Applied Ontology 3 (2008) 219–227
DOI 10.3233/AO-2008-0055
IOS Press

GFO-Bio: A biological core ontology

Robert Hoehndorf [a,b,d], Frank Loebe [b,c], Roberto Poli [a,e], Heinrich Herre [b,c] and Janet Kelso [a]

[a] *Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany*
[b] *Research Group Ontologies in Medicine (Onto-Med), Institute of Medical Informatics, Statistics and Epidemiology (IMISE), University of Leipzig, Leipzig, Germany*
[c] *Department of Computer Science, University of Leipzig, Leipzig, Germany*
[d] *Graduate Program Knowledge Representation, Department of Computer Science, University of Leipzig, Leipzig, Germany*
[e] *University of Trento, Trento, Italy*

Abstract. The rapid increase in the number and use of biological ontologies necessitates developing systems for their integration. In this paper we present a core ontology for biology, and outline its application for integrating biological domain ontologies. Our ontology rests on a foundational ontology, which offers higher-order categories and a theory of levels of reality. The core ontology is implemented in two separate components, each of which adheres to OWL-DL. These can be used independently with efficient DL reasoners, but they will be most effective when used together, which necessitates working with an OWL-Full ontology. The ontology is freely available from our website at: http://bioonto.de/pmwiki.php/Main/GFO-Bio.

Keywords: Core ontology, biomedical ontology, categories, levels of reality, ontology integration, default knowledge, General Formal Ontology

1. Introduction

As the number of ontologies used in the biomedical domain grows, their alignment becomes an increasingly difficult task (Euzenat & Shvaiko, 2007). The Open Biomedical Ontologies (OBO) (Smith et al., 2007) alone contain more than 60 ontologies, continually increasing in number. The majority of biomedical ontologies are domain specific, covering single domains. Categories in these ontologies are linked using relations such as *is-a*, *part-of* or *develops-from* (Smith et al., 2005).

By contrast, little attention has been given to developing core ontologies for biology. A *core ontology* is an ontology that formally describes and defines the basic categories within a domain (Valente & Breuker, 1996; Herre, 2008). Because a core ontology's categories are so general, they are similar to the categories available in foundational ontologies.

A *foundational ontology* contains categories covering all domains of reality (Sowa, 2000; Herre et al., 2006). A primary function of a core ontology is to specialize the concepts and relations from a foundational ontology to concepts that exist in a domain. On the other hand, a core ontology should make the nature of the domain it captures precise in order to distinguish it from other domains.

Based on this understanding of core ontology, we describe the biological core ontology GFO-Bio, which is based on the foundational ontology GFO[1] (Herre et al., 2006). After outlining the relevant

[1] General Formal Ontology, http://www.onto-med.de/ontologies/gfo.

features of GFO, we present the structure of GFO-Bio and its implementation in OWL. We then discuss how GFO-Bio can be used for integrating biological domain ontologies. This integration is grounded in the foundational ontology, and will accommodate a plurality of views on these domains. The paper concludes by comparing GFO-Bio to related approaches and describing current applications.

2. Foundation

We chose to found the core ontology GFO Bio in the General Formal Ontology (GFO) (Herre et al., 2006; Heller & Herre, 2004) because GFO can be distinguished from other foundational ontologies by several features, which make it particularly useful for application in a biological core ontology.

2.1. Object-process integration

The GFO view of static and dynamic entities, of objects and processes, integrates the phenomena of persistence, of presence and features of processes. We consider these characteristics highly relevant for building expressive ontologies, particularly a biological core ontology.

To outline this approach, the intuitive notion of objects, often called endurants or continuants (Masolo et al., 2003; Grenon, 2003), is understood in GFO as a unity of three kinds of entities, namely of a *persistant* (a category), a specific *process* (the extended “life” of the object) and a series of *presentials*. Presentials are entities wholly present at exactly one point in time. The persistence of an object through time is modelled as a persistant whose instances are presentials. These presentials are (logically) distinct individuals and may differ in their properties. They are connected by a process that is comprised of them as its only participants, exactly one for each time point within its temporal extension.

2.2. Higher-order categories

Core ontologies should explain relations between the organization of categories as well as subfields of the domain that they cover. We expect that higher-order categories are one of the instruments that can be used to achieve this purpose. A *category* is an entity that can have instances and can be predicated on other entities, whereas *individuals* cannot be further instantiated. A *higher-order* category is a category whose instances include themselves categories. In contrast, a *simple* category has only individuals as instances.

