BioTopLite: An Upper Level Ontology for the Life Sciences. Evolution, Design and Application

Stefan Schulz[1],* Martin Boeker[2]

[1] Institute for Medical Informatics, Statistics and Documentation,
Medical University of Graz, Austria
[2] Institute of Medical Biometry and Medical Informatics,
University Medical Center, Freiburg, Germany

stefan.schulz@medunigraz.at
martin.boeker@uniklinik-freiburg.de

Abstract:

OBJECTIVE: To present and theoretically underpin recent changes in the upper level ontology BioTopLite2 and to discuss the effectiveness of upper level ontologies in the design process of domain ontologies.

BACKGROUNDs: BioTop is an upper level ontology for the life sciences, based on OWL DL. Experiences with the application of BioTop and changing requirements have required the introduction of a light version (BioTopLite). The usefulness of upper level ontologies is controversial.

METHODS: This paper provides a survey over the evolution of BioTop, use cases, and lessons learnt. It presents the main features of a new version of BioTop, motivated by special domain requirements. In particular it is highlighted how the new version, BioTopLite2 addresses the problem of time-indexed relations between continuants, given the restriction to two-valued relations in OWL.

RESULTS: The new version is optimized to more user-friendliness by reducing the relations to a minimum. BioTopLite is available at *http://purl.org/biotop*.

1 Introduction

The development of ontologies is a tedious and error-prone process. Besides in-depth knowledge of the domain to be represented, ontology engineers should master the representation formalism, and be skilled in building and maintaining modular software artefacts following design specifications. Upper-level ontologies can be understood to guide this process and provide the developers with a sound framework they can rely on and re-use. Another prominent reason for the employment of upper-level ontologies is their standardizing nature which can guarantee for real interoperability of ontology on class and relation levels.

Although upper-level ontologies (ULOs) are often seen as domain-independent, the development of the two most well-established ULOs, viz. DOLCE [BM09] and BFO [GS04] had a focus set on certain areas like cognitive sciences for the former and natural science for the latter. Nevertheless, several ULOs have been created since the mid-nineties, focusing explicitly on biology and medicine. They include the GALEN upper level [RR04], the UMLS semantic network [Cr03], the OBO relation ontology RO [SCK05], GFO-BIO [HLP08], BioTop [BSS08], and the Semanticscience Integrated Ontology (SIO) [Du13].

However, ontology developers and users may wonder whether ULOs have a positive impact on the resulting artefacts or whether it only renders them excessively complex. In our view, this question cannot be answered from a single perspective. On the one hand, ontology quality has many facets (formal correctness, correctness of the representation, completeness, etc.) which are at least partly dependent on specific use cases. On the other hand, the actual usage and significance of an ontology may change over time and can be completely different from the indented use cases. Thus, ontologies should always be evaluated from a perspective broader than just the actual intended use case, taking into account reuse and interoperability as important goals.

The objective of this paper is to describe and assess BioTop, an upper-level ontology, currently being redesigned by the authors. Our assessment is done from different perspectives and takes into account theoretical considerations of its design, together with empirical evidence for its impact on the quality of resulting domain ontologies. We will discuss the rationale of domain ULOs in general and of BioTop in particular. A survey of the development of BioTop will be given, and BioTopLite 2, its most recent version, still in experimental phase, is introduced.

* To whom correspondence should be addressed: stefan.schulz@medunigraz.at

2 BioTop and BioTopLite: Evolution and Design

BioTop version 1 was launched in 2006 as an Upper Domain Ontology using the description logics dialect OWL DL. Its basic design had been inspired by the GENIA ontology for cell signalling, mainly used in natural language analysis. Initially, a series of fundamental design problems had been identified in GENIA. BioTop was designed to go beyond the scope of GENIA, in order to cover a broad range of categories relevant for application in all areas of life sciences.

BioTop was not intended to compete with established ULOs, but rather to integrate with them. Therefore, its developers created bridging ontologies to DOLCE, BFO, and RO, and left BioTop's uppermost hierarchical level deliberately flat. By this mechanism BioTop can be employed as a top-level layer for biomedical ontology without

constraining developers to a certain ULO. However, developers who like to base their ontology on DOLCE, BFO, or RO can combine them.

