Applied Ontology ? (2007) 1–3
IOS Press


# BIOTOP: An Upper Domain Ontology for the Life Sciences

*A Description of its Current Structure, Contents, and Interfaces to OBO Ontologies*

Elena Beisswanger [a,*], Stefan Schulz [b], Holger Stenzhorn [b,c], and Udo Hahn [a]

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

**Abstract.**
In the life sciences, there is an ample need for semantic interoperability of data. Thus shared vocabularies are needed for consistently expressing meta data in terms of semantic annotations as well as for querying bibliographic information systems. In the past years, lots of highly specialized, yet also fragmented terminologies have evolved. However, they lack principled forms of conceptual interlinkage. In order to provide an ontological basis for a seamless integration of such isolated parts of biological knowledge, we here introduce BIOTOP, an upper domain ontology for molecular biology. We describe its structure and contents, as well as its current interfaces to a selected set of OBO ontologies, which contain more detailed terminological knowledge about specific areas of molecular biology, e.g., cell types, molecular functions, biological processes, and chemical compounds.

Keywords: bio-ontology, ontology integration, upper domain ontology, annotation vocabulary, biomedical terminologies

## 1. Biological Terminologies and Ontologies

Biological research and development activities continuously generate vast amounts of experimental data. This data stream feeds model organism specific and cross-species data bases for subsequent fact retrieval, making data interoperability and integration a major topic. Human curators manually add semantic meta data to the experimental data in terms of, e.g., sequence annotation and functional annotation of genes and gene products. Automatic means such as specific annotation editors guiding the annotation process [1] and information extraction and text mining systems [2, 3, 4] increasingly support manual work. Immense efforts have been made to set up ontologies and terminologies serving as meta languages for the annotation task (for a comprehensive survey, see [5]). Many of them are available within the OBO (*Open Biomedical Ontologies*) library.[1] The most prominent resource is the Gene Ontology (GO) [6] covering molecular functions, biological processes, and cellular components.

The OBO ontologies, by and large, were built independently from each other, each dealing with a specific subdomain of biomedicine (anatomy, cell types, molecular functions, biological processes, sequences, chemicals, etc.). Consequently, they lack any deeper form of conceptual integration and interlinkage, though from a scientific perspective the domains they cover are heavily interconnected.

Various approaches have been proposed to detect and formally represent those implicit relations between ontologies to make them accessible for computational purposes. Amongst others, the compositionality of GO terms in particular has been investigated [7, 8, 9, 10] and was exploited to derive computationally usable definitions [11, 12].These are certainly valuable integration efforts. However, we claim that the validity and significance of their results critically depend on grounding the domain ontologies on a formally rigid ontological framework, a so-called *Upper Ontology*.

*Corresponding author: Elena Beisswanger, Jena University Language and Information Engineering (JULIE) Lab, Friedrich-Schiller-Universität Jena, Jena, Germany

[1] `http://obo.sourceforge.net/`

We further claim that a bridge between both, the formal top layer and domain-specific ontologies is needed to guarantee a seamless transition from domain- and application-independent classes (also termed as types, concepts, etc.) and relations in the *Upper Ontology* to very specific classes in the domain ontologies. Such an intermediary layer is defined in terms of an *Upper Domain Ontology* capturing characteristic classes and relations of the respective domain. In this paper, we propose BIOTOP to serve as such a mediating layer for the life sciences domain.

The careful integration of domain ontologies via a common ontological top layer might in particular be beneficial for advanced forms of language technology applications such as text mining from full-texts (rather than abstracts) in terms of semantic meta data-based relation and event extraction (see also Section 4).

### *1.1. Upper (Domain) Ontologies*

An important step towards standardizing biomedical ontologies is due to Smith *et al.* [13] who developed a *Relation Ontology* (RO) needed for the conceptual representation of the biomedical domain. The RO contains consistent and unambiguous formal definitions for the basic relation types (currently up to ten, though this number might still be subject to change in the future) on four major axes — generic taxonomic and partonomic, spatial, derivational and participant relations. It is important to notice that all class-relations provided by RO are defined dependent on relations among the corresponding instances. Furthermore their domain and range is clearly specified as continuants (entities which persist through time) or occurrents (entities which develop over time, e.g. processes). A time parameter is included in the formal definition of a relation, if necessary.

