Waclaw Kusnierczyk

Augmenting Bioinformatics Research with Biomedical Ontologies

PhD thesis

2008

Department of Computer and Information Science
Faculty of Information Technology, Mathematics and Electrical Engineering
Norwegian University of Science and Technology

Abstract

The main objective of the reported study was to investigate how biomedical ontologies, logically structured representations of various aspects of the biomedical reality, can help researchers in analyzing experimental data. High-throughput technologies, such as platforms for microarray gene expression experiments (MAGE), yield increasingly large amounts of 'massive' data; successful analyses of the data require expertise in the application domain, and to support researchers with automated methods explicit domain knowledge representations are often essential. Two attempts to construct such tools are reported here: one successful — eGOn, a web-based tool for mapping the results of microarray gene expression experiments onto the structure of the Gene Ontology; and one unsuccessful — a framework for knowledge- and case-based enhancement of biological association network-building tools.

Ontologies — structured, computer-understandable accounts of expert knowledge — are a relatively new invention. Until only recently, biomedical ontologies were developed with little care for formal semantics, syntactic and semantic compatibility with each other, and the ontological (in a philosophical sense) commitments made. However, for successful integration of resources that use different ontologies to describe their content and services, integration on the ontological level is needed. While one way to achieve this is to apply some of the much-researched techniques of ontology alignment and merging, another approach is to organize the ontology creation movement around a common basis — a unique top-level ontology and a set of design principles. Some of such integrative efforts made by the community of Open Biomedical Ontologies, in which I have participated, are reported. Furthermore, the thesis presents a framework for consistently connecting the Gene Ontology (the most prominent of ontologies covered by OBO) with the Taxonomy of Species, and discusses the benefits of its prospective adoption by the OBO community.

Acknowledgments

This dissertation is submitted to the Norwegian University of Science and Technology (NTNU) in partial fulfillment of the requirements for the degree *doktor ingeniør* (PhD). It reflects and partially reports my work done at the Department of Computer and Information Science (IDI) in the course of a graduate study under the guidance of professors Jan Komorowski, Astrid Lægreid, Agnar Aamodt, and Barry Smith.

I would not have achieved much without support from my supervisors, collaborators, and colleagues. My work at the Norwegian University of Science and Technology (NTNU) in Trondheim was initially part of a research initiative in functional genomics, conducted under the guidance of prof. Jan Komorowski,[1] prof. Astrid Lægreid,[2] and prof. Arne Sandvik.[3] It is these three people, and especially Jan and Astrid, whom I owe thanks for introducing me to the field and for inviting me to be a member of their research team. Special thanks go to prof. Agnar Aamodt,[4] who agreed to become my supervisor after prof. Komorowski had left for Sweden. I am deeply grateful for prof. Aamodt's having introduced me to Artificial Intelligence, Machine Learning, and Case-Based Reasoning; for his faith in my ability to succeed; and for his continuous support in my efforts to find the right environment and an appropriate research target. Last, but not least among my tutors,

[1] Prof. Komorowski is currently the Scientific Director of the Linnaeus Center for Bioinformatics at the Uppsala University, Sweden (http://www.lcb.uu.se/).

[2] Prof. Lægreid is currently a Prorector at NTNU.

[3] Prof. Sandvik is currently a Project Leader at the NTNU Microarray Core Facility.

[4] Prof. Aamodt is currently the Leader of the Division of Intelligent Systems at IDI, NTNU.

prof. Barry Smith[5] deserves particular appreciation for introducing me to the Open Biomedical Ontologies community, for involving discussions, and for inviting me to visit him and his research team in Saarbrücken.

I would like to thank my colleagues at the Institute for their help, discussions, and company. I thank Jörg for helping me in recovering from innumerable annoyances of LaTeX and TeX, and for discussions on these and many other issues. I would also like to thank those at IDI with whom I had the chance to collaborate in teaching — this was a great experience. Kudos to all members of the administration staff for their support in solving mundane problems. I am also grateful to colleagues from the Department of Cancer Research and Molecular Medicine (IKM) and other institutes at the Faculty of Medicine (DMF) at NTNU who helped me or allowed me to participate in their projects. I owe thanks to those members of the Ontolog Forum[6] who, with amazing patience and understanding, responded to my questions, complaints, and claims.

Last not least, the warmest *thank you* to my closest family for all the support and encouragement they have been giving me throughout the whole period of my study and writing.

[5] Prof. Smith is an internationally renowned philosopher-ontologist; currently, he is, among others, the Director of the Institute for Formal Ontology and Medical Information Science (IFOMIS) in Saarbrücken, Germany, and a member of the Scientific Advisory Board of the Gene Ontology Consortium.

[6] http://ontolog.cim3.net/.

Contents

List of Figures

List of Tables

Chapter 1

Outline

This document reflects and partially reports my work done at the Department of Computer and Information Science (IDI), Norwegian University of Science and Technology (NTNU) in the course of a graduate study under the guidance of professors Jan Komorowski, Astrid Lægreid, Agnar Aamodt, and Barry Smith. Chapter 1 provides an outline of the work I have done, and is composed as follows:

– Section 1.1 provides an overview of my research activities, lists publications, and specifies my contributions. Sec. 1.1.4 summarizes the research questions, goals, methods, and results.

– Section 1.2 provides an outline of the structure of the remainder of the thesis.

Much of my work was done as part of a broader collaborative effort; therefore, the plural form 'we' is extensively used throughout this document, and the singular 'I' is reserved to those parts where the entire work was done by me.

1.1 Research Overview

The initial goal of my doctoral research was to build computational models of gastric acid secretion, but as the work progressed and my interests were evolving, biomedical ontologies became the central issue. The work progressed in three distinct, but logically connected phases, each with a different focus, different results, and directed by different supervisors. Only the third phase is covered to a larger extent in this thesis. The earlier phases and the corresponding results are briefly discussed in the following sections, and are only occasionally mentioned in later chapters.

1.1.1 Phase I: Analysis and Mining of Microarray Data

In the first phase, we explored the possibilities of augmenting research in the then emerging field of functional genomics[1] with methods from computer science, specifically data mining[2] and machine learning.[3] In particular, we[4] were interested in investigating similarities between expression profiles of genes which participate in gastric acid secretion under various experimental conditions, and in reusing the discovered patterns for predicting functions of so-called 'unknown' genes (genes with unknown functions). Our goal was to develop computational models of gastric acid secretion and of the pathogenesis of gastrointestinal neoplasia, based on, e.g., the results of microarray gene expression (MAGE) experiments.[5]

On the computational side, my responsibilities included the development

[1] See, e.g., Hieter and Boguski [165] or Winslow and Boguski [386] for a brief introduction, and Campbell and Heyer [67] for an excellent, comprehensive account of this research field.

[2] See, e.g., Witten and Frank [387] or Han and Kamber [158] for a comprehensive introduction, and Duda et al. [96] for an excellent treatment of the closely related field of pattern classification.

[3] See, e.g., Mitchell [257] for a classic, though somewhat superficial introduction.

[4] This part of my work was done in collaboration with a team of molecular biologists, clinicians, and statisticians, under the guidance of proff. Komorowski, Lægreid, and Sandvik.