An important asset of BioTop has been its strong focus on constraining axioms, as an important mechanism for consistency checking, which at the time it was introduced had only been available for DOLCE but not for BFO and RO. This required full class definitions, which changed the initial scope of BioTop to the integration of more classes considered fundamental for the representation of biological entities. Due to its inspiration by GENIA, BioTop first emphasized cell biology. The attempt to provide full definitions led to a further expansion into the realm of biochemistry. Experiences from the @neurIST project [BSK07] and BioTop's use as a top-level ontology in the DebugIT project [SBB10] revealed performance problems, which were mitigated by factoring out most of the chemistry classes into a separate ontology named ChemTop. This module was however not further maintained due to the re-emerging ChEBI ontology [HMD13], which underwent a thorough redesign following the OBO Foundry criteria [SAR07].

To integrate with the large corpus of terminologies provided by the National Library of Medicine, BioTop was aligned with the UMLS semantic network (SN) [Cr03]. This effort included a manual translation of the SN into OWL, with most of the semantic relations of the Semantic Network represented as reifications under *BioTop:Process*. The resulting ontology showed, again, considerable performance problems, so that its intended use for validating UMLS sources had to be postponed. However, the task of covering the whole content of SN provided a good external criterion of drawing a crisp boundary around BioTop [SBH09].

Severe performance issues with BioTop motivated the creation of a “lite” version, which included enough classes, relations, and axioms, in order to address the needs of most life sciences ontologies and, nonetheless, to provide a sound framework and guidance for developers. This version was then released as BioTopLite. It was used in several experimental ontologies by IHTSDO[1] working groups, in which future evolutions of SNOMED CT were tested.

An important design decision of BioTop and BioTopLite addresses the inherent ambiguity of medical terms: “Fracture” may denote both a fracturing process as well as its result, a fractured bone. “Allergy” can be interpreted as an allergic disposition or as allergic manifestation. Such categorial distinctions (as, e.g. proposed by OGMS [CS10]) are often not reflected neither in physicians' discourse and reasoning nor in medical terminologies, and for many clinical reasoning patterns a distinction is not necessary: A fracture of the neck of the femur is a femur fracture, regardless of whether fracture is seen as a process or a material entity. As a result, we added the disjunctive class *Condition* subsuming the classes *Material entity*, *Process*, and *Disposition*. Although this decision breaks the principle of non-overlapping classes on the first hierarchical level it is justified by the requirement of dealing (and reasoning) with ontologically heterogeneous and ambiguous terms in the clinical domain.

[1] http://www.ihtsdo.org

3 Evaluation of BioTop and BioTopLite

Still, no satisfying solution to quantitative or qualitative evaluation of ontologies is available. Even more limited is the situation for ULOs. Therefore, this section is mainly based on practical experiences with the development of ontologies designed for a certain purpose. Quantitative empirical evidence for the superiority or inferiority of the development based on a certain ULO or without any ULO is not available to the best of our knowledge.

Parts of BioTop were used to develop an ontology for the @neurIST project, intended to support the diagnosis and treatment of cerebral aneurism in a distributed environment. In a similar environment BioTopLite was also used as a ULO in the project DebugIT which provided a complex tool chain for the diagnosis and treatment of nosocomial infections.

BioTopLite was intensively used as a reference upper level ontology in GoodOD (Good ontology design), a project in which a comprehensive guideline for good practice ontology design was developed [BJG13]. This guideline was implemented in educational resources and tested in a curriculum with 24 students. As a result of the experiments, a large set of OWL files were collected (14 * 24 = 336), which provided insight into the problem-solving capabilities and typical errors of the test persons when challenged by modelling tasks from the biomedical domain. Both the analysis of the data and related observations shed light on obvious weaknesses of BioTopLite, but also provided positive feedback to the developers in the sense of enforcing principles:

