Proceedings of I-MEDIA '07 and I-SEMANTICS '07
Graz, Austria, September 5-7, 2007

A Semantic Web Framework for the Life Sciences Based on Foundational Ontologies and Metadata Standards

Matthias Samwald
(Section on Medical Expert and Knowledge-Based Systems, Medical University of Vienna
Austria
matthias.samwald@meduniwien.ac.at)

Klaus-Peter Adlassnig
(Section on Medical Expert and Knowledge-Based Systems, Medical University of Vienna
Austria
klaus-peter.adlassnig@meduniwien.ac.at)

Abstract: This article describes an ontology framework called *bio-zen,* which can be used for the representation of information from biomedical research on the Semantic Web. The ontology framework adheres to the OWL DL format and is based on existing foundational ontologies and metadata standards like DOLCE, SKOS and Dublin Core. It is optimised for the usage in distributed environments like the Internet. Novel ontological design patterns in bio-zen allow the unification of good ontological consistency with a flexible, clean and intuitive structure. A unique feature of the bio-zen ontology is that it allows the seamless integration of mathematical descriptions and simulation parameters into qualitative information, making a quick transition from plain data to model simulations possible. A growing number of extension packages are available for the ontology, including concepts from taxonomies such as the Gene Ontology, Medical Subject Headings or the NCBI Taxonomy.

Keywords: Semantic Web, Life Sciences, ontology, simulation, data integration
Categories: J.3, H.3.0

1 Introduction

The development of the *bio-zen ontology framework* is an attempt to represent data and information from research in all facets of the *life sciences* on the Semantic Web. The goal of this project is the unification of information that is now scattered over a multitude of different data structures, exchange formats and databases. Through the use of Semantic Web technologies, the decentralised and barrier-free development and exchange of experimental data, hypotheses and biological models becomes possible. A unique feature of the *bio-zen* ontology is that it allows for a seamless integration of mathematical descriptions and simulation parameters into qualitative information, enabling a quick transition from plain data to model simulations and back.

The development of this ontology addresses several pressing needs that are not fulfilled by current ontologies for the life sciences. For example, most of the currently available Semantic Web ontologies for information exchange in the life sciences (e.g.

BioPAX [biopax][Luciano, 05], MGED ontology [Whetzel, 06]) are based on foundational ontologies. This has several disadvantages:

First, it slows down the development of the ontologies. Without a basic ontological base to build on, every project has to re-invent basic relations and classes (at least implicitly). This is a time-consuming task, especially in projects where many participants are developing the ontology in collaboration and each participant possibly implies different and incompatible ontological foundations.

Second, the absence of a sound ontological basis can result in poor design choices. For instance, a feature that can be observed in many domain-specific Semantic Web ontologies is the use of many different properties where the same information could also be conveyed with very few generic properties like 'part of' or 'participant in' or 'attribute of'. One drastic example is the MGED ontology that contains over 100 properties that could be reduced to just a few properties without losing expressivity [Soldatova, 04]. Such redundancies do not only make it very complicated to understand, use and maintain ontologies; they also complicate the construction of queries and the interoperability between ontologies.

2 Methods

The ontology was designed with the Stanford Protégé ontology editor and its OWL plugin [Protégé].

The basic structure of the ontology was created by mapping classes from the BioPAX ontology to the classes of the *Descriptive Ontology for Linguistic and Cognitive Engineering* (DOLCE [Gangemi, 02][dolce]), a foundational ontology available in OWL DL format. It soon became apparent that the BioPAX ontology is mostly focused on an abstract, conceptual representation, as opposed to a direct ontological description of biological reality. This led to some major reinterpretations of the classes imported from BioPAX.

Where possible without significant loss of expressivity, properties and classes from BioPAX were replaced by more generic properties and classes from the DOLCE ontology. Besides the mapping of BioPAX classes and properties, many new structures were added to make the ontology more expressive.

