

Towards a Top-Domain Ontology for Linking Biomedical Ontologies

Holger Stenzhorn^{a,b}, Elena Beißwanger^c, Stefan Schulz^a

^a Department of Medical Informatics, Freiburg University Hospital, Germany

^b Institute for Formal Ontology and Medical Information Science (IFOMIS), Saarbrücken, Germany

^c Jena University Language and Information Engineering (JULIE) Lab, Germany

Abstract

In this paper we present the ongoing development and extension work on BioTop – a top-domain ontology for linking biomedical domain ontologies. We start by making the case for the application of a common ontology to interface independent biomedical domain ontologies by introducing a set of more general classes. Then we briefly depict the relation of BioTop to the GENIA ontology as starting point of its initial development. Afterwards we propose our distinction of ontologies into top, top-domain and domain ones and describe our approach to the integration of the top ontology BFO into BioTop. Then we present our plans to join the OBO and OBO Foundry repository of ontologies and list its admission principles in relation to our ontology. Some actual BioTop interface classes are shown subsequently. We conclude by detailing on some planned BioTop usages in the area of BioNLP and cancer research and show some further intended improvements.

Keywords:

Biomedical Ontologies

Introduction

The last couple of years have seen a tremendous increase in the amount of data collected within the life sciences, especially in biomedicine and its subfield of genomics research. This in turn has spurred many scientific efforts to analyze and structure the newly gained data and to extract further knowledge from it. In the following we are now focusing on the application of ontologies for this particular task and specifically examine one currently existing drawback: Most existing biomedical ontologies – even when having overlapping content – are developed mostly independently from each other. Also, each ontology embraces only some distinct scenario with a mere

partial view of the overall scientific field. What has therefore been missing so far is an overarching resource to help with linking and interfacing those independent ontologies. With such a facility in place, new methodologies could be conceived to employ ontologies more efficiently in concert and to create synergetic effects. To this end, we have developed the top-domain ontology BioTop, to be presented in the following.

We concentrated our work on interfacing a smaller selection of about 60 ontologies in the Open Biomedical Ontologies (OBO) framework [1]: the Gene Ontology (GO) [2], the Sequence Ontology (SO), the Cell Ontology (CO), the Chemical Entities of Biological Interest (ChEBI) and the Foundational Model of Anatomy (FMA) [3]. At this point we want to stress that the methodology of our work can be easily applied to create additional interfaces to other ontologies in both this particular topic area as well as similar ones. We are highly interested to further investigate the latter in our future research.

Project Background

Relation to GENIA

The initial version of BioTop rested upon the idea to create a comprehensive, formally-based redesign and expansion of the original GENIA ontology [4]. The basic development policy was to follow the fundamental principles of formal rigor, explicitness and precision of ontological axioms and to maintain the clear overall scope of creating a biomedical upper ontology. The implementation was to be based on the Description Logic subtype of the Web Ontology Language (OWL-DL) [5].

The GENIA ontology had originally been developed for and within the biological natural language processing (BioNLP) community and had quickly become a de-facto standard in this field. Its authors claim the ontology to be a formal model of cell signalling reactions in humans and regard its main applica-

tion to serve as basis for creating thesauri and semantic dictionaries in BioNLP applications (e.g. the semantic annotation of named entities in biological literature abstracts). The GENIA authors also consider it as providing a mutual basis for an integrated view over multiple biological databases.

The GENIA ontology itself is very small, containing only 45 distinct terms, arranged in a simple taxonomy with a maximal depth of six levels. It also limits itself to a set of highly general upper-level classes centred on the notions of biochemical substances and their corresponding locations in the organisms.

During our work, we found some non-trivial shortcomings within the GENIA ontology: Firstly, for most classes a proper documentation and/or textual definition was missing or lacked clarity. Secondly, some class names or their particular position in the taxonomy contradicted common biological or ontological intuitions of consistency. Both issues can obviously lead to conflicting interpretations and incorrect applications of the classes. A complete analysis of the deficiencies found in GENIA and our proposed solutions can be found in [6].