- Pragmatic realist view of the domain to be modelled, in which a strict division is made between individual and classes. Classes are no more than sets of individuals, for which necessary and sufficient conditions can be added by means of object properties.
- Agnostic stance with regard to the existence of universals. Although being crucial from a philosophical point of view, it would challenge the modellers' understanding without creating additional benefit.
- Pragmatic approach regarding time: all classes have to be rigid; quantifications are implicitly assumed to hold for all instances in time.
- Compulsive use of top-level classes: domain classes must be placed under these classes, but not parallel to them.
- Flat hierarchy, no top-level classes like *Continuant, Occurrent,* etc. Although the continuant / occurrent distinction (analogous to BFO) underlays BioTop, it is not made explicit as it would confuse rather than help the user.
- Closure of relations (object properties): No additional object properties should be introduced. Relational predicates not covered by existing object properties have to be represented in a reified form as subclasses of Process.
- Continuous consistency checking at design time: The use of a DL reasoner after each modelling step prevented undetected design decisions that violated in-built constraints and would lead to inconsistent and difficult to repair domain ontologies.

The simplicity of these criteria contributed to a quick learning curve. Nevertheless, weaknesses were recorded such as the complexity of the relation hierarchy, as well as certain names which were difficult to understand, such as '**locus of**' or '**inheres in**'.

Since 2011 BioTop has been used in the CELDA project, an ontology of cell types, in vitro as well as in vivo, based on species, anatomy, subcellular structures, developmental stages and origin [CE13].

BioTopLite is currently being used as a top-level in the SemanticHealthNet project [SHN13], which aims at an ontology-based integration of heterogeneous semantic resources to create interoperability between data in the electronic health record. A focus is here on the relationship between information entities and (types of) clinical entities. SemanticHealthNet also takes up the result of extensive analyses of the ontological commitment of SNOMED CT finding/disorder classes, together with ICD classes in a joint activity of the WHO and the IHTSDO on the harmonization between these two terminology systems [RSR13]. It had been found out that the meaning of disease terms and the concepts resp. classes attached to them refer much more to clinical situations as segments of a patient's life in which a clinical condition is fully present than to the clinical conditions themselves as pathological or pathophysiological entities.

These findings on the impact of ULOs are limited; however, they provide some insights which at least can justify the use of ULOs in the development of domain ontologies, particularly by facilitating consensus within teams building domain-level ontologies. To provide the community with tangible evidence for the effectiveness of ULOs in ontology design elaborate experiments would be necessary. In the view of the authors, the current quantitative instruments for the evaluation of ontology are not sufficient to measure those differences reliable.

In addition to benefits on the quality of ontologies there are other important positive effects of an ULO which will be discussed here from the perspective of BioTop and BioTopLite2.

ULOs are supposed to warrant standardization and interoperability. Due to a set of well-defined, mutually disjoint upper-level categories and a set of relations to be regarded as close to complete, ontologies are easier to standardize, and to make interoperable when derived from upper-level ontology. BioTop and BioTopLite address this goal by providing a core set of classes and relations for representation of all areas of the life sciences.