[5] A brief introduction to the microarray gene expression technology is given in Sec. 2.4. For more details see, e.g., Brown and Botstein [60] — one of the earliest articles on microarrays — as well as The Chipping Forecast I, II, and III, collections of review articles published as supplements to Nature Genetics [2].

of a programmatic framework for analyzing microarray experiments, from raw data analysis to pattern classification. The task consisted of two major components:

1. Design and implementation of statistical methods for the preprocessing of raw microarray gene expression data.
2. Design and implementation of classifiers for the functional classification of gene expression patterns.

While the microarray technology was already relatively advanced at that time (though rather immature as compared to the most recent developments in this field), statistical methods for preprocessing and analysis of microarray experiments were largely in an early developmental stage, and ready-to-use software tools were neither easily available nor sufficiently reliable. Software packages supporting statistical processing of MAGE data, such as Bioconductor (Gentleman et al. [125]), Limma (Smyth et al. [337]), or BRB ArrayTools (McShane et al. [250]), were only emerging during those and the next few years. Likewise, methods for the classification and functional analysis of microarray data were under development at that time; relatively simple tools such as Cluster and TreeView (Eisen et al. [100]) were used extensively, and functional annotations with terms from the Gene Ontology (Ashburner et al. [19]) were only becoming popular.

The first of the components mentioned above — statistical preprocessing and analysis — was realized on the basis of statistical tests designed by us as well as on those previously published by others. Implementations were mostly of prototype rather than production quality; We used Perl,[6] R,[7] and occasionally S-Plus[8] as the underlying implementation languages, primarily because of the availability of built-in regular expression operators (Perl),

[6] The *de facto* programming language of bioinformatics, http://www.perlfoundation.org/.

[7] The R environment for statistical computing, http://www.r-project.org/.

[8] http://www.insightful.com/products/splus/. Interestingly, the commercial S-Plus has unusual scoping, unintuitive and inconsistent with the usual lexical scoping rules (adopted in the otherwise quite similar R); this caused a number of my programs to produce wrong results before I discovered this flaw. Insightful, the provider of S-Plus, confirmed it to be an intended property of the language rather than an implementation bug (private conversation with a representative). This 'feature' has not been changed as of 2007.

statistical and graphical subroutines (R), and ease of rapid prototyping. The preprocessing included various sorts of normalization and filtering of raw data, aimed at addressing systematic sources of error such as the effects of printing, labeling, scanning, etc. Subsequent analyses included statistical testing for genes differentially expressed in distinct samples of biological material. Some of the methods designed or adopted by us are briefly discussed in, e.g., Midelfart [253].

The second of the components mentioned above — functional classification — was realized with the use of a supervised data mining technology based on the mathematical theory of rough sets (Rough Set Theory, RST; see, e.g., Pawlak [278], the seminal article on RST). In brief, RST-based classifiers are built by learning from discrete (or discretized) training data in an eager approach, and consist of a set of decision rules that are combined into ensembles and used for classifying previously unseen data. The major benefits of using this technology were its capability of handling noisy and inconsistent data, human-readable format of the classifiers (sets of logical rules), as well as multiclass rather than binary classification. In our studies, rough set classifiers were implemented mainly as prototype Perl scripts or as plugins for the Rosetta system (Øhrn [273]), developed earlier by the Komorowski's team. For more on the use of rough set-based classifiers in bioinformatics, see Øhrn [273], Hvidsten [177], Hvidsten et al. [178, 179], Midelfart et al. [254], Midelfart [253], and Lægreid et al. [226].

The results of studies based on microarray gene expression experiments which involved my contribution have been presented at both Norwegian and international conferences and workshops, as well as published in scientific journals. Specifically, I contributed to the following publications:

P1. Midelfart, **Kuśnierczyk**, et al. (2002). *Learning Yeast Gene Function from Expression Programs and Gene Ontology*. Conference Proceedings of the Winter Meeting of the Norwegian Biochemical Society, Røros, Norway [254].[9]

[9]This work had been earlier presented by Kuśnierczyk and Sonnervik as part of the NTNU course on Advanced Data Mining and Knowledge Discovery in Molecular Biomedicine (SIF80BG), and later at the Computer Science Graduate Students Conference (CSGSC2002) in Trondheim, under

P2. Yadetie, Bakke, Lægreid, **Kuśnierczyk**, et al. (2002). *Analysis of Effects of the PPAR-α Agonist Ciprofibrate on Rat Hepatic Gene Expressions Using cDNA Microarrays*. Proceedings of the 7th IUBMB Conference on Receptor-Ligand Interactions: Molecular, Physiological and Pharmacological Aspects, Bergen, Norway [392].

P3. Yadetie, Lægreid, Bakke, **Kuśnierczyk**, et al. (2003). *Liver Gene Expression in Rats in Response to the Peroxisome Proliferator-Activated Receptor-α Agonist Ciprofibrate*. Physiological Genomics [393].

P4. Nørsett, **Kuśnierczyk**, et al. (2003). *Gene Expression Profiles in Gastric Mucosa During Therapeutic Acid Inhibition*. Proceedings of the 1st ESF Functional Genomics Conference, Praha, Czech Republic [268].

P5. Nørsett, **Kuśnierczyk**, et al. (2003). *Gene Expression Profiles in Hypergastrinemic Rats and Patients*. Conference Proceedings of the Winter Meeting of the Norwegian Biochemical Society, Geilo, Norway [269].

P6. Hofsli, Thomessen, Yadetie, Langaas, **Kuśnierczyk**, et al. (2005). *Identification of Novel Growth Factor-Responsive Genes in Neuroendocrine Gastrointestinal Tumour Cells*. British Journal of Cancer [169].

P7. Nørsett, Bruland, Ween, Hofsli, Thomessen, Misund, Strømmen, **Kuśnierczyk**, et al. (2005). *Systems Biology of the Normal and Diseased Gastrointestinal System*. Proceedings of the 2nd ESF Functional Genomics Conference, Oslo, Norway [267].

P8. Nørsett, Lægreid, **Kuśnierczyk**, et al. (2007). *Changes in Gene Expression of Gastric Mucosa During Therapeutic Acid Inhibition*. American Journal of Physiology — Gastrointestinal and Liver Physiology [270].

My contribution to these studies comprised the design and implementation of tools for statistical and classificatory processing of microarray data, and, in some cases, statistical design of the actual wet-lab experiments. Using our own as well as publicly available gene expression data, we showed, using

the same title (Kuśnierczyk and Sonnervik [225]). The publication by Midelfart et al. [254] was submitted without the knowledge and consent of the original authors.

statistical significance tests, that rough set-based classifiers were a reasonable choice for bioinformatics research in functional genomics, not worse, and in some cases better than other data mining techniques. The tools and methods developed for those studies were typically application-specific, however, and most of them have not been published separately (but see, e.g., Midelfart [253] and Midelfart et al. [254] for an in-depth treatment of the rough set-based approach). Work done during the first phase strongly motivated my further research, but the details are largely irrelevant for what constitutes the major topic of this dissertation (see Sec. 1.1.3).

1.1.2 Phase II: Knowledge-Guided Microarray Data Analysis

Although the techniques used in our studies mentioned above were largely data-oriented, annotation of experimental data with terms from the Gene Ontology did play an important role in the development of our methods for functional classification of genes. As a consequence, I became interested in using computationally amenable forms of both general knowledge (domain knowledge) as well as episodic knowledge (case-specific knowledge) for the purposes of analyzing biomedical data. In this second phase, we[10] were exploring how model-based and case-based reasoning supported with biomedical ontologies can be used to draw biologically interesting inferences from experimental data.