Top, Top-Domain and Domain Ontologies

We propose to distinguish ontologies into three basic types (with their approximate size proportions shown in Figure 1):

- A top ontology (also called top-level or upper ontology) contains only a very small and restricted set of the most high-level, general classes such as “Continuant”, “Occurrent”, “Function” or “Object” – together with some accompanying relations. Examples for this kind of ontologies are BFO [7] and DOLCE [8].
- A top-domain ontology (also called upper-domain (level) ontology) holds the essential core domain classes to interface to both upper and domain ontologies, like “Organism”, “Tissue” or “Cell” in the case of biology. A top-domain ontology can also include more specific relations and further expand or restrict the applicability of relations introduced by the top ontology. An example for this kind of ontologies is BioTop.
- A domain ontology has as its members a multitude of low-level, domain-specific classes to comprehensively describe a certain (aspect of a) domain of interest, e.g. “Antisense RNA Transcription” or “DNA Replication” from the Gene Ontology.

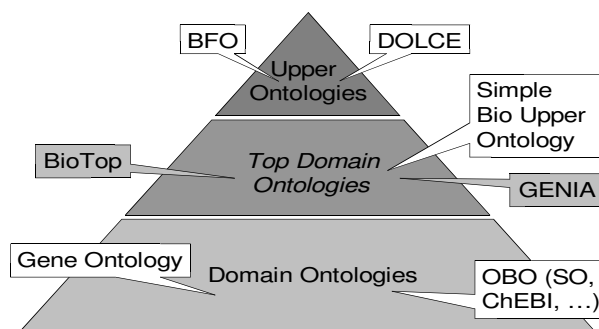


Figure 1 - Ontology Layer Pyramid
(adapted from Alan Rector)

Issues and Goals for the Second Development Phase

When we published the initial version of BioTop we received valuable comments that both encouraged us to continue our work on BioTop but also pointed out some existing shortcomings from the standpoint of biomedical and ontology experts. We also received an open invitation to apply for a membership in the OBO Consortium and (subsequently) in the OBO Foundry. Therefore we set up the following list of basic goals we wanted to achieve in the further development:

- Full inclusion and adoption of the Basic Formal Ontology (BFO) and the OBO Relation Ontology (RO) [9].
- Joining the OBO Consortium as well as the OBO Foundry by fulfilling its given set of principles.
- Improving and expanding the existing interface classes to further link the OBO ontologies.

Methods

Integration of Top Ontologies

Importing BFO and the OBO Relation Ontology (RO)

In the initial version of BioTop we created a set of top-level classes which was based on the terms and definitions found in the publications on both BFO as well as DOLCE (i.e. we initially had a mixture of both top-level ontologies). We also created some additional relations that we based on the ones published in the Relation Ontology (RO).

When the first official version of BFO and RO became available in OWL-DL we decided to modify our ontology by removing the mentioned mixture of BFO and DOLCE top-level classes and to employ the OWL import mechanism to include the now available BFO and RO versions in our ontology.

This integration task was obviously straightforward where actual BFO classes were used before: Here we only had to replace our self-defined class as parent of a given child class with the corresponding class from the BFO import. More problematic were the cases in which DOLCE-inspired classes had been used. Here we had to find either a class in BFO that matched or closely resembled the respective DOLCE class or

we had to remodel the class by introducing one or more mediator classes that were deemed to be ontologically sound.

The inclusion of the RO was also straightforward. We had basically only used the relations already defined by the RO and hence simply had to change the references in our ontology from our relation definitions to the imported ones. For the two additional relations we had defined ourselves, we only needed to change the reference to their parent property since they were directly based (i.e. sub-properties) on ones from the RO.

Because of ongoing discussions to consider the addition of new relations to the RO, the need for keeping our self-defined relations might disappear and hence we could remove them. In case the new relations in the RO do not meet our needs we plan to suggest the addition of our relations to the original RO.

Joining the OBO Consortium and the OBO Foundry

After receiving the above mentioned open invitation to join the OBO Consortium we decided to complete the following steps:

- Select the most important published ontologies contained within the OBO (in regard to their user base size and direct relevance to biomedical research), analyze and compare their current top-level classes with the existing interface classes of the initial BioTop version. Then detect the potential overlap in their respective scope with the scope of BioTop.
- Actively contact the curators and developers of each ontology for which we had developed interface classes and inform them of our plans to join the OBO Consortium and the OBO Foundry and also introduce BioTop as a possible additional layer ontology on top of their ontologies. Additionally invite them to report further ideas on how they think their ontologies and/or BioTop needs to be changed to meet their respective needs.

