From GENIA to BIOTOP

Towards a Top-Level Ontology for Biology

Stefan SCHULZ [a,1], Elena BEISSWANGER [b], Udo HAHN [b], Joachim WERMTER [b], Anand KUMAR [c], Holger STENZHORN [c]

[a] *Department of Medical Informatics, Freiburg University Hospital, Germany*
[b] *Jena University Language and Information Engineering (JULIE) Lab, Germany*
[c] *Institute for Formal Ontology and Medical Information Science (IFOMIS), Saarbrücken, Germany*

Abstract. The increasing need for advanced ontology-based knowledge management in the life sciences is generally being acknowledged but, up until now, the development of biological ontologies lacks adherence to foundational principles of ontology design. This is particularly true of so-called upper-level ontologies such as the GENIA ontology which covers biological continuants and has mainly been devised for corpus annotation in a text mining context. As an alternative, we introduce BIOTOP, an upper ontology of physical continuants in the domain of biology, with a coverage similar to the GENIA ontology. We report on design specifications and modeling decisions for BIOTOP which are based upon formal ontology principles. As a major desideratum, these continuants are described in terms of necessary and sufficient conditions. We accomplished this goal for 85 out of the 146 existing GENIA classes. We use OWL-DL as a formal knowledge representation language and may thus use a terminological reasoner for classification in order to check and maintain consistency during the ontology engineering phase.

Keywords. Bio-Ontologies, Upper-Level Ontologies, OWL-DL

1. Introduction

The rapid increase of scientific knowledge in the life sciences has created an enormous need for advanced knowledge management in this field. As a consequence, many efforts have been devoted to develop description languages to help structure the knowledge of this domain. Whereas cell biology and genomics have only marginally been covered by the traditional clinical vocabularies (such as the roughly 100 sources made available by the Unified Medical Language System (UMLS) [17]), the development of the Gene Ontology [7] and, more generally, the Open Biomedical Ontologies (OBO) framework [13] have put the case of ontology development at the very top of their task agenda.

As with the UMLS, each OBO ontology is independently developed and provides a partial, highly focused view on biology and medicine, fueled by the specific interests of various ontology designers. OBO includes at present (July 2006) 58 ontologies covering

[1]Corresponding Author: Stefan Schulz, Department of Medical Informatics, Freiburg University Hospital, Stefan-Meier-Strasse 26, 79104 Freiburg, Germany; E-mail: stschulz@uni-freiburg.de.

cell types and components, the anatomy and development of several organisms (plants and animals), chemical entities, biological pathways and processes, molecular functions and others. The OBO ontologies, up until now, adhere to a rather simple design pattern: Nodes (called terms) are organized in directed acyclic graphs (DAGs) with labeled edges (relations) such as *Is_A*, *Part_Of*, *Develops_From* and others.

Most of the OBO ontologies were created in a completely informal and ad-hoc fashion which is likely to create conflicting and contradictory interpretations. For example, in the statement *A Part_Of B* (with *A* and *B* being OBO terms which we consider as referring to universals), the assertion that "some instances of A are part of some instances of B" is quite different from the assertion that "all instances of A are part of some instance of B" or that "all instances of B have and instance of A as part" [23,18]. A proposal has recently been made to provide consistent and unambiguous formal definitions of the relational expressions that ontologies in OBO [21] should adhere to.

The necessity of a generalized upper-level to support the interoperability between different domain ontologies and to enforce the consistency in the process of ontology construction and maintenance has been advocated by many researchers though this goal still has not been realized so far. Whilst several proposals for general-purpose upper ontologies exist (e.g., DOLCE [6] and BFO [22]) and are already subject to vivid discussions, this issue is not really on the radar in the biology domain.

Whereas BIO-BFO [8] and Simple Bio Upper Ontology [15] are sketched without any concrete application context, the GENIA upper ontology is most commonly used for the semantic annotation of texts by the biological text mining community. According to its designers, GENIA

> "is intended to be a formal model of cell signaling reactions in human. It is to be used as a basis of thesauri and semantic dictionaries for natural language processing applications such as information retrieval and filtering, information extraction, document and term classification and categorization. Another use of the GENIA ontology is to provide the basis for an integrated view of multiple databases. [24]"

The GENIA ontology limits itself to a set of highly general upper-level categories and is restricted to biological continuants. It contains 45 terms (called "classes") which are arranged in a tree-wise fashion at a maximum depth of 6 nodes. Besides the taxonomic relation *Is_A* it does not contain any further relations or definitory axioms. Instead, so-called "scope notes" informally phrase the meaning of the single classes as natural language statements [24]. As said above, the predominant application of the GENIA ontology targets semantic annotation of named entities in biological literature abstracts [14].