Exploring the Gene Ontology with eGOn In one line of research, we were further investigating the applicability of the Gene Ontology to the analysis of microarray experiments. The success of the GO as a structured vocabulary for the annotation of biomedical data has motivated developers to build tools that would enable researchers to use annotations not only for the purposes of data integration, but also to interpret the data and perform various inferences based on the structure of the GO using, e.g., statistical methods. We have implemented eGOn,[11] a tool for mapping the results of microarray gene expression experiments onto the structure of the Gene On-

[10] This part of my work was done under the guidance of proff. Aamodt and Lægreid.
[11] The acronym 'eGOn' stands for '*e*xplore *G*ene *On*tology'.

tology. In eGOn, a collection of statistical tests allows one to compare gene expression in a number of samples or experiments, and the scope of those tests can be constrained to genes annotated with a user-defined selection of GO terms. Three tests are available; in brief, eGOn includes:

- a test used to find those GO terms for which differentially expressed genes are significantly overrepresented as compared to all genes covered by the study (the *master-target* test); this test is based on the Fisher's exact test for two binomial proportions;
- a test used for two non-overlapping lists of differentially expressed genes (the *mutually exclusive target-target* test); this test is based on similar assumptions as the master-target test;
- a test defined for two overlapping lists of differentially expressed genes (the *intersecting target-target* test); this test is based on the χ-square distribution.

The tool has been publicly available since it was first presented to the public in 2002:

P9. Beisvåg, Jølsum, **Kuśnierczyk**, et al. (2002). *eGOn: A New Tool for Mapping Microarray Data onto the Gene Ontology Structure*. Proceedings of the 1st Workshop on Standards and Ontologies for Functional Genomics (SOFG),[12] Hinxton, UK [33].

The tests mentioned above and other technicalities are discussed in more detail in a recent article (Beisvåg et al. [34]), where eGOn is presented as a component of the GeneTools software suite for gene annotation-related services. The tool is accessible from the GeneTools website[13] of the Norwegian Microarray Consortium, and has also been included in the list of GO-related tools, maintained by the Gene Ontology Consortium.[14] It has been evaluated both theoretically (the correctness of the statistical procedures) and empirically (the usefulness of the tool for a wider audience).

[12] http://www.sofg.org/.
[13] http://www.genetools.microarray.ntnu.no/.
[14] http://geneontology.org/GO.tools.shtml.

Building Biological Association Networks In another line of research, my motivation stemmed from the fact that researchers in functional genomics often come up with lists of genes whose participation in the functioning of the studied organisms is conveniently explained in the form of biological association networks (BANs). Microarray studies have been used to show how the expression of genes depends on a number of positive and negative regulation patterns, reflecting interactions between the respective gene products, as well as interactions between those products and the genes themselves (e.g., at the level of transcriptional regulation). Building complex models of such regulatory systems is one of the problems in focus of the relatively young field of systems biology (SB).[15]

Building biological association networks requires not only expression data from particular experiments, but also data of other sorts (e.g., known intermolecular interactions and metabolic pathways), usually obtainable from public databases. Together, these data reflect various aspects of the investigated biological system. Integration of distributed sources of biological data is a hot research topic on its own (see Ch. 2.5). Numerous services have been designed to facilitate transparent access to data stored in multiple databases — for example, the Gene Cards service (Safran et al. [304]) which provides access and uniformly presents information related to human genes, stored in a number of publicly available databases.

Various tools, such as Pathway Studio (Nikitin et al. [266]), Cytoscape,[16] and other (see, e.g., Hanisch et al. [159], Hoffman and Valencia [168], Jensen et al. [183], and Sohler et al. [338]), are commonly used to build biological association networks. These tools provide automated or semiautomated support for the retrieval, matching, filtering, processing, and presentation of data. However, without substantial manual intervention from the human user, networks built by these tools would either be unilluminatingly simple, or incomprehensibly complex. (Figure 1.1 illustrates this issue with an image of a comprehensive but not comprehensible gene network.) This is because the data retrieval is parametrized by the depth of search: in a

[15] See, e.g., Kitano [202] and others in that volume for an overview, and Alon [11] for a comprehensive introduction to Systems Biology.

[16] http://www.cytoscape.org/.

shallow search, the tool reports only those pieces of information that can be accessed via direct links from the initial nodes; in a deep search, all pieces of information accessible directly an indirectly (in a limited number of steps) are reported. In either case, information that is accessible but irrelevant will be reported, as well as information that is relevant but not accessible will be missed. Borrowing terms from the field of information retrieval,[17] it can be said that neither precision nor recall are perfect with respect to the user's expectations.

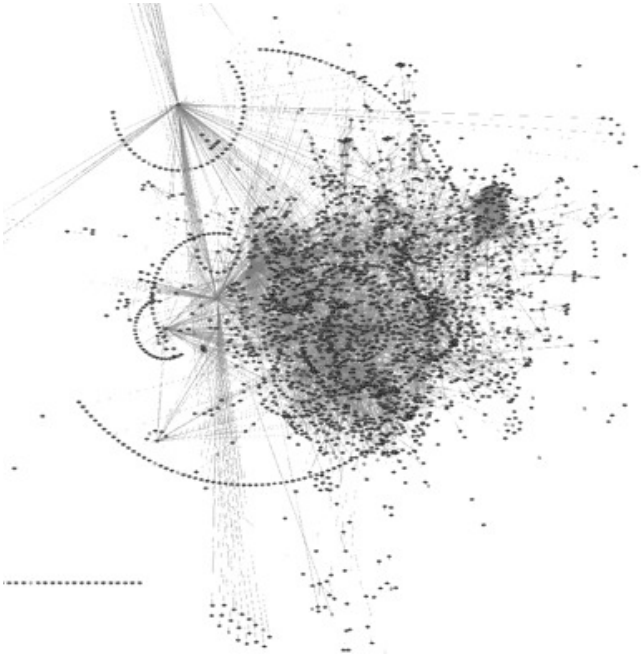

Figure 1.1: A comprehensive but incomprehensible representation of a biological association network. From Cytoscape, http://www.cytoscape.org/.

[17] See, e.g., Bayeza-Yates and Ribeiro-Neto [25] for a slightly dated, but comprehensive introduction.

In our studies, we observed that biologists combine general domain knowledge with intuition and previous experience to heuristically build BANs that are simple enough not to become unreadable, but complex enough to provide a useful explanation of the studied phenomena. They retrieve data (e.g., using network building tools such as those mentioned above) in a number of subsequent steps, at each step deciding which pieces of information retrieved in previous steps should be used in subsequent searches, which databases to query, etc. Based on such observations, I proposed to enrich network building tools with a reasoning module that would, at least partially, dispense the user from manually deciding on what to do at each successive step. Drawing from earlier achievements of the local research team (Aamodt [6, 4, 5], Grimnes [141], Gu [145], Öztürk [274], and Sørmo [339]), we designed a system for combining model-based reasoning (MBR)[18] with case-based reasoning (CBR).[19] In that system, general domain knowledge would come from biomedical ontologies, such as the Gene Ontology and other Open Biomedical Ontologies (OBO),[20] as well as from sources of precompiled interaction networks and pathways, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa et al. [193, 192]), MetaCyc (Caspi et al. [68]), etc. On the other hand, episodic knowledge (cases) would be collected incrementally during interactions with human experts in the course of network building sessions.