GFO distinguishes at least three kinds of categories: universals, concepts and symbols (Gracia, 1999). *Universals* are constituents of the real world; they are associated to invariants of the spatio-temporal world. *Concepts* are categories expressed by linguistic expressions and are presented as meanings in someone’s mind. *Symbols* are categories that can be instantiated by *tokens*. We assume that symbols are always simple categories.

This framework is intended to represent semiotic information, in particular those aspects of the world that pertain to signs and symbols, syntax, semantics and pragmatics. The representation and treatment of this kind of information is not only relevant in social or cognitive sciences, but also in natural sciences, notably in biology, in the context of biosemiotics or zoosemiotics. Important notions and processes in molecular biology and bioinformatics, for example, are related to symbols and sequences of symbols.

2.3. Bridging levels of granularity: Levels of reality

In order to address the problems both of bridging levels of granularity and permitting multiple views on a domain, GFO includes a theory of levels of reality (Poli, 2001). In simplified terms, a *level* is a system of interrelated categories. Moreover, distinct levels are themselves related in particular ways. Three major levels of reality (called *ontological strata*) can be distinguished: the *material stratum*, the *psychological stratum* and the *social stratum*. Each stratum is further organized into sublevels, where scientific fields like physics, chemistry or biology provide suitable starting points for identifying such sublevels.

The theory of levels of reality is the natural setting for elaborating on an articulated theory of the forms of causal dependence. It is grounded on the hypothesis that any ontologically different level has its own form(s) of causality. Material, psychological and social forms of causality could therefore be distinguished (and compared) in a principled way. Aside from the basic causality between phenomena of the same nature, the theory of levels allows upward and downward forms of causality (from a lower level to an upper one and reverse)[2] to be singled out.

The connection between the theory of levels and causality entails the recognition that each level of reality may trigger its own causal chain. This may even be considered as a definition for level of reality: a level of reality is distinguished by its specific form of causality. As a consequence, GFO includes criteria with which to distinguish among levels of reality and levels of granularity.

2.4. Integration of default knowledge

GFO has been used to formalize default knowledge in the context of ontology integration (Hoehndorf et al., 2007). For this purpose, a non-monotonic logic formalism was used. GFO can be extended with axioms requiring a non-monotonic logic. We consider the ability to represent default knowledge coupled with a principled method for integrating it with other forms of knowledge as beneficial features of a foundational ontology to employ as the basis for a biological core ontology. For instance, anatomy is a domain that requires the capability for handling default knowledge.

3. GFO-Bio and its implementation in OWL

GFO-Bio has been formalized primarily in description logic using OWL-DL as the representation format. It includes an ontology of functions (Burek et al., 2006) and has been developed to serve as an ontology for the BOWiki (Hoehndorf et al., 2006), an ontology-based semantic wiki. GFO-Bio, along with a set of tools and additional information, are freely available under a BSD-style license[3] from our website.[4]

There are currently two GFO-Bio components that can be used individually, but they attain their full expressivity only in combination. The main component is gfo-bio.owl, which focuses primarily on simple categories in the biological domain and is illustrated in Fig. 1. The extension component is gfo-bio-meta.owl. It is used for interrelating categories through higher-order categories and relations, and allows a meta-ontological analysis of the categories and relations included in gfo-bio.owl.

[2] Andersen et al. (2000) collects a series of recent studies on the theme.

[3] http://www.opensource.org/licenses/bsd-license.php.

[4] http://bioonto.de/pmwiki.php/Main/GFO-Bio.