In this paper we propose a common upper ontology for biology and adopt the GENIA ontology as the starting point for its development. Taking different traditions of ontology development into account we define a set of best-practice principles and use them for a critique of the GENIA ontology as well as the subsequent design of a new upper ontology of biological continuants. The newly designed ontology is intended to facilitate the interoperability between existing biomedical ontologies, e.g., the Gene Ontology, ChEBI, the Mouse Ontology and other OBO ontologies, but also medical ontologies such as the Foundational Model of Anatomy (FMA) and SNOMED CT. Due to precisely defined axioms this newly created ontology has the potential to be more rigorous, consistent and valid than its precursors.

2. Methodology

2.1. Different Traditions of Ontology Design

One may distinguish three fundamentally different approaches to ontology design due to different traditions, interests and purposes. These different approaches still give rise to misunderstandings and often fruitless discussions. We refer to them as (i) the lexical-cognitivist, (ii) the philosophical-realist, and (iii) the computer science approach.

2.2. The Lexical-Cognitivist Approach to Ontology Design

Natural language constitutes the primary means of communication between domain experts, as used in scientific publications, textbooks, glossaries and dictionaries. The abstraction from word meanings is therefore the most natural way domain experts, such as biologists, chemists or physicians (generally lacking in-depth knowledge in philosophy, logics and computer science) tend to organize their domains of interest. Related to the methodologies developed by lexicographers and librarians, this approach is also supported by the cognitive science community which is more interested in describing the mental representation of reality rather than in the mind-independent reality itself. Prototypical features of concepts (as the entities of thought) therefore guide the enterprise of ontology construction. Evidence for this language and cognition centered view is the preference of the words "terms" or "concepts" for describing the nodes in an ontology, as well as the restriction to inter-concept relations which depict semantic association (of what "normally" has a good degree of plausibility) rather than subscribing to strict formal properties of the relational statements being used. A discussion of semantic underspecifications of concept-to-concept relationships is often regarded as some kind of sophistry. This position is also backed by philosophical positions which dispute the accessibility of a mind-independent reality.

2.3. The Philosophical-Realist Approach to Ontology Design

Regardless of inter-philosophical divergences (which are often difficult to communicate to the outside world), philosophers who dedicate themselves to formal ontologies generally build upon a millenary tradition of metaphysics and logics. Their endeavor of exactly describing entities of being in their essence generally requires a rich inventory of logical constructs. For many purposes, first-order logics is considered as insufficient for adequately describing reality. The claim of describing reality by logical statements is most decidedly raised by the Aristotelian tradition. Accordingly, classifying the world's entities in terms of their genera and differentiae is adopted as a fundamental guideline for the design of formal ontologies.

2.4. The Computer Science Approach to Ontology Design

Computer science has borrowed the term "ontology" from philosophy, using it preferably in the hitherto non-existent plural form. Here, ontologies are mainly conceived as computable abstractions of certain domains of interest, mainly driven by concrete application requirements. Traditionally, only little emphasis has been put on upper ontologies which

has somewhat changed with the advent of the Semantic Web. However, the view prevails that different ontologies represent different and, unfortunately, partly incompatible views of a given reality. Rather than focusing on upper ontologies, computer science ontologists tend to feel more challenged by the tasks of semantic mediation and brokerage. Another contrast to purely philosophical ontologists is the strong focus on computability. Therefore, higher-order logics and even full first-order logics are commonly discarded due to their high computational costs. The attempt of describing more tractable subsets of logic was one of the major driving forces of developing description logics [1].