After most of the ontology development was done, a considerable number of classes and properties from DOLCE that were deemed not useful or too complicated for the scope of the ontology were removed. Among the things that were removed are most of DOLCE's advanced modules, all constructs dealing with 'quality spaces' and all inverse properties.

Properties defined in the RDF version of the Dublin Core metadata standard [dublincore] were used to replace some properties from the BioPAX ontology, mostly for the description of database entry provenance and bibliographic information. The core and extended ontologies of the *Simple Knowledge Organisation System* (SKOS [skos]) were added to *bio-zen* for the description of concepts and taxonomies. Because of their special importance, all concepts from the OBO evidence ontology [evidence] were also included in the core *bio-zen* ontology in the form of SKOS concepts.

A basic design requirement for the *bio-zen* ontology is conformance with the OWL DL standard. While the DOLCE ontology is already valid OWL DL, the SKOS

and Dublin Core are not: SKOS uses a mixture of constructs from OWL and generic RDF; the official version of Dublin Core is pure RDF Schema. Some minor modifications were made to these ontologies to make them conform to OWL DL. All of the ontologies were merged into a single OWL file to avoid problems with ontology imports when a client is disconnected from the internet.

3 Basic principles and design patterns

The ontology is built upon existing foundational ontologies and metadata standards (DOLCE, SKOS, Dublin Core). Through the use of established foundational ontologies and metadata standards, *bio-zen* is rooted in a sound ontological framework, easing the interoperability with ontologies from other domains.

Statements made in *bio-zen* are direct ontological descriptions of biological reality and not of some abstraction of biological reality. This is in contrast to many other bioinformatics projects based on RDF, which are focused on the description of such abstractions, e.g. the organisation of database records. On the contrary, *bio-zen* is focused on the description of *spatio-temporal particulars* (concrete biological things existing in a certain time and space).

Users of the ontology only need to make OWL individuals to represent information. The definition of new classes is not necessary, which helps to keep the class structure clean and simple. It also helps to delineate ontology developers, i.e. people that are educated about the development of ontologies and that make use of ontology editors like Protégé [Protégé], from end-users, i.e. people that are not educated about ontology development and that make use of specialised software or internet portals. Furthermore, this design avoids many other grave problems that arise with OWL ontologies that have an overly complex class structure [Samwald, 06]. If the user of the ontology intends to make statements about some general observed phenomena (e.g. 'Drosophila has two wings'), these general principles are *exemplified* with a certain spatio-temporal-particular that acts as a canonical reference.

A characteristic feature of the ontology is that it integrates two different approaches of information representation in a common framework: 'realist' ontological descriptions and 'conceptual' taxonomies and concept hierarchies. The difference between the two approaches lies in the levels of abstraction: whilst the former focuses on describing reality itself, the latter focuses on describing the conceptualisations humans have made about reality. Unifying both approaches in one common framework makes it possible to combine the specific advantages of each approach in the best way possible. The consistency of the realist approach is complemented with the flexibility of the conceptualist approach. Making a clear distinction between both approaches reduces the susceptibility to inconsistencies.

'Realist' ontological descriptions in *bio-zen* are based on DOLCE and *spatio-temporal particulars* and employ a rich set of classes and properties. 'Conceptual' descriptions are based on SKOS, and mainly use only one class (*'skos:Concept'*) and only a few properties to describe the relations between concepts (*'broader'*, *'narrower'*, *'related'* etc.). The two forms of descriptions are almost fully disconnected. The only property that can be used to connect the two is called *described-by*. It can be used to annotate and identify a *spatio-temporal-particular*

with one or more concepts (Fig. 1). This design pattern and the *described-by* property are innovations of the *bio-zen* ontology.