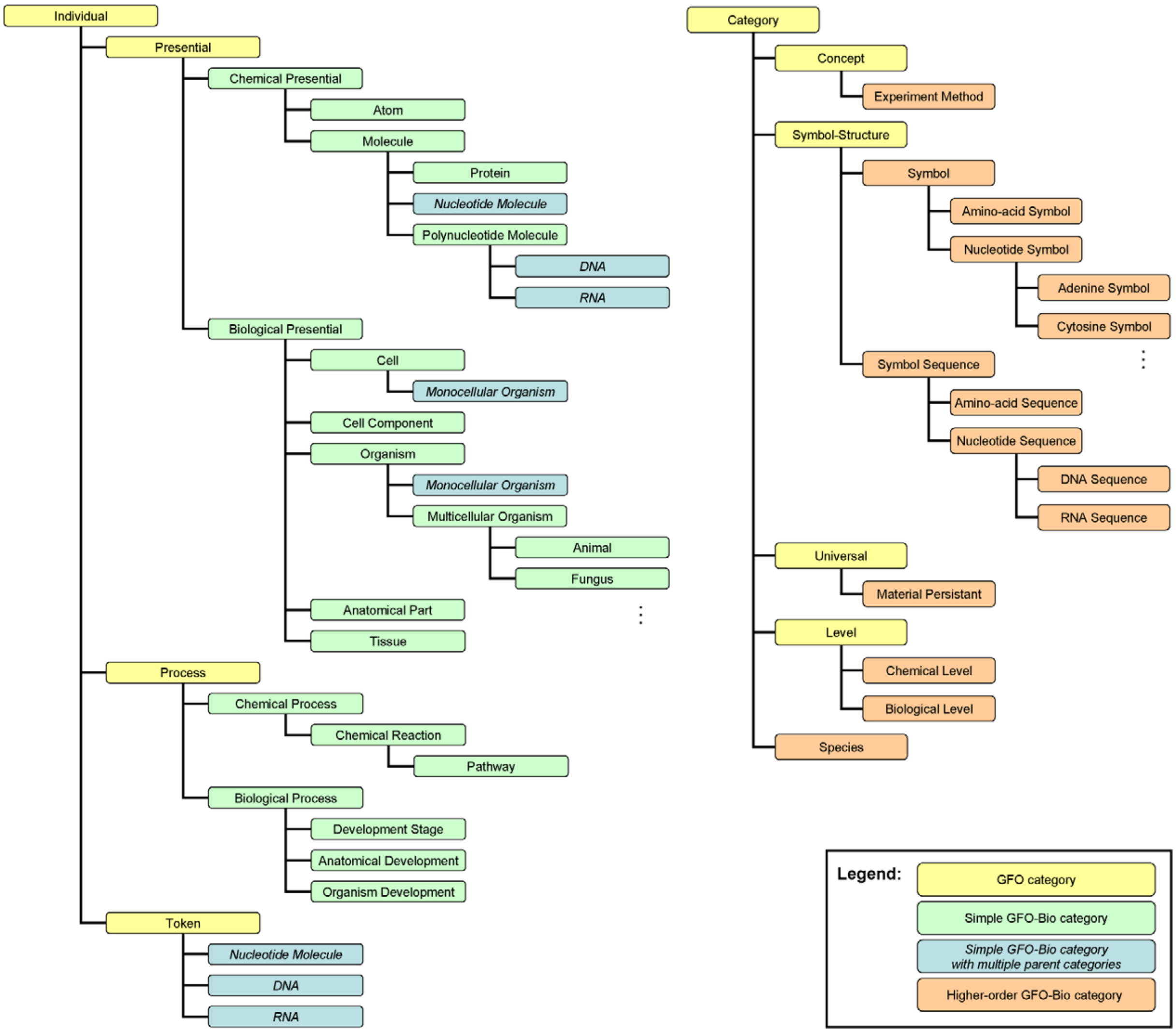

Fig. 1. An extract of GFO-Bio's branch focusing on simple categories.

3.1. *Branch of simple categories*

Most biological domain ontologies are comprised of only *simple categories*, i.e., categories that are instantiated by individuals.[5] For integrating such ontologies, gfo-bio.owl provides a system of categories that are upper categories within the OBO ontologies, like *Cell*, *Organism*, *Plant* and *Biological process*. These categories are aligned with the foundational ontology GFO. With respect to OWL, they are implemented as classes. The only higher-order categories included in gfo-bio.owl pertain to sequences and symbols.

[5] Throughout the article, "individual" refers to the ontological understanding of this term as introduced in Section 2.2. If we refer to the notion of (logical) individuals in the context of OWL, "OWL individual" is employed. "Instance" is used for both notions where disambiguation is clear from context. "Class" refers to OWL classes.

3.2. Branch of higher-order categories

The extension component, gfo-bio-meta.owl, provides a means for meta-ontological analysis and classification of categories from biomedical ontologies.

In its current state, it contains a number of categories – as subcategories in GFO's category classification – whose instances are the categories of specific domain ontologies. For example, the category of *taxonomy of cell* has as instances all subcategories of the *Cell* category.

The relationships between categories in the gfo-bio.owl branch are modelled as assertions between instances of classes within gfo-bio-meta.owl. For example, the category from the celltype ontology *female germ line stem cell* stands in the relation *CC-isa* to *female germ cell*.[6] *CC-isa* corresponds to the `SubClassOf`-relation in gfo-bio.owl, which is lifted to a relation between instances in gfo-bio-meta.owl.