BioTop and the OBO Consortium and Foundry Principles

In order to apply for a membership in the OBO Consortium, an applicant must take several measures to ascertain that the prospective member ontology fulfils a predetermined set of principles [1]. The following list identifies those principles in relation to the application of BioTop to the OBO Consortium:

1. The BioTop ontology is completely open-source and therefore directly accessible and available to everybody. Links to the latest published version – as well as to older versions – can be found on its website. To adopt BioTop in other projects the respective developer or user must only acknowledge its original source and agree not to alter and distribute the modified ontology under its original name and with the same identifiers.
2. The implementation of BioTop is based on OWL-DL which is now accepted as a common and formally defined language by the OBO Foundry and is established as an official standard for building ontologies for the Semantic Web published by the World Wide Web Consortium (W3C). This in turn entails the availability of a wealth of documentation and supporting tools (for

editing as well as classification) and thus allows for a straightforward implementation and adaptation process.

3. BioTop is a top-domain ontology for biomedical research and hence exhibits a clearly defined and delineated subject matter that is distinct from the existing consortium ontologies. Hence it contains only classes and relations necessary to define the higher and (more) general level of this subject field and allows further to link lower level domain ontologies with each other.
4. The string “biotop” is utilized as the unique identifier space for our ontology and serves as the namespace prefix of its OWL-DL implementation. This facilitates avoiding possible naming conflicts as a result of identical class names in BioTop and other ontologies. Also each single class within the ontology holds a unique identifying name to prevent inner-ontology confusion.
5. We include in our ontology precise, plain textual definitions for all classes and relations to resolve the ambiguity that many terms possess in the biomedical sciences. By doing so BioTop cannot only be processed by computer systems but is also understandable for humans and applicable in their regular work.

In addition to these principles the OBO Foundry requires some additional principles to be followed in order to join:

6. We employ a common version control system to make possible the easy identification and retrieval of all available and different ontology versions. This mechanism simplifies greatly the joint collaboration effort by helping to keep track of all changes happening during the everyday development circle, such as the renaming or the deletion of classes or relations for example.
7. The BioTop ontology uses the relations defined by the OBO Relation Ontology through importing its official, published OWL-DL representation. It additionally introduces two new relations that have been formally defined and strictly follow the prescribed pattern of definitions of the original relations.
8. From the beginning we tried to not solely develop the OWL-DL implementation of the ontology but we were also careful not to overlook the need for a comprehensive and comprehensible documentation. We achieved this in two ways, namely by introducing extensive comments and remarks to each class and relation – directly into the implementation – and also by creating descriptive publications targeted at domain experts with no or little background in ontologies.
9. The interest in our work expressed by various researchers after the first release of BioTop showed us that there exists a multitude of potential users for our ontology. By reaching across several independent domain ontologies all users interested in the combination and interoperability of those ontologies can be regarded as possible users of BioTop also.

10. The development started off as a collaborative effort between researchers from two different institutions. Through face-to-face discussions at conferences, postings on mailing lists and personal mail communication, many more people specialized in ontology and biology are currently getting involved and provide input and comments for the continuous BioTop refinement.

Interfaces to Other Ontologies

We tried to achieve a comprehensive coverage of the interface classes in BioTop to link together as many biomedical domain ontologies as possible. But we also tried not to lose track of our original goal: We wanted our ontology to be focused on biomedical matters on a more generic level without any limitations to specific subdomains or some particular species.

Thus we wanted to avoid as much overlap as possible with the given domain ontologies wherever possible and sensible from an ontological point of view. When we found a place of considerable overlap we tried to ascertain whether this overlap problem should be solved on the side of BioTop or rather on the side of the domain ontology. So far this problem has not been tackled in a satisfactory fashion but we plan to contact the responsible curators to further discuss matters and to come up with a principled way of handling such cases.

BioTop sometimes does not provide a direct, logical link to the upper-level classes of the domain ontology. In such cases we obviously see the need to talk to the respective ontology curators to find out whether they should introduce some additional top-level classes to their ontology or whether BioTop is missing any important classes that could be general enough to be relevant to other ontologies as well.

Table 1 lists a small sample of the links we have created so far from BioTop to other domain ontologies and which have been accepted as being valid by the respective curators. A more comprehensive treatment of those links can be found in [6].