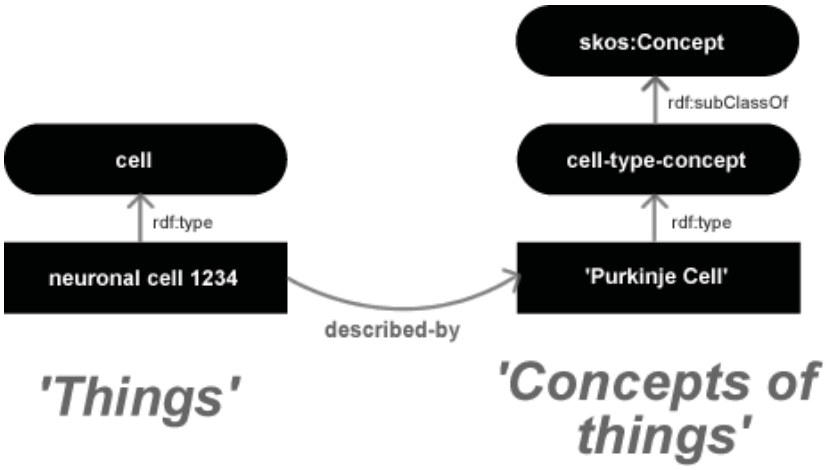

Figure 1: The two ‚worlds' in the bio-zen framework – the world of real things located in a certain space and time and the world of abstract concepts about things. Both worlds can only be connected through the 'described-by' property – otherwise, they are completely separated.

This design pattern has the major advantage that it allows users to put *spatio-temporal-particulars* into a hierarchy without resorting to complicated OWL class hierarchies. Furthermore, it allows users to extend the hierarchy by simply making new instances of *skos:Concept* without needing to define new OWL classes. This keeps the distinction between ontology developers and end-users intact. It gives end-users the ability to create new categorisations while avoiding the potential pitfalls associated with the definition of new OWL classes [Rector, 04].

The concept annotations can be used to express similarity between different *spatio-temporal-particulars* and to 'glue' graphs from different sources (and possibly different ontologies) together (Fig. 2).

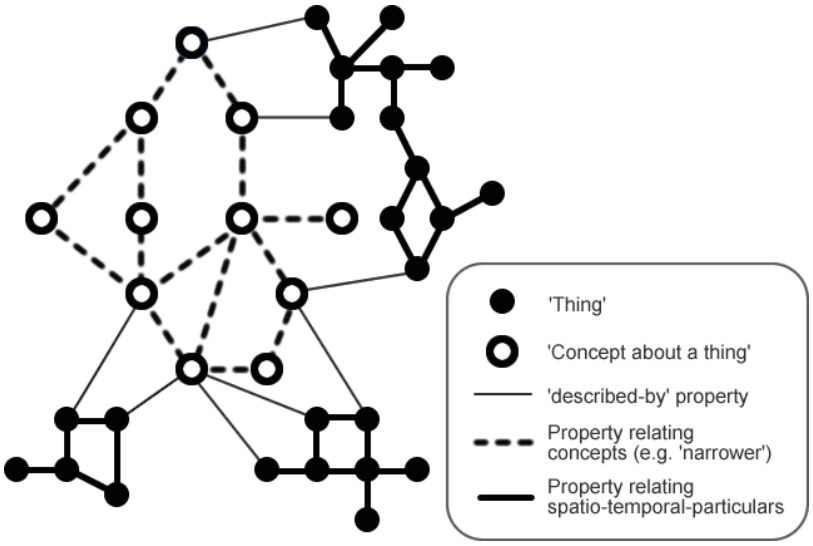

Figure 2: Concepts can act as a kind of ‚glue' between different models. Similarities among different OWL individuals can be inferred from similar concept annotations.

Unnecessary over-specialisation of properties is avoided where possible. Generic properties like 'has part', 'has constituent', 'broader concept than', 'caused by' etc. are the backbone of the whole ontology. The direction of the properties is used consistently, in that they are pointing from the 'larger, containing' thing to the 'smaller, contained' thing (e.g. parts, qualities, participants, features). This allows for the creation of simple query algorithms that can easily capture all parts and details of a resource through iteration.