This approach allows for directly representing OBO-style directed acyclic graphs in OWL together with the possibility of simplifying reasoning over these graphs. It also provides a means for describing sub-domains as instances of higher-order categories. Ontologies represented in the OBO format can be converted into the format required by gfo-bio-meta.owl using a conversion tool that we provide.[7]

3.3. Relating higher-order to simple categories

To re-establish the connection between the distinct formalizations of the relationships between categories and the relationships between individuals, SWRL-like axioms can be added to the combination of the two branches. Table 1 contains a number of examples.

To integrate default knowledge, non-monotonically treated formulas must be added. An answer set program for each relation that is used in the description of default knowledge must be created (Hoehndorf et al., 2007).

Table 1

Examples of axioms for interrelating the branches of GFO-Bio

```
    Individual(?C1 (type OrganismTaxonomy)) ⟺ SubClassOf(?C1 Organism)
```

A category is an instance of the higher-order category *OrganismTaxonomy* if and only if (iff) it is subsumed by the simple category *Organism*.

```
    Individual(?C1 value(CC-isa ?C2)) ⟺ SubClassOf(?C1 ?C2)
```

Two categories stand in *CC-isa* relation in gfo-bio-meta.owl iff they are subclasses in gfo-bio.owl. Essentially, the *CC-isa* relation is a strict synonym of `SubClassOf` in OWL.

```
    Individual(?C1 value(CC-part-of ?C2))
      ⟺ SubClassOf(?C1 restriction(II-part-of someValuesFrom(?C2)))
```

The simple category C_1 stands in *CC-part-of* relation to the simple category C_2 in gfo-bio-meta.owl iff every instance of C_1 (an individual) is related by *II-part-of* to an instance of C_2, i.e., the right-hand axiom is contained in gfo-bio.owl.

Note: The basic notation around the equivalence arrows is OWL abstract syntax, but using variables for named OWL individuals or classes in a SWRL-style notation, i.e., indicated by a leading question mark.

[6] We prefix relations according to their domains and ranges with combinations of I for "individual" and C for "category", cf. Hoehndorf et al. (2007) (Table 1, p. 377.4).

[7] http://bioonto.de/pmwiki.php/Main/GFO-Bio.

4. Integrating ontologies with GFO-Bio

The integration of biomedical ontologies using GFO-Bio can be achieved in several ways, depending on the intended purpose for the integrated system and the desired behavior in terms of computational tractability. First we consider the integration based on the branch of simple categories, i.e., using gfo-bio.owl alone. To integrate a given set of biomedical ontologies, an OWL-DL version in which simple categories are formalized as OWL classes must be acquired or produced for each ontology. The resulting OWL-DL files must then be imported by gfo-bio.owl, and their categories must be defined using categories from gfo-bio.owl. For example, the "Cell" category of a celltype ontology can be declared a subclass of or equivalent to GFO-Bio's "Cell" category.

The consistency of the overall system following integration can be automatically evaluated, and the integrated system will generally yield a more expressive categorial system than each ontology considered individually. This style of formalization produces very large TBoxes, which must be processed by automated reasoners, which is why the computational demands on such a combined system are fairly high.

The second module of GFO-Bio for integrating ontologies, gfo-bio-meta.owl, permits more efficient reasoning, but reduces expressivity with regard to exploiting relations among categories. Starting from a set of biomedical ontologies, OWL-DL files must be produced in which the ontologies' categories are formalized as instances in OWL. Files thus created must be imported by gfo-bio-meta.owl, and the categories from the imported ontology are declared as instances of a class in gfo-bio-meta.owl. For example, the "Cell" category from the Celltype Ontology is represented as:[8]

```
Individual(Cell (type Category))
```

In this representation of the biomedical ontologies, relationships between categories, as expressed in the OBO-style directed acyclic graphs (DAGs), can be directly modelled as relationships between OWL instances. For example, the relationship expressed in the DAG of the Gene Ontology's cellular component ontology, "Membrane *part-of* Cell", is represented as:

```
Individual(Membrane value(CC-part-of Cell))
```

Here, "Membrane" and "Cell" are treated as instances of GFO-Bio's "Category" class, and a relation *CC-part-of* between "Membrane" and "Cell" is asserted ("CC" indicating the ontological category–category reading of the relation).