However, the authors admit that using the same relation types, even with high-level domain and range restrictions, is not sufficient to guarantee interoperability. Hence, an additional common terminological framework must be supplied which empowers a seamless transition from the generic classes continuant and occurrent to other, still fundamental classes. These appear either at the domain-independent level (*Upper Ontology*) or at the generic domain-dependent level (*Upper Domain Ontology*).

For the life sciences, alternative proposals for *Upper Domain Ontologies* already exist, though they are still under development. The OBR framework (*Ontology of Biomedical Reality*) was introduced by Rosse *et al.* [14] in order to integrate domain ontologies from anatomy, physiology and pathology. It applies principles of the domain-independent *Upper Ontology* BFO to the field of biomedicine [15]. Alan Rector's *Simple Bio Upper Ontology*[2] is composed of a class hierarchy and a relation type hierarchy. It is intended to constrain the use of relation types to particular entity classes. GFO-BIO[3] is another *Upper Domain Ontology* for biology based on the top-level ontology GFO [16]. Rosse's and Rector's conceptualizations, unlike GFO-BIO, have a marked focus on medical concept abstractions. We here stipulate that such an approach is too narrow to account for the integration of biomedical ontologies that also cover bio-chemistry and molecular biology.

### *1.2. From* GENIA *to* BIOTOP

While the ontologies mentioned in the previous subsection have no particular application in mind besides connecting fragmented domain ontologies, for natural language processing (NLP) applications such as biomedical information extraction or text mining an ontological *de facto* standard has already been established through the GENIA ontology. [4] It forms the conceptual backbone for named entity annotations in the GENIA corpus [17] and is currently augmented by relation annotations, as well. The underlying ontology is, however, quite fragmentary and certainly not intended to serve as an *Upper Domain Ontology* in the sense outlined above.

[2] http://www.cs.man.ac.uk/~rector/ontologies/simple-top-bio
[3] http://onto.eva.mpg.de/gfo-bio.html
[4] http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/genia-ontology.html

The GENIA ontology is a pure taxonomy composed of (only) 48 classes, informally described by verbal 'scope notes'. It covers biochemical substances, such as **Protein Molecule**, **DNA Molecule**, and **Nucleotide**, and their natural locations, e.g., **Multi-Cell Organism**, **Tissue**, and **Cell Component**. As pointed out by Schulz *et al.* [18], there are major shortcomings with the GENIA ontology. Very briefly, many classes are either poorly or not defined at all, most scope notes are incomplete, non-taxonomic relations are missing, there is no commitment to any formal *Upper Ontology* which leads to a lack of ontological structure, and, finally, the non-standard naming policy for many GENIA classes is rather confusing for biologists.

In order to avoid these shortcomings but still preserve the results of previous work (corpus annotations, in particular), we created the BIOTOP ontology, a major redesign and extension of GENIA intended as an Upper Domain Ontology primarily for molecular biology and biomedicine.

## 2. A Brief Overview of BIOTOP

Our main goal in setting up BIOTOP is to provide an ontologically sound layer for linking and integrating various specific domain ontologies from the life sciences domain. We stipulate in particular, that integrated and, thus, more comprehensive ontologies will enhance the capabilities of advanced NLP applications in the life sciences such as information extraction and text mining.

With these considerations in mind, the structure of the original GENIA ontology was remodelled, some GENIA classes were removed, new BIOTOP classes were introduced and even whole new axes were added, significantly extending the scope of the original ontology. Instead of reusing GENIA's top level distinction between **Source** and **Substance**, the general top level ontology BFO [15] was set on top of BIOTOP. At the relational level, GENIA's exclusive use of a single taxonomic (*is-a*) relation was extended by relation types from the RO. BIOTOP is presently composed of 175 classes, linked by 171 instances of non-taxonomic binary relations taken from nine semantic relation types (including subrelations) and their reciprocal relations (as of February 19, 2008).

### *2.1.* BIOTOP *Classes*

BIOTOP inherits the top-level distinction of BFO between the classes **Continuant** and **Occurrent** and further between **Independent Continuant** and **Dependent Continuant**, the latter depending on the existence of some independent continuant. (For example the function of a protein is a dependent continuant since it cannot exist without a protein, which is an independent continuant.) However, the BFO subclasses of **Independent Continuant** were not incorporated in BIOTOP, since by that BFO enforces a distinction in terms of connection and wholeness. This, as a consequence, requires a commitment to a certain granularity level which could lead to inconsistencies with the inherently cross-granular BIOTOP (see [19]). On the other hand, distinctions missing in BFO were added, e.g. those between **Action**, **State** and **Process**.