Unfortunately, this project had experienced a number of problems on the way, including low quality of the available ontologies, limited accessibility to experts willing to provide the essential input, and implementational issues with the Creek CBR software system.[21] Despite initial progress, this project has never been completed. The early stages have been documented in two peer-reviewed publications:

P10. **Kuśnierczyk** et al. (2004). *Towards Automated Explanation of Gene-*

[18] See, e.g., Davis and Hamscher [87].

[19] See, e.g., Aamodt and Plaza [8] for an introduction, and Kolodner [210] for a comprehensive account.

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

[21] Creek was initially implemented in Common Lisp by Aamodt and colleagues [4], and later reimplemented in Java by Sørmo [339] and fellow students at NTNU. Its recent versions are used for commercial purposes and are not freely available.

Gene Relationships. Proceedings of the 8th Conference on Research in Computational Molecular Biology (RECOMB), San Diego, USA [224].

P11. **Kuśnierczyk** et al. (2005). *Knowledge-Intensive Case-Based Support for Automated Explanation of Biological Phenomena*. Proceedings of the Workshop on Case-Based Reasoning in the Health Sciences, 6th International Conference on Case-Based Reasoning (ICCBR), Chicago, USA [223].

Parts of this research were also presented at the 2005 Winter Meeting of the Norwegian Biochemical Society in Meråker, Norway, as an invited talk on submodeling in systems biology.

1.1.3 Phase III: Biomedical Ontology Engineering

After having gained experience with data-oriented as well as ontology-enhanced analysis of microarray experiments, it seemed appealing and natural to focus further efforts on the biomedical ontologies themselves. The third phase of my work, therefore, was oriented towards principles of ontological engineering in the biomedical domain, and in particular on improving the structure of the Gene Ontology. Prof. Smith involved me in various endeavours related to the goals and work of the biomedical ontology community he was a member of: the Open Biomedical Ontologies initiative,[22] and its successor OBO Foundry.[23] Prof. Smith's earlier work focused on philosophical ontology, including research on a number of issues that have been studied (and disagreed about) since the times of Aristotle (Smith [333]). Recently, this philosophical perspective has been employed in investigations on how to coherently reflect, in the form of ontologies, the knowledge accumulated and used by biologists.

The issue of explicitly representing knowledge[24] has traditionally been of interest for researchers in artificial intelligence (AI) — specifically, those in

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

[23] http://obofoundry.org/.

[24] In his *Knowledge Representation* [340], John Sowa writes: "The words 'knowledge' and 'representation' have provoked philosophical controversies for over two and a half millennia. ... Plato's

the fields of knowledge representation (KR) and its more recent relative ontological engineering (OE; Gómez-Pérez et al. [133]). Unfortunately, there seems to have been a tradition of three different perspectives on ontological activity which did not necessarily work well together. Domain experts (e.g., biologists) have been developing their ontologies with little or no understanding of the logical and computational properties of knowledge representation languages.[25] Many ontologies have been built without precisely specifying the underlying language, as in the case of the early Gene Ontology (see, e.g., Smith et al. [329] and Smith et al. [334]) as well as ignoring the categorial systems developed earlier (see, e.g., McCray [248], Rector et al. [291], Smith [322, 321], and also Burgun [62] and Cimino [77]). Computer scientists have been focusing on the syntactic, semantic, and computational properties — decidability, tractability, etc. — of knowledge representation languages, often with little understanding and care for the needs of domain experts, as well as attempting to make their logical languages as ontology-neutral as possible — as in the case of, e.g., the Semantic Web language RDF[26] or the recent logical formalism IKL.[27] Philosophers have been developing their theories with little regard for the issues of computational tractability, mostly using the undecidable first-order logic (FOL), or other highly expressive formalisms such as modal or higher-order logics (see, e.g., Masolo et al. [244] for the modal axiomatization of DOLCE[28] and some other top-level ontologies). Often, they were occupied with abstracting as much as possible from the details specific to particular domains.

Central to the OBO approach to ontology development is close collaboration between biologists, computer scientists, logicians, philosophers, and experts of other relevant professions. I became involved in a number of ac-

student Aristotle shifted the emphasis of philosophy from the nature of knowledge to the less controversial, but more practical problem of representing knowledge." In the context of biomedical ontology, it is argued that it is not knowledge, but that what the knowledge is about that should be represented (Bodenreider et al. [51], Ceusters and Smith [70], Ceusters et al. [71], Smith [319, 321, 322], etc.)

[25] See, e.g., Aranguren et al. [18] for a recent discussion of why it is essential to take care of formal semantics while constructing a domain ontology.

[26] Resource Description Framework, http://www.w3.org/RDF/.

[27] IKRIS Knowledge Language, http://www.ihmc.us/users/phayes/IKL/GUIDE/GUIDE.html.

[28] Descriptive Ontology for Linguistic and Cognitive Engineering, http://www.loa-cnr.it/DOLCE.html.

tivities undertaken by the OBO community; together with prof. Smith and colleagues, I participated in, among others, the development of representational, terminological, and nomenclatural standards for OBO ontologies. Most recently, I have been involved in efforts aimed at cleanly linking various domain ontologies, such as the Gene Ontology, anatomy ontologies, etc., with the taxonomy of species. These efforts are reflected in the following publications:

P12. **Kuśnierczyk** (2006). *Nontological Engineering*. Proceedings of the 4th Conference on Formal Ontology in Information Systems (FOIS), Baltimore, USA [218].

P13. Smith, **Kuśnierczyk**, et al. (2006). *Towards a Coherent Terminology for Principles-Based Ontology*. Proceedings of the 2nd International Workshop on Formal Biomedical Knowledge Representation (KRMED), Baltimore, USA [331].

P14. Schober, **Kuśnierczyk**, et al. (2007). *Towards Naming Conventions for Use in Controlled Vocabulary and Ontology Engineering*. Proceedings of the 10th Bio-Ontologies SIG Workshop, ISMB/ECCB, Vienna, Austria [311].[29]

P15. **Kuśnierczyk** (2007). *Taxonomy-Based Partitioning of the Gene Ontology*. To appear in Journal of Biomedical Informatics [221].

P16. **Kuśnierczyk** (2007). *Taxonomic Partitioning of the Gene Ontology* Proceedings of the Dagstuhl Seminar Towards Interoperability of Biomedical Ontologies, Schloß Dagstuhl, Germany [220].[30]

P17. **Kuśnierczyk** (2007). *The Logic of Relations Between the Gene Ontology and the Taxonomy of Species*. Proceedings of the 10th Bio-Ontologies SIG Workshop, ISMB/ECCB, Vienna, Austria [219].

P12 presents a critical assessment of the literature on ontological engineering, and provides evidence for the claim that there is much terminological

[29] An extended report on this effort is in preparation for journal publication.

[30] This article has not been peer-reviewed, but it was presented and discussed at a forum of over 20 leading experts in the fields of knowledge representation and biomedical ontology.

confusion in this field, which reflects a deeply rooted incoherence between the philosophical-ontological views adopted by individual developers.[31] P13 presents an attempt undertaken to standardize both the terminology and the underlying philosophical theory of existence, much needed for assuring interoperability between various ontologies developed under the guidance of OBO Foundry.