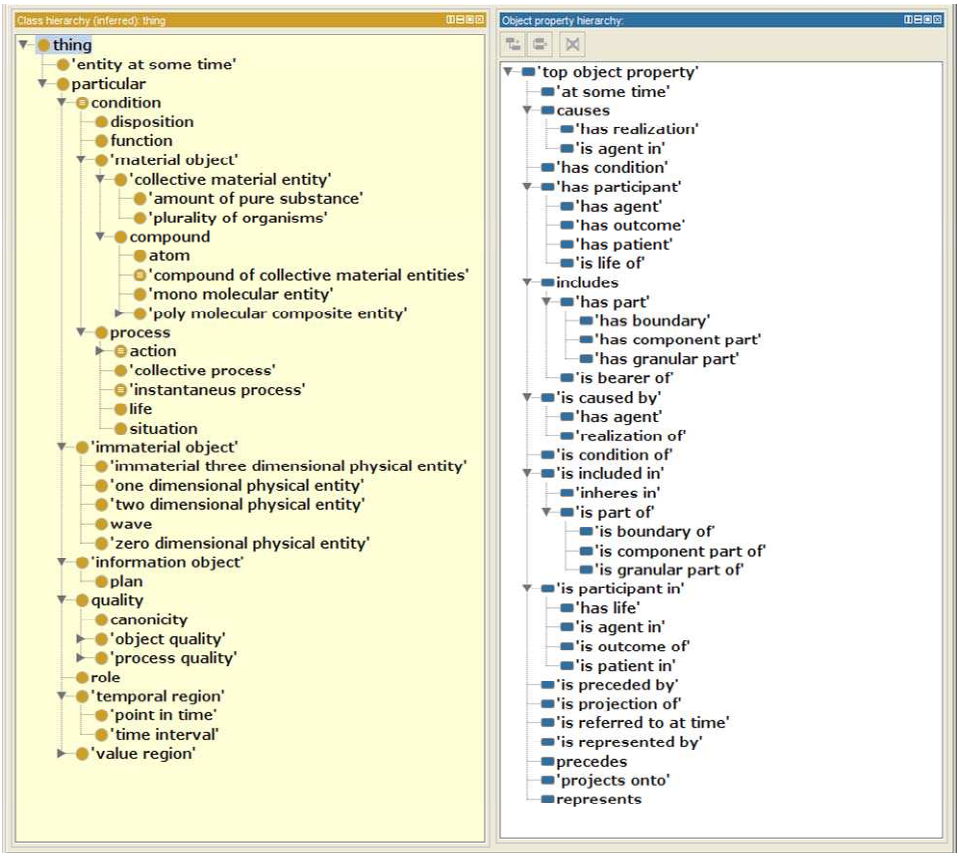

Fig. 1: Most classes (inferred view) and all relations (object properties) in BioTopLite 2 as Protégé 4 screenshots

4 Additional features and modification of BioTopLite2

These and other factors have motivated a major redesign of BioTopLite, viz. additional classes deriving from current use cases, the demand for a simpler hierarchy of object properties, the demand for more intuitive labels, and a principled approach to the representation of time-relevant entities. Figure 1 visualizes most of the BioTopLite 2 classes and relations. In the following, we briefly describe the changes as compared to the predecessor version.

Additional classes

Use cases from current biomedical terminologies suggested to include the class *Life*, the process in which an organism is involved from birth to death. The meaning can be broadened to all material entities, but also to immaterial and information entities. In

medical diagnosis, the references are often time segments of a (biological) life, during which a condition exists, as empirical investigations have shown [SRR12]. For instance, the term "gastric ulcer" would therefore refer to the life segment, called "clinical situation" in which a gastric ulcer process unfolds, or in which a gastric ulcer structure is present. We have therefore added the classes *Life* and *Situation* to BioTopLite2, which has now 53 instead of 49 classes.

Simplified relation (object property) hierarchy

The first BioTop version distinguished between process parts and object parts, as well as between parts and proper parts. It had turned out that this complicated the use of the ontology. In the new version, there is only one relation part '**has part**' / '**is part of**'. Thus, the number of relations was reduced from 51 to 37, despite some additional ones, which connect the classes *Life* with *Material Object*, as well as *Situation* with *Condition*. These relations, 'is life of' and 'has condition' are shortcuts to be used in simple axioms that substitute more complicated ones, which cannot be fully expressed in DL.

Substitutions of Domain / Range axioms

The simplification of the relation hierarchy resulted in the fusion of different relations (such as part-of between processes and between objects). The necessary constraints cannot be fully expressed by domain / range restriction axioms, e.g. that '**is part of**' cannot obtain between a process and a material object. It was therefore decided to refrain from the use of domain / range axioms. Instead, constraining axioms were included at the class level and as general class inclusion axioms, see Fig. 2. The total number of axioms grew from 530 to 572.