Initially, BIOTOP, like GENIA, focused on molecular entities. Thus major parts of BIOTOP are subordinated to **Independent Continuant**$_{BFO}$. While the BIOTOP classes **Organism**, **Tissue**, **Cell**, **Cellular Component**, and **Atom** correspond to classes in the GENIA **Source** branch, the subclasses of **Mono Molecular Entity**$_{BioTop}$ correspond to the GENIA **Substance** branch.

However, BIOTOP gradually moved beyond the scope of GENIA providing a hierarchy of biological processes subsumed by **Process**$_{BFO}$ (a subclass of **Occurrent**$_{BFO}$), a hierarchy of biological functions subsumed by **Function**$_{BFO}$ (a subclass of **Dependent Continuant**$_{BFO}$), as well as several qualities and roles, such as **Physical Mass** and **Canonical State** subsumed by **Quality**$_{BFO}$, and **Signalling Role** subsumed by **Role**$_{BFO}$.

Following design considerations of advanced knowledge representation languages such as OWL, the *Web Ontology Language* [20], we want to support automatic terminological classification [21] as much as possible. Hence, we introduced existential and universal restrictions in class definitions, in terms of necessary, and, wherever possible, necessary and sufficient conditions. For example, the class **Nucleotide** is restricted by four necessary conditions:

1. **Nucleotide** *has-component* only (**Heterocyclic Base** or **Phosphate** or **Ribose**)
2. **Nucleotide** *has-component* exactly one **Heterocyclic Base**
3. **Nucleotide** *has-component* exactly one **Ribose**
4. **Nucleotide** *has-component* some **Phosphate**

### *2.2.* BIOTOP *Semantic Relation Types*

BIOTOP has a taxonomic backbone based on the subsumption relation *is-a* which relates subclasses to their parent classes. Additionally, it borrows semantic relation types from the RO, namely *proper-part-of*, *located-in*, *derives-from*, *has-participant*, and their reciprocal relations. The partonomic relation *proper-part-of* (and its reciprocal *has-proper-part*) is taken as transitive, non-reflexive, and asymmetric relation.

In addition to the RO, first the relation *has-inherence* (and its reciprocal *inheres-in*) was introduced to express the relation between physical objects and their inherent (biological) functions. Second, the relation *realization-of* (with its reciprocal *has-realization*) is used to link the realization of a function to the corresponding function. Third, two subrelation pairs of *has-part* were introduced, *viz has-grain* / *grain-of* (according to [22]) and *component-of* / *has-component*. Both relations are not transitive. *Has-grain* allows to define collectives as mass entities composed of their constituent singletons, such as populations of cells, amounts of protein molecules, etc. *Has-component* relates compounds to their constituent components based upon a non-overlapping and exhaustive partition, like a protein chain is related to its constituent amino acid monomers. A collective remains the same when one adds or removes a grain (e.g. a population of T-cells remains a population of T-cells when we remove a single T-cell). However, the sortal identity of a compound changes as soon as a single component is added or removed (e.g., when we remove an amino acid from the peptide chain of a protein this might change the over-all nature of the protein).

## 3. BIOTOP's Interfaces to OBO Ontologies

As an *Upper Domain Ontology* for the biomedical field, BIOTOP contains foundational and uncontroversial statements about the basic kinds of molecular biology and biomedicine and provides classes as interfaces to domain ontologies kept within the OBO framework. Using these classes to integrate different domain ontologies, BIOTOP can be used as common top level for the OBO. As a side effect, we expect the mapping of the OBO ontologies to BIOTOP classes to reveal hidden inaccuracies in the modelling practice of the single OBO ontologies, such as the conflation of classification axes. In the following, we suggest how the Gene Ontology (GO) [6], the Cell Ontology (CO) [23] and parts of the ChEBI ontology [24], as exemplars for all OBO ontologies, could be integrated using BIOTOP as an interface. We propose matches and subsumption relations between BIOTOP and OBO ontology classes that need to be adjusted and confirmed by the respective ontology developers in subsequent revisions.