Paper P14 discusses a number of terminological, nomenclatural, and typographical conventions worked out by, among others, the Metabolomics Standards Initiative (MSI)[32] and the Proteomics Standards Initiative (PSI)[33] ontology working groups. In that article, we explore how OBO could be improved in this respect, and propose for wider adoption a series of patterns that have already been tested in practice by the aforementioned groups. The activity has been directed towards reviewing the existing documentations in an effort to distill both common and conflicting conventions. The aim of this analysis was to overcome the present diversity and fragmentation and to determine what conventions should be commonly applied. In P14, we describe the results of that study: naming conventions that, we believe, should provide robust labels for controlled vocabularies and ontologies.

Articles P15–P17 reflect my work on establishing semantically clear links between the Gene Ontology (and, in an extension, other OBO ontologies) and the Taxonomy of Species. P15 discusses the problem of subsetting the Gene Ontology with respect to various criteria, and explores the usefulness of the so-called 'GO slims' (custom, hand-made subsets of GO terms) for automatically producing species-specific subsets. It also describes, informally, a framework designed to give support for systematically linking the GO and the Taxonomy of Species.[34] P16 presents a formalization of that framework, and P17 includes further extensions to what is covered in P16. Recently, the approach presented in papers P15–P17 has been suggested by the OBO Consortium as a generic framework for annotating terms in OBO ontologies

[31] Quite often, this incoherence seems to reflect negligence or ignorance rather than consciously made decisions.

[32] http://msi-ontology.sourceforge.net/; see also Sansone et al. [306].

[33] http://www.psidev.info/; see also Hermjakob [164].

[34] Specifically, the NCBI implementation of the Linnaean Taxonomy. See Appendix B for more details and discussion.

with terms in the NCBI Taxonomy.[35]

In another article, I present an attempt to logically formalize the meaning of gene and gene product annotations with terms in the Gene Ontology:

P18. **Kuśnierczyk** (2007). *What Does a GO Annotation Mean?* Proceedings of the 10th Bio-Ontologies SIG Workshop, ISMB/ECCB, Vienna, Austria [222].

That paper describes work in an initial stage; the formalism used in that article is not discussed in much detail. GO annotations have been informally discussed in Hill et al. [166] and Blake et al. [47]. The content of P18 is not further discussed in this thesis.

Parts of the material included in Ch. 2 have also been presented at the 2007 Workshop on Ontologies, Standards and Best Practices in Ghent, Belgium, as an invited talk on biomedical ontologies — foundations and principles of design.

1.1.4 Research Summary

This section briefly characterizes the goals, questions, methods, and results of the three phases of my research introduced above.

Phase I In this phase, the ultimate goal was:

G1. To develop computational models of gastric acid secretion and gastrointestinal neoplasia.

One of the central interests of the reseach group was to use microarray data, and thus it was important to address the following questions (in addition to questions related directly to the application domain):

[35] http://www.obofoundry.org/wiki/index.php/Species_specificity.

Q1. How can microarray data be used for computational modeling in gastrointestinal research?

Q2. How should microarray experiments be designed and analyzed to deliver high-quality data and meaningful results?

Q3. Which of the available machine learning and data mining technologies are suitable for the modeling?

Furthermore, one of our primary interests was in applying the rough set-based technology, and thus the following question was also in place:

Q4. How suitable for the purpose are rough set-based classifiers, and what improvements can be made to further adapt the technology to the task?

The research methodology used included:

M1. Laboratory procedures, such as hybridization and scanning of microarray slides according to established protocols where available (I have participated in the laboratory part of some of the experiments).

M2. Statistical procedures, including exploratory analysis, filtering, normalization, and hypothesis testing according to established standards where available.

M3. Data mining procedures, including implementation and application of rough set-based classifiers, evaluated according to established standards, e.g., using receiver-operating characteristics (ROC).[36]

Scientific contributions made in this phase include:

C1. Extending scientific knowledge about the pathogenesis of gastric acid secretion and gastric cancer, as documented in P2–P8. My contribution includes design and implementation of statistical methods for

[36] See, e.g., Fawcett [112] for an introduction.

processing raw data, as well as application of various rough set-based and other classifiers.

C2. Extending knowledge about the applicability and performance of rough set-based classifiers in the context of functional genomics, specifically gene function prediction from microarray gene expression data, as documented in P1.[37] My contribution includes design and implementation of rough set classifiers, as well as the assessment of statistical significance of the results of subsequent classifications.

Phase II In this phase, the ultimate goal was:

G2. To design and build a tool (or extend an existing one) that would enable automated construction of biological association networks using reasoning over general domain knowledge and past experience (user profiles).

Questions that had to be addressed include:

Q5. What sources of general domain knowledge are available, what is their quality, and how can they be accessed?

Q6. What sources of task-specific, episodic knowledge are available, and how can such knowledge be represented?

Q7. How does a biologist choose relevant pieces of information while constructing a biological association network, what sorts of knowledge are needed for making the decisions?

Since our group had a growing interest in the then newly developed Gene Ontology, the following more specific questions were also relevant:

Q8. How can the Gene Ontology be used to analyze gene expression patterns, how does the structure of the ontology challenge the established statistical methodology?

[37] As well as Kuśnierczyk and Sommervik [225]; see footnote 9 on page 4.

Q9. How can the Gene Ontology (and other OBO ontologies) be used to guide automated construction of biological association networks?

The research methodology used included:

M4. Critical review of existing structured sources of biomedical knowledge (i.e., ontologies), assessment of how state-of-the-art automated text mining can provide additional information.[38]

M5. Collection of episodic knowledge by direct observation and discussion of experts' performance and experiences during network construction sessions.

M6. Experimentation with various statistical and machine learning approaches to analyze and classify gene expression data based on the structure of the Gene Ontology, with theoretical and empirical validation.

M7. Implementation of some of the ideas and assessment of their utility based on external users' experiences.

Contributions made in this phase include:

C3. Extending the field of biomedical data mining with a novel method and a tool for analyzing gene expression patterns based on annotations with Gene Ontology, as presented in P9. My contribution includes conceptual support during the design and implementation phases.

C4. Design and partial implementation of a system for automated reasoning over large biological association networks, as presented in P10 and P11. Most of the work here was done by me.

[38] I have marginally participated in biomedical text mining-related research conducted by my colleagues at NTNU (see, e.g., Sætre [303]), but this activity was not essential for my further work.

Phase III In this phase, the ultimate goal was:

G3. To improve the structure of and interoperability between biomedical ontologies, with particular focus on the Gene Ontology, its relatives in the Open Biomedical Ontology framework, and their relationships with the Taxonomy of Species.[39]

Questions that had to be answered included:

Q10. To what extent are the existing biomedical ontologies based on common design principles, and what are they, if any?

Q11. Is it desirable and feasible to impose a standard upper-level ontology which all the biomedical ontologies would have to explicitly refer to?

Since one of the key achievements of the research team led by prof. Smith was the Basic Formal Ontology (BFO),[40] it was natural to pose the following questions:

Q12. To what extent can the existing biomedical ontologies be forced to modify their structure and content in order to be BFO-compliant?

Q13. Would such enforcement be desirable and beneficial?

Q14. How should BFO be modified or extended, if necessary, in order to be better suited for the purpose?