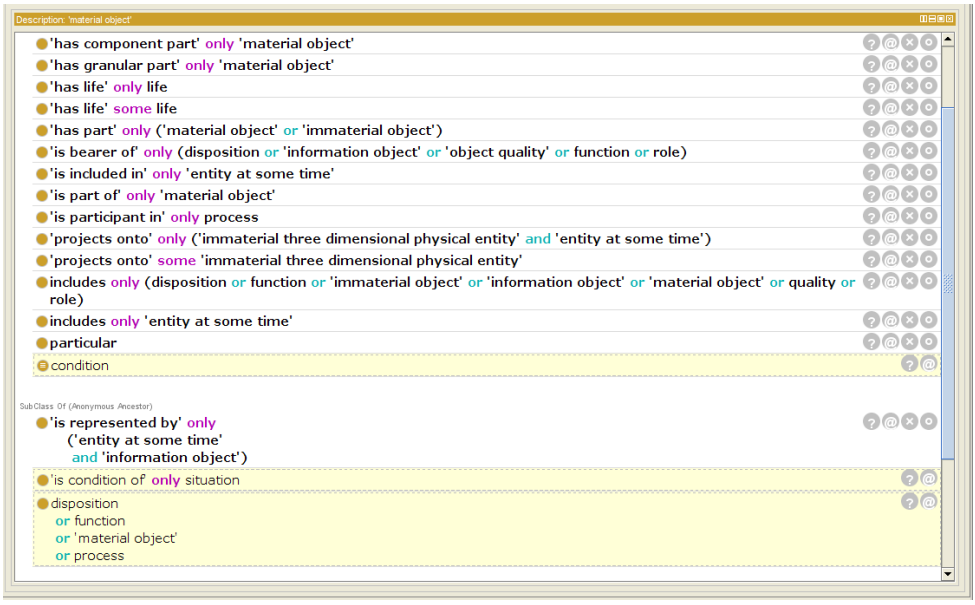

Fig. 2: Example of class-level axioms in BioTopLite 2 in Protégé, specifying implications and constraints for the class '*Material Object*'. The numerous value restrictions are the price to be paid for the parsimony of relations

More intuitive labels

Whereas the class labels remained (roughly) the same, relation labels were changed towards better intelligibility. For instance, the relation pair 'has locus' / 'locus of' was changed into 'is included in' and 'includes'. Linguistically, all relation labels correspond to verb phrases, consisting of full verbs in present tense ('**includes**'), partly with preposition ('**projects onto**'), or the auxiliary verb forms '**has**' or '**is**', followed by a complement, such as '**is part of**' or '**has life**'.

Representation of time relevant entities

In three-dimensionalist ontologies such as BioTop, but also BFO and DOLCE, a known issue is the representation of relations between continuants, i.e. objects that exist during time, undergo temporal change and have no temporal parts like processes or time intervals. OWL-DL does not account for the representation of time. This has two severe consequences, viz. that both instantiation and relations are not temporally qualified. In the first case this provokes ambiguities when the same individual instantiates different (non-rigid) classes at different times. For instance, we may want to express that the classes *Butterfly* and *Caterpillar* are disjoint and that an individual x first instantiates the first and then the second class. As we cannot express time and would therefore create an inconsistent ontology:

x rdf:Type *B*
x rdf:Type *C*
B subClassOf not (*C*)

A similar problem arises with relations between individuals: Assuming y receives x's kidney k as a transplant, the assertions:

x '**has part**' k
y '**has part**' k

would imply that k is both part of x and y, from which one could draw the wrong conclusions that both bodies overlap.

BioTopLite 2 mitigates the lack of ternary, time-dependent relations in OWL-DL by introducing time-dependent entities. The class *Entity at some time* is of no real ontological relevance but it has proven useful as a means to enforce that instances of time-dependent classes be placed in a temporal context. Class-level axioms are such that the reasoner infers that the relata must be of the type *Entity at some time*, e.g.:

'*Material object*' subClassOf '**is included in**' only '*Entity at some time*'

Instances of '*Entity at some time*' are related to a temporal reference by the relation '**is referred to at time**', and the relation between an atemporal entity and its temporalized "snapshot" is expressed by the relation '**at some time**'. The above example (see also Fig. 3) could then be expressed as