All of the classes and properties of the ontology are labelled with the *rdfs:label* property. Many other ontologies (e.g. BioPAX) solely rely on the URIs of their resources as human-readable descriptors. Since URIs are essentially not meant to contain meaningful information, this is a grave shortcoming that hampers the development of user-friendly application interfaces.

The ontology allows the seamless integration of mathematical descriptions into existing information. Mathematical formulas can be used to describe the correlation of different qualities (e.g. concentrations of molecules) over a certain timeframe. The mathematical formulas can be expressed in MathML. Physical measurements are strictly represented according to the International System of Units (e.g. kilogram, second, Kelvin) and floating point numbers. This gives *bio-zen* the power to act as a modelling language similar to the popular *Systems Biology Markup Language* (SBML [Hucka, 03]).

Contrary to other ontologies in the field, molecular interactions are modelled as stochastic processes involving populations of molecules, not as single events that involve single molecules. This approach is much closer to biological reality in most occasions, and avoids some grave consistency problems associated with the other approach.

The ontology has basic support for fuzzy logic - like constructs. A property called *'realness'* can be used to assign a measure of uncertainty to every *spatio-temporal-particular* in the ontology. The value of this property should be a floating point number between 0 and 1. This is a metric for how certain it is that the described entity really exists in nature. If a user applies a realness-value of 1 to an entity this essentially means that he or she is sure that the entity exists in nature. Lower values mean that he or she is less certain. A value of 0 implies that there is no known evidence that the entity exists in nature.

Another property called *'interestingness'* can be used to make subjective ratings about how interesting a given resource is. This might seem unusual but is a very important feature, as not all proven and true facts are relevant for scientific discourse. The realisation of such a system for fuzziness values and ratings based on simple datatype properties does produce only a small triple overhead and can be queried with good performance, as compared to other design patterns that make use of RDF reification. As most relations between biological entities in *bio-zen* are reified to start with, the use of RDF reification is not necessary.

4 Ontology specifications and status

The ontology is designed to conform to the OWL DL standard, which guarantees computability and eases the development of tools that work with the ontology. Currently, the core ontology defines around 130 classes, 40 datatype properties and 60 object properties. This includes entities that have been taken over from DOLCE, SKOS, Dublin Core and the evidence code ontology.

There are several extensions available in the form of OWL files. These extensions represent concepts from other biomedical ontologies and taxonomies as concepts based on the SKOS ontology. All of the current extensions are derived from taxonomies that are part of the *Open Biomedical Ontologies* repository [OBO]. Currently, the following extension packages are available:

The *Gene Ontology* [Ashburner, 00] extension (defines 20.000 concepts), the *Medical Subject Headings* [MeSH] extension (23.000 concepts), the *NCBI Taxonomy* [Wheeler, 00] extension (340.000 concepts), the *celltype ontology* [Bard, 2005] extension (800 concepts), the *sequence ontology* [Eilbeck, 05] extension (1000 concepts) and the *INOH Molecule role ontology* [inoh] extension (7.200 concepts).

With all extension packages taken together, *bio-zen* is among the largest structurally coherent ontologies currently available in OWL. Further extension packages will be added in the future.

The ontology, all extensions and a manual can be downloaded from http://neuroscientific.net/index.php?id=download

5 Discussion

The experience of mapping BioPAX to DOLCE showed that such a mapping process can reveal undiscovered problems in the original ontology. Ontology developers in the field of the life sciences should therefore be encouraged to conduct such mappings

to test the consistency of their ontologies. It also became apparent that the real value of foundational ontologies lies in the definition of a small set of the most basic concepts (e.g. 'occurent', 'continuant', 'process', 'part of'). The value of a foundational ontology can be severely decreased by unnecessary details and specialisations.