Furthermore, we focused on the issue of explicitly connecting the Gene Ontology with the Taxonomy of Species, attempting to answer the following questions:

[39] One should rather speak of *a* taxonomy of species, as there are a few competitive approaches to the classification of organisms. However, the Gene Ontology explicitly refers to the NCBI Taxonomy, one of a few implementations of the Linnaean Taxonomy of Species, thus my study was focused on that taxonomy as a reference. See Ch. 4 and Appendix B for more details.

[40] http://www.ifomis.uni-saarland.de/bfo/.

Q15. How does OBO address the issue of relations between GO terms and taxa, is the solution clean and efficient?

Q16. How could this problem be solved alternatively in order to improve the structure and usability of the Gene Ontology?

Q17. How suitable for the task is the Linnaean Taxonomy, what are its structural properties as a hierarchical classification system, what are its implementations and how reliable they are?

The approach to address these questions was based on:

M8. Critical review of existing biomedical ontologies, comparison of their design principles (if any), identification of weaknesses; the assessment process included extensive communication with both authors and users of OBO ontologies.

M9. Critical review of the Basic Formal Ontology, comparison with a few other top-level ontologies, assessment of its applicability in the biomedical domain; the process included extensive communication with both authors and users of BFO.[41]

M10. Analysis of the structure and content of the Gene Ontology and the Taxonomy of Species,[42] with focus on how relations between these two are addressed in species-specific subsets of the GO.[43]

M11. Experimentation with various designs for a framework that would allow to connect the GO and the Taxonomy in an explicit and consistent manner.

These efforts had the following results:

[41]While BFO is currently in the process of adoption as the top-level ontology for OBO, it is by no means accepted by all potential users, and there are hot debates on many fundamental issues addressed, or addressable, by BFO. See http://groups.google.com/group/bfo-discuss.

[42]The NCBI Taxonomy database, http://www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy.

[43]The so-called 'GO slims'; see Ch. 4.

C5. A collection of design principles for biomedical ontologies have been proposed, as documented in papers P12–P14. I have contributed to gathering and presenting evidence for the thesis that the lack of such standards is undesirable, as well as to defining the core terminology for OBO ontology developers, and to clarifying basic principles for ontology design.

C6. A novel framework for connecting the GO and the Taxonomy has been proposed and discussed with several members of the Gene Ontology Consortium. As discussed in papers P15–P17, the framework addresses the issue of partitioning of the GO in various taxonomic contexts in a more structured and flexible way than it is currently done using GO slims, and should thus augment the usability of the GO, as well as provide support for species-dependent error checking at annotation time. Following a recent decision of the Gene Ontology Consortium, the framework is currently being experimentally incorporated into a non-public copy of the GO, and an empirical evaluation will follow.[44] The framework has also been suggested a candidate for a plugin to the ontology curation tool OBO-Edit (Day-Richter et al. [88]).

C7. Paper P18 presents an attempt to formalize the meaning of GO annotations. Its main contribution is to show that the informal explanation given in Hill et al. [166] is not entirely unambiguous, and that further specifications are necessary to provide for a shared understanding.

1.2 Thesis Overview

The thesis reflects mostly the third phase of my research (Sec. 1.1.3). The rest of this document is organized as follows:

– Chapter 2 provides an extensive introduction to the problem of integration of biomedical data. It discusses sources of the problem —

[44] Personal communication with Jennifer Deegan, GO curator, most recently in October 2007. See also http://www.obofoundry.org/wiki/index.php/Species_specificity.

in particular, high-throughput screening technologies (Sec. 2.4); the need for integration (Sec. 2.5); attempts to standardize storage and communication protocols (Sec. 2.6); and the role biomedical ontologies are intended to play in solving the problem (Sec. 2.7).

– Chapter 3 focuses on practices in ontological engineering. It provides evidence that there is, in many cases, a dose of uncertainty as to what elements of an ontology represent, and which of the available representational elements should be used to represent various entities in the domain (Sec. 3.2). Section 3.3 describes an effort undertaken by various parts of the OBO community, aimed at building a coherent basis for the development of biomedical ontologies. It is an effort I have joined only recently, and the work is still in progress.

– Chapter 4 introduces the issue of dependency of terms in the Gene Ontology terms on taxonomic contexts, the problem of subsetting the Gene Ontology accordingly to a chosen context, and the need for a systematic approach to link terms in the GO with terms in the Taxonomy (Sec. 4.1 and 4.2). Section 4.3 examines GO slims, the currently used approach to (manually) subsetting the Gene Ontology.

– Chapter 5 introduces a framework for systematically linking GO terms with taxa. Section 5.2 specifies a set of quantification patterns that can be used to define a GO term's validity, specificity, and relevance for a taxon. Section 5.3 shows how these patterns can be propagated up- and downwards along the hierarchy of both the GO and the Taxonomy, and Sec. 5.4 provides examples of how the patterns can be used to automatically partition the Gene Ontology.

– Chapter 6 summarizes the thesis, reconsiders the research questions and the way they have been addressed, and discusses ideas for further research.

– Appendix A introduces a logical formalism designed to formally present the properties of the framework. Section A.2 specifies a simple syntax and semantics, and Sec. A.3 defines rules of inference.

– Appendix B takes a closer look at the Taxonomy of Species and its implementations (i.e., taxonomic databases). It reviews the ontological problem of species (Sec. B.2), taxonomic classification and nomenclature schemes (Sec. B.3), and discusses some further problems with the Taxonomy (Sec. B.4).

Chapter 2

Background

This chapter provides background information on bioinformatics and the challenge of data integration in molecular biology, and a stepwise introduction to biomedical ontologies as a proposed solution to problems related to the integration. Some of the questions addressed here are: What are the reasons underlying the rapid increase in the amount of experimental data in molecular biology? How are the data made accessible, how can they be integrated and turned into information useful for biologists? How can the data be made understandable to automated agents, how can biomedical ontologies help?

The chapter is structured as follows:

- Section 2.2 explains what bioinformatics, computational molecular biology, and other related fields of research are concerned with.
- Section 2.3 briefly discusses the terms 'data', 'information', and 'knowledge', which are often used in the bioinformatics literature.
- Section 2.4 introduces the idea of high-throughput screening, a collection of technologies that allow biologists to produce massive amounts of data.

- Section 2.5 reviews attempts targeted at integration of the data so that it can be uniformly presented to the human user, irrespectively of the distributed nature of the underlying sources.
- In Sec. 2.6 we take a look at the efforts made to standardize the storage formats and communication protocols, intended to ensure that automated agents can easily access established and newly appearing resources with minimized support from the human user.
- Section 2.7 introduces biomedical ontologies as a promise to equip both human users and computer systems with knowledge necessary for efficient and successful navigation through the vast amounts of data.

2.1 Introduction

Scientific inquiry inevitably leads to the production of information that, in order to be reusable, has to be stored in some form — as images, free-text scientific publications, structured entries in relational databases, etc. However, the fact that it is *stored* is not yet a guarantee for its reusability; it has to be *accessible* in such a way that a potential user can browse, search, filter, or otherwise select parts of the information relevant to the particular study at hand. Furthermore, the data retrieved from a database have to be interpreted in the right way to be useful as an information bearer and to provide support for contributions to the scientific knowledge.