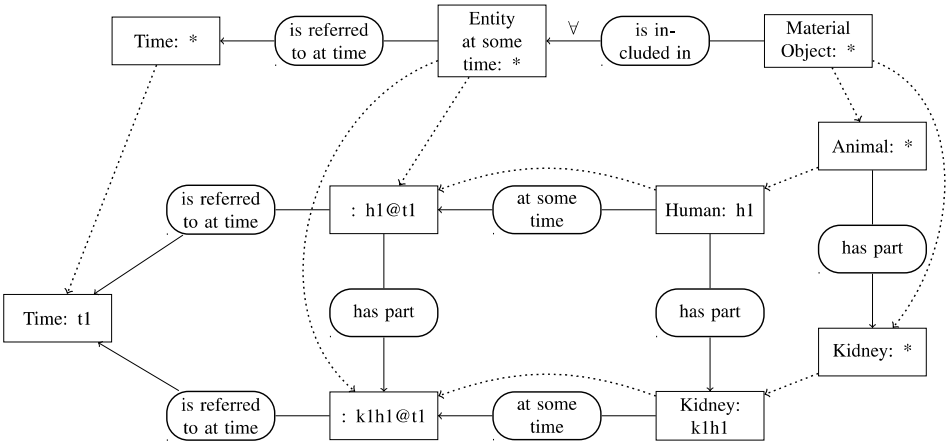

Fig. 3: Conceptual graph of temporally qualified entities. The DL class definitions are depicted on the top and on the left. In the middle, a homomorphic diagram for individuals is embedded.

h1@t_1 '**has part**' k1h1@t_1
h2@t_2 '**has part**' k2h2@t_2

with h1@t_1 and h2@t_2 being individuals in both classes *Human* and '*Entity at some time*', whereas k1h1@$t_1$and k2h2@t_2 are individuals in both classes *Kidney* and '*Entity at some time*'. Additionally the following assertions hold:

h1@t_1 ' **is referred to at time**' t_1
h2@t_2 ' **is referred to at time**' t_2
k1h1@t_1 ' **is referred to at time**' t_1
k2h2@t_2 ' **is referred to at time**' t_2
h1 '**at some time**' h1@t_1
h2 '**at some time**' h2@t_2
k1h1 '**at some time**' k1h1@t_1
k2h2 '**at some time**' k2h2@t_2

The introduction of temporally qualified entities grants more flexibility regarding the description of classes, in case it is a necessary criterion that something was related at some time. For instance, the axiom

'*Structured biological entity*' subClassOf
'**at some time**' some ('**is part of**' some *Organism*)

expresses that for each atemporal instance of '*Structured biological entity*' there is at least one temporal phase in which it is part of some Organism. This is the BioTopLite 2 way of expressing 'parthood at some time', which is intransitive, in contrast to parthood without reference to temporal phases such as

'*Cell nucleus*' subClassOf '**is part of**' some *Cell*

Since the relation '**is part of**', in BioTopLite 2, is constrained to obtain either between processes or temporally qualified objects, it must be interpreted as

'*Cell nucleus at some time*' subClassOf '**is part of**' some '*Cell at some time*'

This means that for each cell nucleus at any time there is some cell it is part of. This comes close to generic parthood (always part of some entity of a certain kind but not necessarily the same entity), as axiomatised in the OBO relation ontology as the meaning of the (transitive) class-to-class relation *A Part-of B* [SCK05]. The extension of BioTopLite in this way increases its complexity. However, general modelling with BioTopLite is not altered by this extension and provides developers with additional expressivity. Although complex, the constellations between temporal classes can be used in the form of few patterns by ontology developers.