2.5. Principles of Ontology Building and Critique

A reasonable starting point for the ontological analysis of the biological upper-level is given by the following principles [5]: (i) select a set of foundational relations, (ii) define the ground axioms for these relations, (iii) establish constraints across the basic relations, (iv) define a set of formal properties induced by these formal relations, (v) introduce the basic categories and classify the relevant kinds of domain entities accordingly, and, finally, (vi) elicit the dependencies and interrelations among the basic categories. In our case, most of these basic categories are borrowed from the upper ontologies BFO [22] and DOLCE [6] enriched by principles introduced by Rector *et al.* [16]. Accordingly, we adopt the generally accepted, mutually exclusive divisions between universals and particulars on the one hand, and between continuants and occurrents on the other. Particulars (individuals) are the concrete and countable entities in the world (e.g., "my hand") whereas universals are entities which are instantiated by particulars (e.g., "hand"[2]). Orthogonal to this dichotomy, a fundamental distinction between continuants and occurrents is also commonly introduced. The GENIA ontology has no explicit category for occurrents[3] and hence its focus is put on the representation of continuants.
Furthermore we subscribe to the canonical relations[4] recently adopted by OBO [21]: **Instance_of** relates an individual entity to a certain class. *Is_A* relates two classes in terms of taxonomic subsumption. The relation **part_of** and its inverse **has_part** relate individuals in terms of parthood.[5] Furthermore, **derives_from** holds between an individual which was either identical or part of another individual at some instant in time. Finally, **has_function** and its inverse **inheres** hold between individual material entities (such as molecules) and their inherent (biological) functions. As a subcategory of dependent continuants we introduce here the important notion of *biological function*. Although function is not addressed directly by the current state of the GENIA ontology, it will prove necessary for a complete definitory framework of GENIA classes.

---

[2] In the context of this paper the term *universal* will be considered synonymous with the terms *class* and *type*. We refrain from the use of the term *concept* due to its multiple, partly contradictory senses. Our distinction between universals and particulars is made explicit by strict naming conventions: names of universals use *Upper Case* initials, while names of particulars are written in *lower case* letters.

[3] In practice, annotators have been using the residual category "other" for tagging occurrents.

[4] We use the following naming conventions: Relations in which one or more individuals are involved are expressed by means of **bold face expressions and lower case initials**. Relations involving classes only come with *Upper Case Initials and Italic Fonts*.

[5] We understand parthood as proper parthood in the sense of formal mereology [20], i.e., a transitive, irreflexive and asymmetric relation.

2.6. Analysis and Reconstruction of GENIA

Our approach to design a new ontology covering the existing GENIA classes rests on the following steps:

1. We analyze each GENIA "scope note" in terms of its definitory value, both under an intensional (i.e., the definition) and an extensional (i.e., the subordinate classes) point of view. We hereby focus on how the linguistic expressions contain sufficient information to delimit the meaning of the associated term and the extension of the class it refers to.
2. Under the assumption of the current GENIA ontology being a taxonomy we analyze it with regard to proper classification principles. Keeping in mind that a major purpose of GENIA is to unambiguously assign exactly one semantic label to each text entity under scrutiny, this requires a mono-hierarchical classification tree with pair-wise disjoint and exhaustive classes at each classificatory level.
3. We logically redefine the classes, exploiting both the associated scope notes and canonical biological knowledge. As we are aware of the fact that a comprehensive ontological account often requires a highly expressive language, we do not a priori impose any restriction on that language. However, wherever computationally expensive formalizations result, we transform them into a simplified representation using OWL-DL, according to the preferences of the computer science approach to ontology implementation. The expressivity problems can most likely be solved by integrating rules through the Semantic Web Rule Language (SWRL) [11] in our BIOTOP implementation. This framework built on top of OWL-DL allows to combine class definitions with rules and, by doing so, makes it feasible to express complex facts that cannot be expressed using class definitions alone. A caveat is that the rules must be applied carefully to avoid excessive computational costs. If applied with care, however, they can certainly improve the existing coverage of the domain. Hence, their use will be investigated as a future step in the development of BIOTOP.
4. A major requirement rarely met by any existing biological ontology is the introduction of true definitions. This means that both the necessary (i.e., getting from a class to its conditions) and the sufficient conditions (i.e., getting from the conditions to a specific class) for class membership which need to be described. The latter is one of the main requirements in order to fully exploit the inferential power of description logic reasoners such as RACER [10]. Machine reasoning is then used for checking the logical consistency of the ontology. Any inconsistency found will then require additional change iterations. We expect that abstraction from full first-order logic will lead to a loss of expressivity which we intend to counterbalance by the introduction of auxiliary constructs.
5. The interfaces to existing ontologies such as the Gene Ontology, CHEBI, etc. are identified. Besides this, the new ontology should exhibit a sufficient granularity and coverage to support a mapping to the classes of the GENIA ontologies without ambiguities. This would meet the requirements of the text mining community for which GENIA has evolved as a kind of a quasi standard.