The mapping process also necessitated minor changes to some of the foundational ontologies in order to make them valid OWL DL. There seem to be no widely accepted agreements on whether such minor modifications to existing ontologies are acceptable or not and how they should be handled. Such an agreement will become necessary when ontology use and therefore also ontology *reuse* on the web will increase.

A project related to *bio-zen* that is worth mentioning is the *Ontology of Biomedical Investigation* [OBI]. The development of OBI is based on the *Basic Formal Ontology* [BFO]. Compared to the BFO, DOLCE has the advantage that it already has an established formalisation in OWL, while the formalisation of the BFO in OWL is still under development. As many of the concepts of the BFO have a counterpart in DOLCE, interoperability between both ontologies should be easily achievable when necessary.

6 Conclusions

The *bio-zen* ontology is among the first functional Semantic Web ontologies for molecular biology that are based on widely accepted foundational ontologies. It is also the first metadata format that attempts to unite taxonomies, ontologies, qualitative data and mathematical modelling in a coherent data structure. It introduces new design patterns, e.g. strategies of avoiding overuse of OWL classes and the ability to easily bind together incompatible ontologies through concept annotations. These design patterns might prove useful for future ontology developments in many different knowledge domains, not only the life sciences.

References

[Ashburner, 00] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet (2000), 25(1):25-29.

[Bard, 2005] J. Bard, S.Y. Rhee, M. Ashburner. An ontology for cell types. Genome Biol 2005; 6(2):R21.

[BFO] http://ontology.buffalo.edu/bfo/

[biopax] http://www.biopax.org/

[dolce] http://www.loa-cnr.it/DOLCE.html

[dublincore] http://dublincore.org/documents/dces/

[Eilbeck, 05] K. Eilbeck, S.E. Lewis, C.J. Mungall, M. Yandell, L. Stein, R. Durbin et al. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol (2005) 6(5):R44.

[evidence] http://obo.sourceforge.net/cgi-bin/detail.cgi?evidence_code

[Gangemi, 02] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, L. Schneider, Sweetening ontologies with DOLCE, in: Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02), Springer-Verlag, London, UK, 2002, pp. 166-181

[Hucka, 03] M. Hucka, A. Finney, H.M. Sauro, H. Bolouri, J.C. Doyle, H. Kitano et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 2003; 19(4):524-531.

[inoh] http://www.inoh.org/download.html

[Luciano, 05] J.S. Luciano. PAX of mind for pathway researchers. Drug Discov Today (2005), 10(13):937-942.

[MeSH] http://www.nlm.nih.gov/mesh/

[OBI] http://obi.sourceforge.net/

[OBO] http://obo.sourceforge.net

[Protégé] http://protege.stanford.edu/

[Rector, 04] A. Rector, N. Drummond, M. Horridge, J. Rogers, H. Knublauch, R. Stevens et al. OWL Pizzas: Practical Experience of Teaching OWL-DL: Common Errors & Common Patterns. 3257 ed. 2004.

[Samwald, 06] M. Samwald, Classes Versus Individuals: Fundamental Design Issues for Ontologies on the Biomedical Semantic Web, in: Proceedings of the STC2006, AKA, Berlin 2006, pp. 335-340

[skos] http://www.w3.org/2004/02/skos/

[Soldatova, 04] L.N. Soldatova, R.D. King. Are the current ontologies in biology good ontologies? Nat Biotechnol (2005), 23(9):1095-1098.

[Wheeler, 00] D.L. Wheeler, C. Chappey, A.E. Lash, D.D. Leipe, T.L. Madden, G.D. Schuler et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res (2000), 28(1):10-14.

[Whetzel, 06] P.L Whetzel, H. Parkinson, H.C. Causton, L. Fan, J Fostel, G. Fragoso et al., The MGED Ontology: a resource for semantics-based description of microarray experiments. Bioinformatics (2006), 22(7):866-873.