Molecular biology and its close relatives, computational molecular biology ('computational biology' for short) and bioinformatics are currently at a stage in which data of various sorts are produced with ever increasing speed, which in turn motivates the development of computational approaches to processing, analyzing, interpreting, and presenting the data as information in a form and amount comprehensible for the human user. It is typical for publications that present new bioinformatics tools to underline the overabundance of data and the need for integration; terms such as 'information explosion' are used by their authors (e.g., Lloyd et al. [238]) to stress the

shifting of modern biology from the so-called 'one gene one postdoc' approach to genomic analyses that include the simultaneous monitoring of thousands of genes. It is also widely recognized that the amounts of genomic and other data, which are already too large to be studied by human researchers in detail element by element, will continue to increase at an even increasing pace (Brazma [56]).

It is clear that there is an increasing need for efficient access to concise and integrated biomedical information to support data analysis and decision making. The complexity of biological systems, and the vast amount of information now available at the level of genes, proteins, cells, tissues and organs, requires the development of mathematical models that can define the relationship between structure and function at all levels of biological organization (Hunter et al. [176]). It appears, however, that knowledge discovery in the widely scattered resources relevant for biomedical research is often a cumbersome and non-trivial task, one that requires significant training and effort (Rebhan et al. [290]). In an attempt to remedy the situation, various tools emerge that facilitate interpretation of biological data in a batch mode, rather than on a gene-per-gene basis. Such tools, however, often leave the investigator with large volumes of apparently unorganized information (Beisvåg et al. [34]). The use of explicitly represented general biological knowledge, e.g., in the form of biomedical ontologies, is often proposed as a solution to the data integration problem (Fedoroff et al. [113], Smith [323]).

2.2 Bioinformatics and Computational Biology

Bioinformatics is a young research field, focused on the development of computational tools that help biologists process, store, and draw inferences from the data they generate in wet-lab experiments. Bioinformatics and computational biology involve the use of techniques including applied mathematics, informatics, statistics, computer science, artificial intelligence, chemistry, and biochemistry to solve problems in biomedical research and clinical practice. While the terms 'bioinformatics' and 'computational biology'

(as well as 'computational molecular biology', 'biomolecular informatics', etc.) are often used interchangeably, the Biomedical Information Science and Technology Initiative Consortium (BISTI)[1] of the National Institutes of Health (NIH) has proposed the following definitions that should help in distinguishing the fields:

> *Bioinformatics*: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.
>
> *Computational Biology*: The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.[2]

Some authors (e.g., Krane and Raymer [213]) underline that the primary interest of bioinformaticians is to *develop* algorithms that can be applied in analyses of biomedical data, while the primary interest of computational biologists is to *apply* such algorithms in particular biological studies. Both fields are interested in algorithms and computation, but for one of them these are research targets, for the other they are research tools.

Fundamental textbooks on bioinformatics and computational biology are *Introduction to Computational Biology* (Waterman [379]) and *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology* (Gusfield [154]). Other interesting sources are Baldi and Brunak [27], Krane and Raymer [213], Pevzner [281], Jones and Pevzner [186], and Ewens and Grant [108], and other. For a recent overview of trends in bioinformatics, see, e.g., Perez-Iratxeta [280], or Altman [13].

Related to bioinformatics and computational biology is *systems biology*, focused on studying interactions between components of complex biological

[1] http://www.bisti.nih.gov/ .

[2] http://www.bisti.nih.gov/CompuBioDef.pdf, NIH Working Definition of Bioinformatics and Computational Biology, July 17, 2000.

systems.[3] Research in systems biology is highly dependent on the availability of the computational methods of bioinformatics. For an introduction to systems biology, see, e.g., Kitano [203, 202] and Alon [11]; for more on the application of computer science methods in systems biology, see Dubitzky [95], Burrage et al. [64], Lilburn et al. [236], Larrañaga et al. [231], etc.

2.3 Data, Information, Knowledge

'Data', 'information', and 'knowledge' are three terms often used in the context of bioinformatics. Intuitively, knowledge is usually distinguished from data and information, while the terms 'data' and 'information' are often used interchangeably. It has been argued (see, e.g., Aamodt and Nygård [7]) that the unclear distinction between data, information, and knowledge impairs their combination and utilization for the development of integrated systems. There are a number of attempts to provide concise, comprehensible definitions, and while we are not willing to engage in a lengthy discussion, it is desirable to provide at least a brief introduction of these terms.

From the knowledge management perspective the following definitions have been proposed:

> *Data*: unorganized and unprocessed facts; a set of discrete facts about events — structured records of transactions.
>
> *Information*: an aggregation of data that makes decision making easier; facts and figures based on reformatted or processed data.
>
> *Knowledge*: human understanding of a specialized field of interest that has been acquired through study and experience. (Awad and Ghaziri [20])

[3] In a review of three recently published books on systems biology, Eric Werner says, "Systems biology is not as new as many of its practitioners like to claim. It is a mutated soup of artificial life, computational biology and computational chemistry, with a bit of mathematics, physics and computer science thrown in" (Werner [382]).

In his *Semantic Conceptions of Information*, Floridi [115] provides an in-depth discussion of data, information, and the differences between them. He provides the following definitions:

> *Diaphoric Definition of Data*: A datum is a putative fact regarding some difference or lack of uniformity within some context.
>
> *General Definition of Information*: σ is an instance of information, understood as semantic content, if and only if:
>
> – σ consists of one or more data;
> – the data in σ are well-formed;
> – the well-formed data in σ are meaningful.

This, of course, leaves open the question of what 'well-formed' and 'meaningful' mean; the reader is referred to the original text for further discussion. Note that, according to this definition, anything can be regarded as data — just any non-uniformity in the real world is a datum. From the perspective of artificial intelligence, Aamodt and Nygård [7] suggest to define data, information, and knowledge as the roles, or in terms of the roles that syntactic entities — e.g., symbolic assertions — may play in decision-making processes of reasoning agents:

> *Data*: syntactic entities, patterns with no meaning. They are input to an interpretation process, i.e., to the initial step of decision making.
>
> *Information*: interpreted data, data with meaning. It is the output from data interpretation as well as the input to, and output from, the knowledge-based process of decision-making.
>
> *Knowledge*: learned information, information incorporated in an agent's reasoning resources, and made ready for active use within a decision process; it is the output of a learning process.

In philosophy, knowledge is usually defined as *justified true belief* — this is the so-called 'JTB' account of knowledge (Steup [348]) — with a number of further modifications intended to address various problems inherent in this definition.[4] In essence, according to the JTB view knowledge is a belief

[4]For example, the Gettier problem (Steup [349]).

which is true, and the believing agent has good reasons to hold the belief. In the definition above, the belief is justified by the process of learning, though there is no explicit demand on the belief's being correct.

2.4 Biomedical Data

As in the case of any other empirical science, research in biology is based on experimentation and involves gathering of data as an essential activity. While the data come from observations of phenomena in the real world, they are usually treated as an intermediate result that confirms or negates hypotheses made *a priori*, and from which theories are inferred by means of abduction.[5] It is general truths about what generally holds in reality — what some might wish to call the 'laws of Nature' — rather than the individual phenomena, events, etc. reflected in the data, that are the ultimate target of scientific inquiry. Traditionally, scientific publications only occasionally include pieces of raw experimental data; observations are usually summarized statistically, and conclusions are presented as theories — generalized claims.