3. Analysis of GENIA

3.1. Analysis of Scope Notes

A general impression of the scope notes is that besides cursory hints to related terms, they do not contain sufficient definitory information. A reason for this may be that the annotators using GENIA were too familiar with these terms and hence believed that no additional information was required. Summarizing some of the typical shortcomings, Table 1 reveals that only a quarter of all classes are fully defined by their scope note. Half of the GENIA classes are incompletely described by just enumerating their subclasses or listing examples. Yet another quarter does not even have a scope note.

3.2. Analysis of GENIA's Ontological Structure

A formally correct taxonomic classification is done on the basis of the ontological nature of the entities. Classes in an ontology stand for universals (or logical expressions denoting universals), whilst instances correspond to entities which cannot be instantiated [5]. Whereas it is straightforward to assume classes such as organism, cell, individual DNA (desoxyribonucleic acid) molecule to be instantiated by concrete entities (e.g., "this individual cell under this microscope"), we also observed numerous oddities which arise with regard to other classes such as source, cell type, tissue, protein family or group. identified the following kinds of classes which require deeper ontological inquiry.

3.2.1. Source and Substance

The division between "Source" and (chemical) "Substance" constitutes the uppermost partition of the GENIA ontology. Whereas "Substance" refers to chemical substances involved in biochemical reactions, "Sources" are defined as "biological locations where substances are found and their reactions take place". They are subdivided into natural (such as organism, cell) and artificial sources (such as cell line). As much as it may be acceptable that for specific purposes biological objects are not distinguished from the space they occupy, biological location can hardly be accepted as a suitable upper-level distinction. For example, "Natural Source" subsumes different kinds of entities (cell, cell component) which also occur in artificial sources, e.g., cell lines. Our suggestion is therefore to treat "Source" as a role and not as top-level class.

Feature	Occurrences	Class	Scope Note
No Definition	11	Carbohydrate	
Examples Only	18	Amino Acid Monomer	An amino acid monomer, e.g., tyr, ser
Partial Definition	2	Artificial Sources	Cultured, immortalized or otherwise artificially processed sources
Full Definition	10	Domain or Region of Protein	A tertiary structure that is supposed to have a particular function, e.g., SH2
Enumeration of Subclasses	4	Organism	Organisms include multi- and mono-cell organisms

Table 1. Analysis of GENIA scope notes

3.2.2. Cell Type

"Cell Type" occurs as a sibling of "Organism" and "Tissue" and is vaguely described in the corresponding scope note as "a cell type, e.g., T-lymphocyte, T-cell, astrocyte, fibroblast". Here the question arises whether the attribute "type" is merely a notational flavor or conveys an additional meaning, e.g., a metaproperty instantiated by universals instead of individuals [5]. An instance of "Cell Type" would therefore not be an individual cell but rather a universal such as "Fibroblast" or "Leukocyte". But in turn this argument would equally justify the creation of classes such as "Tissue Type" or "Organism Type". In any case, such classes would specialize the class "Natural Source" since sources are defined as biological locations and a "Cell Type" is definitely not a biological location. Hence we suggest to ignore the meta-level reading and read "Cell Type" as "Cell".

3.2.3. Family or Group

A similar problem can be found with classes labeled "Family or Group" (in the DNA, RNA and Protein branch) defined by GENIA as "a family or a group of proteins, e.g., STATs[6]". Such a class definition addresses the need of a reference to instances of a human-made classification scheme for proteins rather than to instances of biological classes. That again, would correspond to a meta-class reading leading to conflicts with the parent classes "Protein" and "Substance". We may argue that such classification schemes follow biological functions, locations and other roles (e.g., structure proteins, enzymes, or transport proteins) and because of this an account for this phenomenon by a separate branch of the ontology (e.g., "Role", "Function", "Entity of Classification") would be required.