Table 1 – Example Links from BioTop to OBO Ontologies

BioTop	OBO Ontologies
Biological Process	Biological Process (GO)
Protein Function	Molecular Function (GO)
Cell Component	Cell Component (GO)
Cell	Cell (CO), Cell (FMA)
Atom	Atoms (ChEBI)
Subatomic Particle	Elementary Particles (ChEBI)
Organic Compound	Organic Molecular Entities (ChEBI)
Tissue	Tissue (FMA)
DNA, RNA	DNA, RNA (SO)
Protein	Protein (SO)

Figure 2 depicts some small, restricted portions of the BFO,

BioTop and GO ontologies to demonstrate the layering and interfacing in between them. In this particular example the domain-specific GO class “Transcription” is linked to the BioTop class “Biological Function” which in turn is linked to the generic, domain independent class “Function” in BFO.

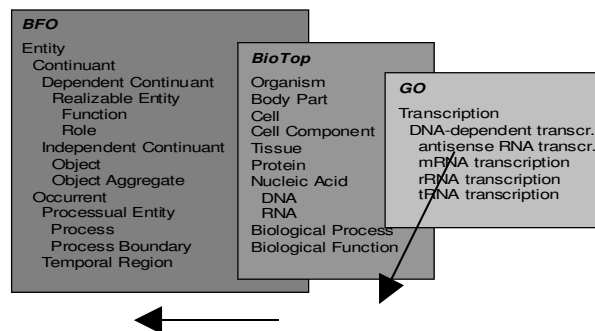


Figure 2 – Classes showing the Ontology Interfacing

Outlook and Future Plans

Usage Scenarios

For the further evaluation and application of BioTop we are currently pursuing the following two main scenarios.

Natural Language Processing and Named Entity Annotation

The European project BOOTStrep [11] will apply the BioTop ontology for various natural language processing purposes in the biology domain (BioNLP). Amongst them is its application to improve the quality of semantic annotation of biological text corpora, i.e. the class names from the BioTop ontology are used as the vocabulary to semantically annotate named entities automatically identified in literature abstracts. The resulting annotated corpora are subsequently employed as a training material for statistical methods which in turn build the basis for more complex BioNLP applications such as meaning disambiguation, relation extraction or anaphora resolution.

Involvement in the Development of Other Ontologies

After successfully joining the OBO Consortium and the OBO Foundry we hope that BioTop will become even more publicly visible and subsequently adapted in the development processes of new and existing ontologies in the context of OBO and elsewhere. By doing so, the developers would acknowledge their (and also our) belief in the necessity for interoperability among individual ontologies, as we have stressed above. An application of BioTop by a larger user base will furthermore facilitate the discovery and resolution of residual problems, bugs and shortcomings in the current BioTop implementation.

Further Improvements

Amelioration and Expansion of Interface Classes

Future versions of BioTop will incorporate continuously improved interface classes. We are discussing the suitability and applicability of the current interface classes with the curators of several OBO domain ontologies. This will lead to the addition of new or the removal of existing BioTop classes, as well as to the clarification of class definitions or documentations.

One specific interest lies in the present development of an ontology for clinical trials within the OBO Consortium. We believe that BioTop could be applied in this context as a link between biological ontologies on one side and medical ones on the other ameliorating the creation of classes that are both of biological and medical interest. To this end, the first author participates in the creation of an ontology for clinico-genomic trials on nephroblastoma (a specific type of kidney cancer found in children) as part a European research project [12]. This ontology will serve after its completion as major input and the basis for the mentioned OBO clinical trial ontology.

Re-Integration of DOLCE

As mentioned above, the initial BioTop version contained a mixture of classes from both BFO and DOLCE at its top level. We now believe that it would be an interesting experiment to reintroduce DOLCE through including its official OWL-DL version as a second top-layer (in addition to keeping the BFO classes). Firstly, this could perhaps allow us to elicit whether the two ontologies are indeed equivalently applicable as the top-level of our ontology. Doing so could also identify places in BioTop where and why one top ontology might excel the other. Secondly, through the addition of DOLCE we might attract users who are accustomed to this top ontology and are therefore reluctant to use an ontology (solely) based on BFO.

Related Work

We are aware of two other projects that are currently engaged in setting up a top-domain ontology for biology. These are the Simple Bio Upper Ontology [13] and GFO-Bio [14]. It seems that at the time of this writing both projects exist in an experimental implementation stadium only and have not produced any publications. Nevertheless we intend to contact both groups for discussions and a possible cooperation.