In contrast, in gfo-bio.owl, "Membrane" and "Cell" are created as classes, and in addition to declaring the equivalence of "Cell" to GFO-Bio's cell category, the following restriction is produced (in line with Horrocks, 2007):

```
SubClassOf(Membrane restriction(II-part-of someValuesFrom(Cell)))
```

While neither the first nor the second step alone require more than the description logic fragment of OWL, together they produce an OWL-Full (McGuinness & van Harmelen, 2004) ontology. While certainly the least tractable case, this combination provides the most comprehensive integrated system of categories.

[8] In all examples, namespace identifiers are omitted for readability.

5. Discussion

Several other core ontologies are available for biology, among them the Simple Bio Upper Ontology[9] (SBUO) and the BioTop ontology (Schulz et al., 2006). Both incorporate the aspects of a core ontology that GFO-Bio's simple categories branch covers: they provide well-defined categories that can be used to classify individuals. The main differences between GFO-Bio and alternative approaches pertain to the properties of GFO as outlined in Section 2: including higher-order categories, treating semiotic information, bridging levels of granularity and integrating objects and processes.

Because BioTop and the SBUO are biological core ontologies based on or inspired by the foundational ontologies BFO (Grenon, 2003) and DOLCE (Masolo et al., 2003), neither includes higher-order categories. Higher-order categories are used in GFO-Bio to model symbols and sequences, model persistence through time and explicate the intension of the relations used in biomedical ontologies.

A major difference between GFO-Bio and both BioTop and SBUO pertains to the capabilities for representing default knowledge. Default knowledge in GFO-Bio can be encoded in relations that hold between categories (Hoehndorf et al., 2007). The intension of these relations can be defined using answer set programs that define them as holding by *default*. Answer set programming (Lifschitz, 2002) provides a non-monotonic knowledge representation formalism for use with GFO-Bio. For representing defaults and exceptions, a non-monotonic logic is more versatile than monotonic logics. Currently, both BioTop and SBUO deal with "normality" and "non-normality" using OWL-DL exclusively, which is a monotonic logic. It is thus unclear whether default knowledge can be incorporated into these systems. Further, the lack of higher-order categories makes it difficult to apply the method for representing defaults used in GFO-Bio to these ontologies.

A further difference between GFO-Bio and BioTop is the treatment of biological symbols and sequences.[10] In GFO-Bio, sequences are categories that can have instances (the tokens). They are entities *sui generis* and do not depend on any other entity, whereas in the BioTop ontology, they are generically dependent[11] continuants that depend on the existence of a molecule. For example, the instance of a DNA sequence in BioTop requires the existence of some DNA molecule that exhibits this sequential structure. However, the sequences used in biological research are not always the sequence of some molecule. It is unlikely that the "canonical" sequence of human chromosome 5 is exhibited by any DNA molecule, due to sequencing errors, the presence of mutations, variations or similar. It is not clear how sequencing errors, variations or mutations are represented in BioTop. The same holds for randomly or artificially created sequences that are studied as entities in their own right.

Because BioTop and the SBUO are similar in many respects, efforts are ongoing to reconcile both ontologies or merge them into one. In addition, a mapping between BioTop and the UMLS Semantic Network exists which provides axioms for UMLS classes. Their main application area remains the integration of OBO ontologies.

GFO-Bio is currently being applied as a built-in ontology for the BOWiki[12] (Hoehndorf et al., 2006), a semantic wiki that uses an OWL reasoner in conjunction with an included ontology. As a background ontology for large knowledge bases like those developed using the BOWiki, GFO-Bio provides for flexible trade-offs with respect to the scalability of reasoning. That means, for some purposes (e.g.,

[9] http://www.cs.man.ac.uk/∼rector/ontologies/simple-top-bio/.

[10] SBUO does not contain an explicit sequence class, but the more general *Pattern* class, which is left unexplained.

[11] A category C is generically dependent on the category D, if, necessarily, whenever an instance c of C exists, then some instance d of D exists.

[12] http://bowiki.net.

certain forms of type-checking), parts of the knowledge base can be formulated in such a way that they are less demanding with respect to OWL reasoning, by making use of GFO-Bio's higher-order categories.

Also, in the context of text-mining, terms may refer to both categories (e.g., "mouse" as the category of all mice) and individuals (e.g., Jerry, the "mouse" used in a particular experiment). Using GFO-Bio for classifying the results of text-mining together with GFO-Bio's support for non-monotonic reasoning allows to disambiguate these cases.

To summarize, the strengths of GFO-Bio lie in the representation of levels of granularity (through levels of reality), the representation of knowledge pertaining to symbols and sequences and the detailed elaboration of the intension of relations between categories using higher-order categories, rules and answer set programming. We believe that this provides a versatile and detailed representation of both non-canonical and canonical information, and for extended support of multiple, alternative views on domains, especially pertaining to the definition of relations.