3.2.4. Other

Residual categories, although repeatedly criticized [3,2], are characteristic for classification systems since they allow for an exhaustive, non-overlapping coverage of a given domain even for those entities which do not fall into the properly defined categories. GENIA's use of residual categories (e.g., "Other Natural Source", "Other Organic Compound") is however quite inconsistent because residual categories are only present in some partitions but missing in others (e.g., "Natural Source"). Although residual classes are ontologically irrelevant (i.e., their instances lack a common property), they can nevertheless be formalized as the logical complement to the union of their siblings. However, they may be misused for classifying those instances which are simply underspecified due to missing information and hence degrade the quality of classification.

3.2.5. Masses, Aggregates and Collectives

Many kinds of biological and chemical entities occur as collectives of uniform objects (e.g., cell collections or H_2O molecules). More complex aggregations of cells and intracellular matrices are present in biological tissues. A prototypical example is "Tissue", described in GENIA as "a tissue, e.g., peripheral blood, lymphoid tissue, vascular endothelium". That is not a proper definition but merely an enumeration of possible subclasses. For instance, "Tissue" in a biological context denotes an aggregate of cells and intracellular substances. Due to this fact it is not clear what exactly is an instance of

[6] Signal Transducers and Activators of Transcription

"Tissue". The main difficulty here is to make a clear commitment to the referents of such mass or collection terms. In principle, there are good arguments to refer to either (i) the totality of the mass/collective (e.g., all red blood cells (RBCs) in an organism), (ii) any portion of it (e.g., the RBCs in a lab sample) or (iii) the minimal constituent (e.g., a single RBC). So far there is no biological ontology which sufficiently accounts for the distinction between single objects and collectives.

4. Design of the BIOTOP Ontology

The design of BIOTOP (Biological Top-Level) was done by two of the authors with good knowledge in description logics as well as molecular biology. For ontology engineering, we used the Protégé ontology editor [12] supported by the RACER terminological reasoner [10] for consistency checking. This framework required a restriction to the OWL-DL language specification. BIOTOP contains a total of 146 classes (85 fully defined), 12 relations and 171 restrictions. The ontology successfully classifies on a middle-end laptop computer in about four minutes. It is available for download from `http://morphine.coling.uni-freiburg.de/~schulz/BioTop/BioTop.html`. In the course of engineering the BIOTOP ontology, several design decisions were taken which we discuss next.

4.1. Relations

In addition to the class-level taxonomy-building *Is_A* relation, we introduced the mereological relations **proper_part_of** and **has_proper_part** which relate individuals. Although the OBO relations proposal prefers the reflexive reading (e.g., "my body is part of itself") [21], we adopt the irreflexive variant for two reasons. Firstly, reflexivity is counterintuitive in biology since the common language use of 'part' excludes identity. Secondly the OWL-DL language specification does not support reflexive relations. Just as proposed by Simons [20], taking **proper_part_of** as a primitive is just a matter of convention. The relations **proper_part_of** and **has_proper_part** are subrelations of **located_in** and **location_of**, respectively [21]. The refining criteria for distinguishing **proper_part_of** from **located_in** are complex and discussed in [19]. Two subrelation pairs of **has_proper_part** were introduced, *viz.* **has_grain** and **grain_of** (according to [16]) as well as **component_of** and **has_component**, respectively, both relations being intransitive. The relation **has_grain** allows for the definition of collectives (i.e., amounts of cells, molecules, etc.) in terms of their constituent objects. The relation **has_component** relates compounds to their constituent components. An example for this is the relation between a protein chain and its constituent amino acid monomers. The criterion for the assignment of this subrelation is based on the notion of a partition: all parts related by **has_component** are mutually non-overlapping and sum up to the whole entity. We can formally deduce this relation from **has_proper_part** as follows (using $\sum$ for the mereological sum [25] and the RCC relations **po** for proper spatial overlap and **dc** for spatial disconnection [4]):

$$\textbf{has\_component}_P(a, b_0) \leftrightarrow \tag{1}$$

$$\exists a, b_0, ..., b_n : \bigcap_{\nu=0}^{n} \textbf{has\_proper\_part}(a, b_\nu) \cap \left[\bigcap_{\nu=0}^{n-1} \bigcap_{\mu=\nu+1}^{n} \neg \textbf{po}(b_\nu, b_\mu) \right] \cap \sum_{\nu=0}^{n} b_\nu = a$$