The Gene Ontology (GO) is composed of three independent branches which relate to BIOTOP in the following way. The **Molecular Function**$_{GO}$ branch is subsumed by the class **Molecular Function**$_{BioTop}$ and the **Biological Process**$_{GO}$ branch is subsumed by **Biological Process**$_{BioTop}$. The subsumption relations are due to the fact that, unlike BIOTOP, GO restricts the meaning of "molecule" to "gene product" (which is a protein or RNA molecule), and restricts the biological process branch to processes in which gene products are involved. For the **Cellular Component**$_{GO}$ branch not the whole class hierarchy but only large parts of it are subsumed by **Cellular Component**$_{BioTop}$. This is because cellular components in BIOTOP are defined as proper parts of cells, whereas **Cellular Component**$_{GO}$ also subsumes classes going beyond the scope of this definition, such as **Extracellular Region**$_{GO}$ and **Cell**$_{GO}$ (which matches **Cell**$_{BioTop}$).

The top node of the Cell Ontology (CO), **Cell**$_{CO}$, matches **Cell**$_{BioTop}$. In addition, BIOTOP provides links to several subclasses of **Cell**$_{CO}$. If these links would be formally represented this would enrich the formal semantics of the CO. For example, linking **Eukaryotic Cell**$_{CO}$ to **Eukaryotic Cell**$_{BioTop}$, the CO class and all its subclasses would inherit the necessary condition that they must either be a eukaryotic

organism (this applies, e.g., to yeast cells) or be the proper part of a eukaryotic organism (this applies, e.g., to animal cells), as specified in BIOTOP.

The ChEBI ontology covers *Chemical Entities of Biological Interest*. Though the ChEBI classes are named with plural noun forms, the accompanying textual definitions reveal that, in fact, the singular noun form is meant. Disregarding the misleading names, **Molecular Entities**$_{ChEBI}$ and all its subclasses are subsumed by the union of **Mono Molecular Entity**$_{BioTop}$ with **Poly Molecular Entity**$_{BioTop}$. In addition, some of the ChEBI classes match direclty BIOTOP classes, amongst them **Atoms**$_{ChEBI}$ matching **Atom**$_{BioTop}$, **Simple Proteins**$_{ChEBI}$ corresponding to **Entire Protein Molecule**$_{BioTop}$, and **Nucleic Acids**$_{ChEBI}$ to **Nucleic Acid Molecule**$_{BioTop}$. The ChEBI branch starting with **Biological Role**$_{ChEBI}$ needs deeper consideration covering classes that match subclasses of **Material Entity**$_{BioTop}$ which are either restricted by a role or a function specification. For example the class **Food**$_{ChEBI}$ would be defined in BIOTOP as being equivalent to a subclass of **Material Entity**$_{BioTop}$ that has the existential restriction *has-inherence* some **Food Role**$_{BioTop}$.

## 4. Conclusion and Outlook

We demonstrated the need for an integration layer for biomedical ontologies. Considering this we introduced BIOTOP as an *Upper Domain Ontology* whose basic design was inspired by the GENIA ontology. Since then, it has grown to cover all foundational entity types of the whole realm of the life sciences. BIOTOP is intended as a bridge linking various domain-specific biomedical ontologies with the top-layer ontology BFO. Linking-up domain ontologies with BIOTOP is a complex task in itself, for which we have already taken the first step. By now we mapped the fundamental classes of the domain ontologies GO, CO, and ChEBI to BIOTOP classes. An enormous effort is still needed to complete the task for all relevant OBO ontologies. After its completion the OBO ontologies can comprehensively be used. This would allow for cross-ontology consistency checking, inferencing, and other value-adding inference services, vital e.g., for proper reference resolution in biomedical documents. Hahn *et al.* [25] already showed the value of proper taxonomic and partonomic reasoning for information extraction from biomedical documents. Quite recently, Poprat and Hahn [26] provided preliminary empirical evidence for the hypothesis that the use of composite, high-coverage terminological resources (such as the UMLS[5] or the NCI Thesaurus [27]) is far more advantageous for various forms of reference resolution than the use of heavily focused stand-alone ontologies (such as e.g., the Cell Ontology).

Solid experimental evidence that shows whether the BIOTOP redesign of is really better suited for e.g., corpus annotation and information extraction than the original GENIA source is still lacking. By now there is evidence that semantic annotation of scientific documents profits from basing the annotation vocabulary on a formally sound ontology such as BIOTOP [28]. To preserve the compatibility of future annotations based on BIOTOP with existing annotations from the GENIA corpus, BIOTOP contains, as an additional feature, mappings to GENIA classes.