Acknowledgement

We are grateful to Hernán Burbano for his advice and discussions, and to Christine Green for help with the preparation of the English manuscript.

References

Andersen, P.B., Emmeche, C., Finnemann, N.O. & Christiansen, P.V. (2000). *Downward Causation: Minds, Bodies and Matter*. Aarhus, Denmark: Aarhus University Press.

Burek, P., Hoehndorf, R., Loebe, F., Visagie, J., Herre, H. & Kelso, J. (2006). A top-level ontology of functions and its application in the Open Biomedical Ontologies. *Bioinformatics*, *22*(14), e66–e73.

Euzenat, J. & Shvaiko, P. (2007). *Ontology Matching*. Berlin: Springer.

Gracia, J.J.E. (1999). *Metaphysics and Its Tasks: The Search for the Categorial Foundation of Knowledge*. Albany, NY: State University of New York Press.

Grenon, P. (2003). BFO in a nutshell: A bi-categorial axiomatization of BFO and comparison with DOLCE. IFOMIS Report No. 06/2003, Institute for Formal Ontology and Medical Information Science (IFOMIS), University of Leipzig, Leipzig, Germany.

Heller, B. & Herre, H. (2004). Ontological categories in GOL. *Axiomathes*, *14*(1), 57–76.

Herre, H. (2008). The ontology of medical terminological systems: Towards the next generation of biomedical ontologies. In M. Healy, A. Kameas & R. Poli (eds), *Theory and Applications of Ontology (TAO)* (Vol. 2, forthcoming). Berlin: Springer.

Herre, H., Heller, B., Burek, P., Hoehndorf, R., Loebe, F. & Michalek, H. (2006). General Formal Ontology (GFO) – A foundational ontology integrating objects and processes (Version 1.0). Onto-Med Report No. 8, Research Group Ontologies in Medicine, Institute of Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig.

Hoehndorf, R., Prüfer, K., Backhaus, M., Herre, H., Kelso, J., Loebe, F. & Visagie, J. (2006). A proposal for a gene functions wiki. In R. Meersman, Z. Tari & P. Herrero (eds), *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops: Proceedings, Part I*, Montpellier, France, October 29–November 3 (Vol. 4277, pp. 669–678). Berlin: Springer.

Hoehndorf, R., Loebe, F., Kelso, J. & Herre, H. (2007). Representing default knowledge in biomedical ontologies: Application to the integration of anatomy and phenotype ontologies. *BMC Bioinformatics*, *8*(1), 377.1–377.12.

Horrocks, I. (2007). OBO Flat File Format syntax and semantics and mapping to OWL Web Ontology Language. Technical report (Editor's Draft, March 04), University of Manchester (http://www.cs.man.ac.uk/∼horrocks/obo/syntax.html).

Lifschitz, V. (2002). Answer set programming and plan generation. *Artificial Intelligence*, *138*(1/2), 39–54.

Masolo, C., Borgo, S., Gangemi, A., Guarino, N. & Oltramari, A. (2003). WonderWeb Deliverable D18: Ontology library (final). Technical report, Laboratory for Applied Ontology, ISTC-CNR, Trento, Italy.

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

Poli, R. (2001). The basic problem of the theory of levels of reality. *Axiomathes*, *12*(3/4), 261–283.

Schulz, S., Beisswanger, E., Hahn, U., Wermter, J., Kumar, A. & Stenzhorn, H. (2006). From Genia to BioTop: Towards a top-level ontology for biology. In B. Bennett & C. Fellbaum (eds), *Formal Ontology in Information Systems: Proceedings of the Fourth International Conference (FOIS 2006)*, Baltimore, MD, USA, November 9–11 (pp. 103–114). Amsterdam: IOS Press.

Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A.L. & Rosse, C. (2005). Relations in biomedical ontologies. *Genome Biology*, *6*(5), R46.1–R46.15.

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

Sowa, J.F. (2000). *Knowledge Representation: Logical, Philosophical and Computational Foundations*. Pacific Grove, CA: Brooks/Cole.

Valente, A. & Breuker, J. (1996). Towards principled core ontologies. In B.R. Gaines & M.A. Musen (eds), *Proceedings of the 10th Knowledge Acquisition Workshop (KAW'96)*, Banff, AB, Canada, November 9–14 (pp. 301–320).