The relation **has_grain** can be formalized in a similar way:

$$\textbf{has\_grain}(a, b_0) \leftrightarrow \exists a, b_0, ..., b_n : \bigcap_{\nu=0}^{n} \textbf{instance\_of}(b_\nu, B) \cap \tag{2}$$
$$\bigcap_{\nu=0}^{n} \textbf{has\_proper\_part}(a, b_\nu) \cap \left[\bigcap_{\nu=0}^{n-1} \bigcap_{\mu=\nu+1}^{n} \textbf{dc}(b_\nu, b_\mu) \right] \cap \sum_{\nu=0}^{n} b_\nu = a$$

Whereas a compound's sortal identity depends on the exact sum of its components, a collective identity does not. If one removes a single blood cell from a given blood sample then the type of the sample still remains the same. But if a nucleotide is removed from a gene sequence then it instantiates a different type. Another criterion is that grains unlike components are not spatially connected. However, this requires a clear-cut conceptualization of connection. Another difference between grains and components can be found in the relation between components and compounds depending on a partition (see subscript *P* in formula 1). There may be different ways to dissect an entity into compounds. Consider a human skeleton which is normally partitioned into its 206 bones. A more coarse-grained partition (e.g., considering skull and pelvis single components), however, is also possible. Also, a DNA sequence can be partitioned either into nucleotides or into tri-nucleotide units called codons with each coding for a single amino acid. Finally, the arrangement of components is fundamentally relevant to the nature of the compound, whereas the arrangement of grains is irrelevant for the collective. (This issue is not considered in the above formula.)

Since it is not possible to directly translate the above formula into OWL-DL, these considerations need to be added via primitive classes. Future versions of the BIOTOP ontology may discard those primitive classes and instead apply SWRL rules at this point.

4.2. Collectives

The introduction of collectives as classes of their own, in contrast to their constituent objects, is justified by the ontological difference between these two kinds of entities and the referential ambiguity which can commonly be observed in texts. From a cognitive point of view, a distinction between masses and collectives is plausible, since humans perceive them in a different way and therefore use different language constructs (e.g., "some blood", "du sang", "Blut" vs. "erythrocytes", "des érythrocytes", "Erythrozyten"). This is the reason why DOLCE makes an ontological distinction between "Collection" and "Amount of Matter". We consider such a distinction arguable since is depends on the scale of granularity and type distinction. Due to the atomicity of matter, actually any amount of matter can be described as a collective of particles. We even refrain from an upper distinction between collectives and count entities because any material continuant can be regarded as a collection of elementary particles.

4.3. New Classes

In order to (at least partly) fulfill our objective of describing ontology classes in terms of full definitions, we introduced additional classes, many of which are only textually addressed in the GENIA scope notes. An example of this is the class "Particle". It was

originally meant to represent the classical notion of molecule or atom as constituent of matter. As a property of such a class we required that it should not be homomerous, i.e., no part of a particle itself should be a particle. Classifying the ontology under this constraint immediately led to a series of inconsistencies. A closer analysis of chemical entities revealed that it is indeed highly problematic to classify chemical entities in terms of unity [9]. Whereas at the level of small molecules this could still be accounted for by additional subdivisions (e.g., amino acid molecule vs. amino acid residue) this is nearly impossible for the domain of macromolecules in which several flavors of chemical bonds (i.e., hydrogen bonds, polar bonds and ionic bonds) are responsible for a broad and continuous range of cohesive forces. We therefore dropped the notion of a whole and consequently the requirement of non-homomerity for particles.

A further example of a newly introduced class is "Heterocyclic Base" which is used for the definition of "Nucleotide". Compared to other ontologies, the number of fully defined classes (i.e., definitions in terms of both necessary and sufficient attributes) is quite high. Interestingly, there are no such definitory statements in any of the current OBO ontologies.