Our initial intention for BioTop was to improve the interoperability between different biomedical domain ontologies by having a common top-domain ontology. The creation of several top-domain ontologies in this field would obviously be counterproductive and hence some cooperation is essential to achieve a unified solution. Then it would be ideal to have a dedicated workshop to gather the views from more experts in the field to reach a consensus about a single top-domain ontology (which could be based on BioTop or have it as a source).

Conclusion

In this paper we described our current efforts to further develop and extend the biomedical top-domain ontology BioTop. We made the case why an overarching ontology with general classes is important and needed to link independent domain ontologies. We described our integration of BFO into BioTop, discussed our intention to join the OBO Consortium and the OBO Foundry and listed their principles in relation to BioTop. Then we showed some actual interface classes and concluded by detailing on planned BioTop usages and related projects.

Availability

All BioTop material (including its OWL-DL implementation) is available from its website <http://www.ifomis.org/biotop>.

Acknowledgments

This work was supported by the European Union Network of Excellence “Semantic Interoperability and Data Mining in Biomedicine” (NoE 507505) (<http://www.semanticmining.org>) as well as the European Commission STREP project “BOOT-Strep” (FP6 - 028099) (<http://www.bootstrep.eu>).

References

- [1] OBO Consortium. Open Biological Ontologies (OBO), 2004. [<http://obo.sourceforge.net>] – Last accessed: March 25, 2007.
- [2] Gene Ontology Consortium. Creating the Gene Ontology Resource: Design and Implementation. *Genome Research*, 11(8):1425–1433, 2001.
- [3] Rosse C, and Mejino JVL. A Reference Ontology for Biomedical Informatics: the Foundational Model of Anatomy. *J Biomed Inform.* 36:478-500.
- [4] Ohta T, Tateisi Y, and Kim JD. The Genia Corpus: An Annotated Research Abstract Corpus in the Molecular Biology Domain. In Proc. of the 2nd International Conference on Human Language Technology Research (HLT 2002), San Francisco: Morgan Kaufmann, 2002; pp. 82–86.
- [5] Horrocks I, Patel-Schneider PF, and van Harmelen F. From SHIQ and RDF to OWL: The Making of a Web Ontology Language. *Journal of Web Semantics*, 1(1):7–26, 2003.
- [6] Schulz S, Beißwanger E, Wermter J, and Hahn U. Towards an Upper-Level Ontology for Molecular Biology. In Proc. of the American Medical Informatics Assn. Annual Conference (AMIA 2006), Washington, 2006; pp. 694–698.
- [7] Grenon P, Smith B, and Goldberg L. Biodynamic Ontology: Applying BFO in the Biomedical Domain. In Proceedings of the Workshop on Medical Ontologies, Number 102 in Studies in Health Technology and Informatics, Amsterdam: IOS Press, 2004; pp. 20–38.

- [8] Gangemi A, Guarino N, Masolo C, Oltramari A, and Schneider L. Sweetening Ontologies with DOLCE. In Proceedings of the 11th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2002), Sigüenza, Spain, 2002; pp.166–181.
- [9] Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL, and Rosse C. Relations in Biomedical Ontologies. *Genome Biology*, 6(5):R46 (1:15), 2005.
- [10] OBO Foundry Consortium. Open Biological Ontologies Foundry (OBO Foundry), 2006. [<http://www.obofoundry.org>] – Last accessed: March 25, 2007.
- [11] BOOTStrep Consortium. Bootstrapping Of Ontologies and Terminologies Strategic Research Project (BOOTStrep), 2006. [<http://www.bootstrep.eu>] – Last accessed: March 25, 2007.
- [12] ACGT Consortium. Advancing Clinico-Genomic Trials on Cancer (ACGT), 2006. [<http://www.eu-acgt.org>] – Last accessed: March 25, 2007.
- [13] Rector AL, Stevens A, and Rogers J. Simple Bio Upper Ontology, 2006. [<http://www.cs.man.ac.uk/~rector/ontologies/simple-top-bio>] – Last accessed: March 25, 2007.
- [14] Höhndorf R. GFO-Bio: A Biomedical Core Ontology, 2006. [<http://onto.eva.mpg.de/gfo-bio.html>] – Last accessed: March 25, 2007.

Address for correspondence

Holger Stenzhorn, Department of Medical Informatics, Freiburg University Hospital, Stefan-Meier-Str. 26, 79104 Freiburg, Germany