference point for a broad collection of biology’s entities.

4 Discussion

Ontologies capture knowledge about a domain or field of interest. This knowledge is used by humans and computers in a variety of tasks. One of those tasks should be to ‘find out’ about the field of interest of the ontology or group of ontologies without being an ontologist. Biology, for example, has developed a wide range of ontologies covering the domain from genotype to phenotype and then beyond [4, 16].

Most presentations and uses of ontologies are for people knowingly using ontologies—presentation as either some sort of graph [12] or search that helps to find classes or ontology [5, 10, 14] and with the task of finding the appropriate ontology or the appropriate term for tasks such as annotation of data [10, 17]. The task of finding out about domain knowledge for humans to read and learn appears to be neglected.

To address the task of finding out about ‘stuff’ in biology, OBOPedia exploits the situation that:

- The OBO cover a broad range of biomedical knowledge that will provide enough content for the encyclopaedia.
- The OBO are available in the same format and importantly having common annotation properties and types of annotation.

- The OBO are intended to be clearly delineated, so each entry should be unique (this is not always true; for instance, anatomical entries will be duplicated, but for different organisms). This means OBOPedia entries should not be repeated.
- The OBO principles demand each class in the ontology has a natural language definition; these are mostly orientated to domain users rather than ontologists, which should increase the accessibility of OBOPedia to its intended audience.
- The rich synonyms within OBO ontologies offer the potential for many avenues of access for users of OBOPedia.

Many of these features capitalise on some of the OBO Foundry principles⁶. OBOPedia is, therefore, restricted to OBO and there is no intention for it to use ontologies outside the OBO regime.

What OBOPedia seeks to do is present the collected knowledge of OBO's ontologies as a catalogue of biology's entities in an encyclopaedic form for non-ontology users. It has been evaluated by two kinds of users. A technical evaluation was conducted by people in the domain of computer science; this elicited general usability issues. Biologists were also used to evaluate OBOPedia. As well as usability issues, this group, as the intended primary users, found the knowledge afforded by OBOPedia a useful way of finding out about entities in biomedicine. The broader claims of OBOPedia have not been fully evaluated, but feedback so far suggests the approach is feasible.

There are other ways of accessing the content of a corpus of bio-ontologies [1, 5, 18]. They all appear to be aimed at an audience seeking to find the ontology or its content, rather than learn about the knowledge in the ontology as a human seeking domain knowledge. Ontobee seems to be closest to what OBOPedia seeks to support; it has a page for a term from an ontology, highlighting the label, synonyms and the natural language definition. It also has a large quantity of ontological information which is something that OBOPedia avoids. BioPortal [14, 18] has a repository of over 400 biomedical ontologies. It offers a range of facilities to developers of ontologies and to those that use ontologies as ontologies. The type of presentation of ontology knowledge and role of OBOPedia is not currently part of BioPortal's remit. Nevertheless, what OBOPedia could be implemented using BioPortal's web services [18] and offered as part of BioPortal's web presence—as long as OBOPedia's 'non-ontology' view of knowledge captured in an ontology is part of BioPortal's web presence's role. OBOPedia does in no way negate the role of resources such as Ontobee and BioPortal; they have different roles and different audiences.

OBOPedia is intended for use by biologists or users wishing to find out about biology and not to find out about the ontology itself. Some users will wish to find out about the biology content of an ontology, but not in an ontological presentation: What an entity is and how it is known to be related to another entity. This is what an ontology does, but just putting an ontology or collection of ontologies 'as an ontology' in front of a biologist or someone wishing to find out about biology is not necessarily sufficient for them to find out about the knowledge content in an ontology. OBOPedia is an attempt to look sideways at the problem of giving access to the considerable effort that has been made to collect what is known about biology into an easily accessible form. We can see

⁶ <http://wiki.obofoundry.org/wiki/index.php/Category:Principles>

that an encyclopaedic form can do this job. As a community we should think of how we can bring the semantic view of biology to as wide an audience as possible. Applications like kupkb [11] do this—using an ontology without knowing one is doing it, but offering working biologists a useful thing that is facilitated by ontology.