4.4. Rearranged Classes

Some classes in the original GENIA ontology are misleading. For instance, "Amino Acid" subsumes any compound which contains amino acids though the term is regularily used for amino acid monomers. Hence we introduced the classes "Amino Acid Monomer" and "Amino Acid Polymer" in order to avoid confusion. Generally, there seems to be a major confusion in the domain concerning monomers, polymers and subdivisions of polymers. The prototypical example for this is DNA. According to the GENIA ontology, the term DNA refers to one or more of
1. a DNA monomer constituted by a base, desoxyribose and a phospate residue;
2. one polymer constituted by DNA monomers, bound together by covalent bonds;
3. two complementary strands of DNA polymers (cf. 2), joined by hydrogen bonds;
4. any subdivision of item 2 or 3, provided it is made up of more than one DNA monomer.
In BIOTOP we therefore made a sortal distinction between DNA monomer (according to item 1), full DNA (according to item 2) and DNA which corresponds to item 4. Double strands are considered to be of different types.

4.5. New Branches

As already pointed out, the "Family or Group" categories from the original GENIA ontology are improperly arranged in the hierarchy. In GENIA these categories were included to denote terms such as "enzyme" or "membrane protein". In a statement such as "the enzyme E", "enzyme" refers to a biological function whereas "E" refers to an amount of molecules. What is meant here is that "E" exercises the function "enzyme". In order to account for this peculiarity we introduced an additional branch named "Non-Physical Continuant" which subsumes "Biological Function" together with "Biological Location". Just as in the GENIA ontology, BIOTOP does not elaborate on biological processes, events, or actions. In the current version it only contains one single class named "Occurrent". An enhancement towards a more detailed description of this kind of entities will constitute an important issue of future work.

4.6. Mapping to GENIA

In order to guarantee downward compatibility, the original GENIA ontology was added as an additional layer, in a separate step. To this end, all terminal GENIA nodes (i.e., those which are used for semantic annotation) were added as jointly exclusive classes and linked to the BIOTOP classes by *Is_A* relations. Consistency is assured by applying the terminological reasoner.

4.7. Interfacing with Other Ontologies

Several BIOTOP classes can be used as links to other existing ontologies. For example, "(Bio)Molecular Function", "Cellular Component" and "Biological Process" provide links to the homonymous branches of the Gene Ontology. The same can be applied to the CHEBI ontology. "Molecular Function$_{BioTop}$" interfaces with "Biological Role$_{ChEBI}$", "Atom$_{BioTop}$" and "Compound$_{BioTop}$" with "Molecular Entities$_{ChEBI}$" and "Subatomic Particles$_{BioTop}$" with "Elementary Particles$_{ChEBI}$". "Organism$_{BioTop}$", "Tissue$_{BioTop}$" and "Body Part$_{BioTop}$" can finally be linked to species-specific OBO ontologies, to the Foundational Model of Anatomy (FMA) and to clinical terminologies.

5. Discussion and Conclusion

In this paper we introduced design principles and modeling decisions for the biological top-level ontology BIOTOP which is based on the GENIA ontology/annotation vocabulary as a semantic glue for connecting existing biomedical ontologies. BIOTOP has been devised as a rather expressive model which makes use of the full range of OWL-DL constructs. Future applications of BIOTOP will include the provision of semantically precise classes to improve the quality of semantically annotated corpora (while keeping downward compatibility to GENIA) and the assurance the consistency of biological ontologies in the further development of OBO and clinical terminologies. The latter goal may be partially impaired by the high computing demands of BIOTOP as a consequence of its expressiveness. We also plan to augment the current purely OWL-DL based implementation with SWRL rules. By doing so we believe to overcome the still existing expressivity gaps (stemming from the insufficient OWL-DL constructs) and hence to achieve better domain coverage. Necessary further steps will be BIOTOP's enhancement in the domain of biological functions and processes and the (semi-automatic) generation of natural language definitions in order to facilitate its usage and to assure its adequacy.

Acknowledgments. This research was supported by the European Network of Excellence "Semantic Mining" (NoE 507505). The second, third, and fourth author were additionally funded by the BOOTStrep project under grant FP6-028099, both within the EC's 6th Framework Programme.

References

[1] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *The Description Logic Handbook. Theory, Implementation, and Applications*. Cambridge, U.K.: Cambridge University Press, 2003.

[2] O. Bodenreider, B. Smith, and A. Burgun. The ontology-epistemology divide: A case study in medical terminology. In Achille C. Varzi and Laure Vieu, editors, *Proceedings of FOIS 2004*, pages 185–195.