BIOTOP is implemented in OWL-DL [20]. The ontology is under continuous development and the current version can be downloaded from `http://www.purl.org/biotop`. A discussion group has been established to debate topics relating to the theoretical background and implementation issues of BIOTOP (cf. `http://groups.google.com/group/biotop`).

## Acknowledgements

This work is supported in the context of the BOOTSTREP project (`www.bootstrep.org`), funded by the European Commission under contract FP6 – 028099.

[5] `http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html`

## References


[1] Michael Bada, Daniele Turi, Robin McEntire, and Robert Stevens. Using reasoning to guide annotation with Gene Ontology terms in GOAT. *ACM SIGMOD Record. Special Issue: Data engineering for life sciences*, 33(2):27–32, June 2004.

[2] Irena Spasic, Sophia Ananiadou, John McNaught, and Anand Kumar. Text mining and ontologies in biomedicine: Making sense of raw text. *Briefings in Bioinformatics*, 6(3):239–251, 2005.

[3] Aaron M. Cohen and William Hersh. A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6(1):57–71, 2005.

[4] Lars Juhl Jensen, Jasmin Saric, and Peer Bork. Literature mining for the biologist: from information retrieval to biological discovery. *Nature Reviews Genetics*, 7(February):119–129, 2006.

[5] Olivier Bodenreider and Robert Stevens. Bio-ontologies: Current trends and future directions. *Briefings in Bioinformatics*, 7(3):256–274, 2006.

[6] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–29, May 2000.

[7] Philip V. Ogren, Kevin B. Cohen, George K. Acquaah-Mensah, Jens Eberlein, and Lawrence Hunter. The compositional structure of GENE ontology terms. In Russ B. Altman, A. Keith Dunker, Lawrence Hunter, Tiffany A. Jung, and Teri E. Klein, editors, *PSB 2004 – Proceedings of the Pacific Symposium on Biocomputing 2004*, pages 214–225. Hawaii, USA, January 6-10, 2004. Singapore: World Scientific Publishing, 2004.

[8] Anita Burgun and Olivier Bodenreider. An ontology of chemical entities helps identify dependence relations among Gene Ontology terms. 2005.

[9] Simen Myhre, Henrik Tveit, Torulf Mollestad, and Astrid Laegreid. Additional Gene Ontology structure for improved biological reasoning. *Bioinformatics*, 22(16):2020–2027, August 2006.

[10] Michael Bada and Lawrence Hunter. Enrichment of OBO ontologies. *Journal of Biomedical Informatics*, 40(3):300–315, 2007.

[11] Chris J. Wroe, Robert Stevens, Carole A. Goble, and Michael Ashburner. A methodology to migrate the Gene Ontology to a description logic environment using DAML+OIL. In Russ B. Altman, A. Keith Dunker, Lawrence Hunter, Tiffany A. Jung, and Teri E. Klein, editors, *PSB 2003 – Proceedings of the Pacific Symposium on Biocomputing 2003*, pages 624–635. Kauai, Hawaii, USA, January 3-7, 2003. Singapore: World Scientific Publishing, 2003.

[12] Christopher J. Mungall. Obol: integrating language and meaning in bio-ontologies. *Comparative and Functional Genomics*, 5:509–520, 2004.

[13] Barry Smith, Werner Ceusters, Bert Klagges, Jacob Köhler, Anand Kumar, Jane Lomax, Chris Mungall, Fabian Neuhaus, Alan L. Rector, and Cornelius Rosse. Relations in biomedical ontologies. *Genome Biology*, 6(5):R46 (1:15), 2005.

[14] Cornelius Rosse, Anand Kumar, José Leonardo V. Mejino, Daniel L. Cook, Landon D. Detwilern, and Barry Smith. A strategy for improving and integrating biomedical ontologies. In Charles P. Friedman, editor, *AMIA 2005 – Proceedings of the Annual Symposium of the American Medical Informatics Association. Biomedical and Health Informatics: From Foundations, to Applications to Policy*, pages 639–643. Washington, D.C., USA, October 22-26, 2005, 2005.