Acknowledgments

We acknowledge the input of the participants in OBOPedia’s evaluation.

References

1. Xiang, Z., Mungall, C., Ruttenberg, A., He, Y.: Ontobee: A Linked Data Server and Browser for Ontology Terms. In: Proceedings of the 2nd International Conference on Biomedical Ontology (ICBO). pp. 279–281 (2011)
2. 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., Sherlock, G.: Gene Ontology: tool for the unification of biology. *Nature Genetics* 25(1), 25 – 29 (2000)
3. Baker, P., Goble, C., Bechhofer, S., Paton, N., Stevens, R., Brass, A.: An Ontology for Bioinformatics Applications. *Bioinformatics* 15(6), 510–520 (1999), <http://www.cs.man.ac.uk/~stevensr/papers/bioinformatics-ontology99.doc>
4. Bodenreider, O., Stevens, R.: Bio-Ontologies: Current Trends and Future Directions. *Briefings in Bioinformatics* 7(3), 256–274 (2006), <http://bib.oxfordjournals.org/content/7/3/256.full>
5. Côté, R., Reisinger, F., Martens, L., Barsnes, H., Vizcaino, J.A., Hermjakob, H.: The ontology lookup service: bigger and better. *Nucleic Acids Research* 38(suppl 2), W155–W160 (2010), http://nar.oxfordjournals.org/content/38/suppl_2/W155.abstract
6. Foster, A., Ford, N.: Serendipity and information seeking: an empirical study. *Journal of Documentation* 59(3), 321–340 (2003), <http://dx.doi.org/10.1108/00220410310472518>
7. Guarino, N.: Formal Ontology and Information Systems. In: Proceedings of FOIS98., pp. 3–15 (1998)
8. Horridge, M., Bechhofer, S.: The OWL API: A Java API for Working with OWL 2 Ontologies. In: OWLED 2009, 6th OWL Experienced and Directions Workshop. Chantilly, Virginia (October 2009)
9. Jensen, L., Bork, P.: Ontologies in Quantitative Biology: A Basis for Comparison, Integration, and Discovery. *PLoS Biology* 8(5), e1000374 (2010)
10. Jonquet, C., Musen, M., NH, S.: Building a biomedical ontology recommender web service. *Journal of Biomedical Semantics* 1(S1) (2010)
11. Jupp, S., Klein, J., Schanstra, J., Stevens, R.: Developing a Kidney and Urinary Pathway Knowledge Base. *Journal of Biomedical Semantics* 2(Suppl. 2), S7 (2011), <http://www.jbiomedsem.com/content/2/S2/S7>
12. Katifori, A., Halatsis, C., Lepouras, G., Vassilakis, C., Giannopoulou, E.: Ontology visualization methods a survey. *ACM Computing Surveys (CSUR)* 39(4), 10 (2007)
13. Lewis, C.H.: Using the “Thinking Aloud” Method In Cognitive Interface Design(Technical report). Tech. Rep. RC-9265., IBM. (1982)

14. Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D.L., Storey, M.A., Chute, C.G., Musen, M.A.: Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research* (2009), <http://nar.oxfordjournals.org/content/early/2009/05/29/nar.gkp440.abstract>
15. Rector, A.: Title of the Deliverable: Barriers, approaches and research priorities for integrating biomedical ontologies (2008), deliverable D6.1 for the Semantic Health project.
16. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., et al.: The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology* 25(11), 1251–1255 (2007)
17. Stevens, R., Lord, P.: Application of ontologies in bioinformatics. In: Staab, S., Studer, R. (eds.) *Handbook on Ontologies in Information Systems*, pp. 347–485. *International Handbook on Information Systems*, Springer, second edn. (2008)
18. Whetzel, P.L., Noy, N.F., Shah, N.H., Alexander, P.R., Nyulas, C., Tudorache, T., Musen, M.A.: Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic acids research* 39(suppl 2), W541–W545 (2011)
19. Wikipedia: *Encyclopedia* (2015), <http://en.wikipedia.org/wiki/Encyclopedia>, (visited on 8 April 2015)