[3] J. J. Cimino. Auditing the Unified Medical Language System with semantic methods. *Journal of the American Medical Informatics Association*, 5(1):41–45, 1998.

[4] A. G. Cohn. Formalising bio-spatial knowledge. In Chris Welty and Barry Smith, editors, *Proceedings of FOIS 2001*, pages 198–209.

[5] A. Gangemi, N. Guarino, C. Masolo, and A. Oltramari. Understanding top-level ontological distinctions. In *Proceedings of the IJCAI-01 Workshop on Ontologies and Information Sharing*, pages 26–33, 2001.

[6] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider. Sweetening ontologies with dolce. In *Proceedings of EKAW 2002*, pages 166–181.

[7] Gene Ontology Consortium. Creating the Gene Ontology resource: Design and implementation. *Genome Research*, 11(8):1425–1433, 2001.

[8] P. Grenon, B. Smith, and L. Goldberg. Biodynamic ontology: Applying BFO in the biomedical domain. In *Ontologies in Medicine*, number 102 in Studies in Health Technology and Informatics, pages 20–38, 2004.

[9] N. Guarino and C. A. Welty. Identity, unity, and individuality: Towards a formal toolkit for ontological analysis. In *Proceedings of ECAI 2000*, pages 219–223.

[10] V. Haarslev and R. Möller. RACER: A core inference engine for the Semantic Web. In *Proceedings of the 2nd International Workshop on Evaluation of Ontology-based Tools, Located at ISWC 2003*, pages 27–36, 2003.

[11] I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, and M. Dean. SWRL: A Semantic Web Rule Language Combining OWL and RuleML, 2004. `[http://www.w3.org/Submission/SWRL]` Last accessed: May 5th, 2006.

[12] N. Fridman Noy, R. W. Fergerson, and M. A. Musen. The knowledge model of PROTEGE-2000: Combining interoperability and flexibility. In *Proceedings of EKAW 2000*, pages 17–32.

[13] OBO. Open Biological Ontologies (obo), 2005. `[http://obo.sourceforge.net]` Last accessed June 26th, 2005.

[14] T. Ohta, Y. Tateisi, and J.-D. Kim. The GENIA corpus: An annotated research abstract corpus in molecular biology domain. In *HLT 2002 – Proceedings of the 2nd International Conference on Human Language Technology Research*, pages 82–86.

[15] A. Rector, R. Stevens, and J. Rogers. Simple bio upper ontology, 2006. `[http://www.cs.man.ac.uk/∼rector/ontologies/simple-top-bio]` Last accessed: May 5th, 2006.

[16] A. L. Rector, J. Rogers, and T. Bittner. Granularity, scale and collectivity: When size does and does not matter. *Journal of Biomedical Informatics*, 39(3):333–349, 2006.

[17] UMLS. *Unified Medical Language System*. Bethesda, MD: National Library of Medicine, 2005.

[18] S. Schulz and U. Hahn. Parthood as spatial inclusion: Evidence from biomedical conceptualizations. In *Proceedings KR 2004*, pages 55–63.

[19] S. Schulz, A. Kumar, and T. Bittner. Biomedical ontologies: What *part-of* is and isn't. *Journal of Biomedical Informatics*, 39(3):350–361, 2006.

[20] P. Simons. *Parts: A Study in Ontology*. Oxford: Clarendon Press, 1987.

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

[22] B. Smith and P. Grenon. The cornucopia of formal-ontological relations. *Dialectica*, 58(3):279–296, 2004.

[23] B. Smith, J. Williams, and S. Schulze-Kremer. The ontology of the Gene Ontology. In *Proceedings of the 2003 Annual Symposium of the American Medical Informatics Association*, pages 609–613, 2003.

[24] Tsujii Laboratory. Genia project home page, 2003. `[www-tsujii.is.s.u-tokyo.ac.jp/GENIA]` Last accessed: May 5th, 2006.

[25] A. C. Varzi. Mereology. In Edward N. Zalta, editor, *Stanford Encyclopedia of Philosophy*. Stanford: The Metaphysics Research Lab, 2003. `[plato.stanford.edu]` Last accessed: May 5th, 2006.