4 Conclusion

Seven years' experience with different versions of BioTop used in different projects has shown the need for adaptation of an upper level ontology to the user's context. This does not mean to abandon fundamental axioms but rather the provision of additional, domain-specific classes and relations. Inherent ambiguities with in a domain's discourse can be addressed by disjunctive classes and shortcut relations. Only the former ones can be easily defined inside OWL-DL, whereas the latter ones require a richer logic.
The new version of BioTopLite addresses the problem of time-indexed relations, for which OWL does not provide a straightforward solution. Our proposal is to regard instantiations of continuants as inherently time-indexed, which is enforced by the new BioTopLite class *Entity at some time*. This approach allows not only for eliminating ambiguities in the instantiation of non-rigid classes. It also offers a straightforward pattern to distinguish between those relationships that hold at some time and those which (generically) hold at all times. The proposed solution is still experimental and requires more in-depth theoretical elucidation and feedback from ontology engineering practice.
We presented some findings on the impact of BioTop and BioTopLite on the quality of ontologies; although these results cannot easily be generalized, they provide some insights which at least can justify the use of ULOs in the development of domain ontologies, particularly by facilitating consensus within interdisciplinary teams building domain-level ontologies. To provide the community with tangible evidence for the effectiveness of ULOs in ontology design further research is necessary.

Acknowledgements

This work was partly supported by the Deutsche Forschungsgemeinschaft (DFG) within the GoodOD project (Good Ontology Design), grant JA 1904/2-1 and SCHU 2515/1-1.

References

[BM09] Borgo S, Masolo C. Ontological Foundations of DOLCE. In Staab S, Studer R (eds.), Handbook on Ontologies (Second Edition), Springer Verlag, 2009: 361-382.

[GS04] Grenon P, Smith B. (2004) "SNAP and SPAN: Towards Dynamic Spatial Ontology", Spatial Cognition and Computation, 4: 1, 69-103.

[RR04] Rector A, Rogers J. Patterns, Properties and Minimizing Commitment: Reconstruction of the GALEN Upper Ontology in OWL. Proc. of the EKAW*04 Workshop on Core Ontologies in Ontology Engineering, http://ceur-ws.org/Vol-118/

[Cr03] McCray AT. An upper-level ontology for the biomedical domain. Comp Funct Genomics. 2003; 4 (1): 80-84

[SCK05] Smith B et al. Relations in bio-medical ontologies. Genome Biology 2005; 6 (5): R46. Epub 2005 Apr 28.

[HLP08] Hoehndorf R et al. GFO-Bio: A biological core ontology. Applied Ontology, 2008, 3 (4), 219-227.

[BSS08] Beisswanger E et al. BioTop: An upper domain ontology for the life sciences. A description of its current structure, contents and interfaces to OBO ontologies. Applied Ontology, 2008; 3 (4): 205-212.

[Du13] Dumontier M (2013). The Semanticscience Integrated Ontology (SIO) http://code.google.com/p/semanticscience/wiki/SIO .

[BSK07] Boeker M et al. The @neurIST Ontology of Intracranial Aneurysms: Providing Terminological Services for an Integrated IT Infrastructure. AMIA Annu Symp Proc. 2007:56–60.

[SBB10] Schober D et al. The DebugIT core ontology: semantic integration of antibiotics resistance patterns. Stud Health Technol Inform. 2010;160(Pt 2):1060-1064.

[HMD13]Hastings J et al. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic Acids Res. 2013 Jan;41

[SAR07] Smith B et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology 2007 Nov; 25 (11): 1251-1255.

[SBH09] Schulz S et al. Alignment of the UMLS semantic network with BioTop: methodology and assessment. Bioinformatics. 2009 Jun 15; 25 (12): i69-76.

[CS10] Ceusters W, Smith B. Foundations for a realist ontology of mental disease. J Biomed Semantics. 2010 Dec 9;1(1):10

[BJG13] Boeker M et al. Effects of Guideline-Based Training on the Quality of Formal Ontologies: A Randomized Controlled Trial. PLoS ONE 2013 8(5): e61425.

[CE13] CELDA – An Ontology for the Comprehensive Representation of Cells in Complex Systems http://cellfinder.org/about/ontology

[SHN13] SemanticHealthNet Network of Excellence (2013). http://www.semantichealthnet.eu/

[RSR13] Rodrigues JM et al. Sharing Ontology between ICD 11 and SNOMED CT will enable seamless re-use and semantic interoperability. Accepted for MEDINFO 2013.

[SRR12] Schulz S, Rector A, Rodrigues JM, Spackman K. Competing interpretations of disorder codes in SNOMED CT and ICD. AMIA Annu Symp Proc. 2012; 2012: 819-827.