[15] Pierre Grenon, Barry Smith, and Louis Goldberg. Biodynamic ontology: Applying BFO in the biomedical domain. In Domenico M. Pisanelli, editor, *Ontologies in Medicine – Proceedings of the Workshop on Medical Ontologies*, number 102 in Studies in Health Technology and Informatics, pages 20–38. Rome, Italy, October 2003. Amsterdam: IOS Press, 2004.

[16] Heinrich Herre, Barbara Heller, Patryk. Burek, Robert Hoehndorf, Frank Loebe, and Hannes Michalek. General Formal Ontology (GFO): A foundational ontology integrating objects and processes. Part I: Basic principles. Technical report, Research Group Ontologies in Medicine (Onto-Med), University of Leipzig, 2006.

[17] Tomoko Ohta, Yuka Tateisi, and Jin-Dong Kim. The GENIA corpus: An annotated research abstract corpus in molecular biology domain. In M. Marcus, editor, *HLT 2002 – Human Language Technology Conference. Proceedings of the 2nd International Conference on Human Language Technology Research*, pages 82–86. San Diego, Cal., USA, March 24-27, 2002. San Francisco, CA: Morgan Kaufmann, 2002.

[18] Stefan Schulz, Elena Beisswanger, Udo Hahn, Joachim Wermter, Anand Kumar, and Holger Stenzhorn. From GENIA to BIOTOP: Towards a top-level ontology for biology. In Brandon Bennett and Christiane Fellbaum, editors, *Formal Ontology in Information Systems. Proceedings of the 4th International Conference – FOIS 2006*, number 150 in Frontiers in Artificial Intelligence and Applications, pages 103–114. Baltimore, Maryland, USA, November 9-11, 2006. Amsterdam: IOS Press, 2006.

[19] Stefan Schulz, Martin Boeker, and Holger Stenzhorn. How granularity issues concern biomedical ontology integration. In *MIE'08 – Medical Informatics Europe 2008. Proceedings of the 21th International Congress of the European Federation for Medical Informatic (to appear)*, 2008. Accepted for publication.

[20] Ian Horrocks, Peter F. Patel-Schneider, and Frank van Harmelen. From $\mathcal{SHIQ}$ and RDF to OWL: The making of a Web ontology language. *Journal of Web Semantics*, 1(1):7–26, 2003.

[21] Robert MacGregor. The evolving technology of classification-based knowledge representation systems. In John F. Sowa, editor, *Principles of Semantic Networks. Explorations in the Representation of Knowledge*, pages 385–400. San Mateo, CA: Morgan Kaufmann, 1991.

[22] Alan L. Rector, Jeremy Rogers, and Thomas Bittner. Granularity, scale and collectivity: When size does and does not matter. *Journal of Biomedical Informatics*, 39(3):333–349, 2006.

[23] Jonathan Bard, Seung Y. Rhee, and Michael Ashburner. An ontology for cell types. *Genome Biology*, 6(2):R21, 2005.

[24] Kirill Degtyarenko, Paula de Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. CHEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36(Database-Issue):344–350, 2008.

[25] Udo Hahn, Martin Romacker, and Stefan Schulz. Discourse structures in medical reports – watch out! The generation of referentially coherent and valid text knowledge bases in the MEDSYNDIKATE system. *International Journal of Medical Informatics*, 53(1):1–28, 1999.

[26] Michael Poprat and Udo Hahn. An investigation into the reusability of biomedical terminologies for the resolution of referential expressions. In Hagit Shatkay, Lynette Hirschman, Alfonso Valencia, and Christian Blaschke, editors, *BioLINK 2007 – Proceedings of the BioLINK SIG 2007. The Annual Meeting of the ISMB BioLINK Special Interest Group on Text Data Mining, in Association with ISMB 2007*, pages 39–42. Vienna, Austria, July 19, 2007, 2007.

[27] Nicholas Sioutos, Sherri de Coronado, Margaret W. Haber, Frank W. Hartel, Wen-Ling Shaiu, and Lawrence W. Wright. NCI thesaurus: A semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics*, 40(1):30–43, 2007.

[28] Ai Kawazoe, Lihua Jin, Mika Shigematsu, Roberto Barerro, Kiyosu Taniguchi, and Nigel Collier. The development of a schema for the annotation of terms in the BioCaster disease detecting/tracking system. In Olivier Bodenreider, editor, *Proceedings of the KR-MED 2006 Workshop*, volume 222, pages 77–85. Baltimore, MD, USA, Nov 8, 2006, 2006.