While the interest of science in general truths has not changed, recent developments in, among others, database and networking technologies allow for the storage and reuse of virtually all of the experimental data being produced, without the need for immediate generalization. Researchers are encouraged, and in increasingly many situations even required, to supplement their publications with raw (unprocessed) experimental data. Recent advances in molecular biology, genetics, biochemistry, and other related fields have been accelerated by the availability of newer, faster, and increasingly more automated technologies. Where only a few decades ago most of the experimental work was done manually and in a very time-consuming manner, complicated experiments can now be run on computerized devices with only a few mouseclicks. There has been a decrease in the time needed to perform measurements of various properties of biological entities — mul-

[5] Abduction is usually defined as inference to the best explanation, a non-monotonic or defeasible inference (Josephson and Josephson [187]).

ticellular organisms, cells, organelles, biomolecules; the number of such entities that can be simultaneously observed and individually characterized within a single experiment has increased substantially.

Computational systems biology, or simply systems biology (SB),[6] a relatively new field of study — or, according to some, a whole new research paradigm — focuses on the observation and description of biological entities as complete functioning systems. This new approach is contrasted with the traditional reductionistic practice of describing a complex system by means of combining descriptions of its components observed one at a time. While the latter approach has successfully identified most of the components of living organisms and many interactions between them, it does not seem to offer enough convincing insights and methods to comprehend how the properties of whole systems emerge from the properties of their parts studied separately. Rather than from the reductionist viewpoint,[7] the pluralism of causes and effects in biological networks is better addressed by observing, through quantitative measures, multiple components simultaneously, and by rigorous data integration with mathematical models. Such a systemwide perspective (that is, the systems biology approach) on component interactions is required so that whole networks of components can be quantitatively understood and rationally manipulated (Sauer et al. [307]).

The success of the systems biology approach depends on the availability of experimental and computational technologies that allow one to generate, process, analyze, and present huge amounts of data. One of the cornerstones of the development of SB was the invention of automated high-throughput screening technologies (HTS);[8] In brief, high-throughput screening is the process of quantitative testing of a large number of entities — organisms, genes, chemical structures — in particular, molecules of mRNA, proteins, etc. Compared to traditional screening methods, HTS is characterized by its simplicity, rapidness, low cost, and high efficiency. Various technologies are now available, and the screening of more than 100,000

[6] For an introduction, see the seminal articles by Kitano [202, 203]; for an overview of recent activities in SB, see the website of the Institute for Systems Biology, http://www.systemsbiology.org/.

[7] For more on genetic reductionism and related issues, see, e.g., Waters [380].

[8] See, e.g., Burbaum and Sigal [61], Fan et al. [111], Shendure et al. [315].

samples per day is not unusual. Another important aspect of HTS is that many variables can be tested simultaneously for the same sample; it is thus possible, for example, to assess the expression of tens of thousands of genes from cells of a single line subjected to diverse experimental conditions (Liu et al. [237]).

Without a doubt, the application of HTS technologies has already dramatically increased the amount of data produced by biological experimentation, and the increase rate is not linear; parallel high-throughput experiments are generating increasing data volumes at an ever more rapid pace (Hekkelman and Vriend [163]). It is not uncommon nowadays to speak of *massive* biological data, and data sets from biological experiments are often the motivation for work presented on conferences and workshops dedicated to mathematical and computational approaches to handling massive data, e.g., the Workshop on Algorithms for Modern Massive Data Sets (MMDS)[9] or the International Conference on Very Large Databases (VLDB).[10]

One of the best known and widely established HTS technologies for molecular biology are microarrays (MA).[11] Various sorts of microarrays can be distinguished, depending on the type of biological material used, for example:

- DNA microarrays, one of the most important applications for arrays so far in monitoring of gene expression (Lockhart and Winzeler [239]);
- protein microarrays, which allow for rapid and multiplex screening of thousands of samples on a single microarray with applications in, e.g., drug screening, metagenomics,[12] and high-throughput enzyme assays (Angenendt et al. [15]);
- cell microarrays, with which it is possible to study transfection with

[9] http://www.stanford.edu/group/mmds/.

[10] http://aitrc.kaist.ac.kr/~vldb06/index.html.

[11] See., e.g., Schena et al. [309] for an early introduction, and Duyk [97] for a more recent account.

[12] Metagenomics is the study of genomes from samples obtained from organisms found in their natural environments, as opposed to artificially cultured ones; see, e.g., Eisen [99].

thousands of different RNAi[13] reagents on a microarray slide (Wheeler et al. [384]);

– tissue microarrays, where portions of thousands of different tissue samples can be analyzed on one microscope glass slide simultaneously (Simon et al. [316, 317]).

Microarrays based on other types of material, such as those focused on protein-DNA interaction, protein-RNA interaction, on-chip translation, protein binding, etc., are under development (Hoheisel [171]). The availability of high-throughput quantitative screening has led to the invention of a whole range of new fields of holistic biological inquiry, such as genomics (the study of entire genomes, complete sequences of an organism's genetic material); proteomics (the study of proteomes, complete collections of proteins expressed by a genome); transcriptomics (the study of entire collections of mRNA molecules — transcripts); metabolomics (the study of metabolomes, whole collections of small molecules — metabolites);[14] interactomics (the study of interactomes, entire sets of interactions between proteins), etc. Other examples include glycomics, lipidomics, spliceosomics, phenomics, reactomics, etc. The Cambridge Healthtech Institute maintains an -omics and -omes glossary and a taxonomy[15] with statistics of Google search hits and references to publications on new omics fields, under the motto "we have entered the 'omic' era in biology". The top-class scientific journal Nature maintains what they call the 'omics Gateway':

> "Biology has become an increasingly data-rich subject, and NPG[16] is committed to helping the community mine those data for novel insight. Many of the emerging fields of large-scale, data-rich, biology are designated by the suffix "-omics" added onto previously used terms. The importance to the life science community as a whole of

[13] RNA interference (RNAi) is the phenomenon of blocking transcription of genes by short interfering RNA molecules (siRNA); see, e.g., Mello and Conte Jr. [251] and other articles in the RNA Interference volume of Nature Insight.

[14] "Metabolomics is the newest 'omics' science" (Claudino et al. [82]).

[15] http://www.genomicglossaries.com/content/omes.asp.

[16] Nature Publishing Group.

> such large-scale approaches is reflected in the huge number of citations to many of the key papers in these fields. The Omics Gateway provides life scientists a convenient portal into publications relevant to large-scale biology from journals throughout NPG."[17]

While the underlying principle of omics fields is a holistic view of their domains of study, they inevitably divide the domain of molecular biology into smaller subdomains studied separately. Recently, the term 'omeomics' has been proposed as a name for the science that successfully integrates various omics approaches, and 'omeome' as a name for the catalog of all omics sciences.[18]

DNA microarrays are one of the earliest introduced and most popular microarray HTS technologies. With DNA microarrays, one can observe (and quantify) the expression of tens of thousands of genes simultaneously — indeed, the expression of whole genomes, such as the human genome with its approximately 25,000 protein-coding genes (the International Human Genome Sequencing Consortium [180]). In brief, a typical microarray gene expression experiment (MAGE) involves the following steps:

- collection and preparation of biological material from experimental animals, plants, or cell cultures;
- purification of the material, extraction of the messenger RNA (mRNA) fraction, and reverse transcription (RT) of the mRNA to complementary DNA (cDNA), and labelling of the material with fluorescent probes;
- distribution of the cDNA onto a collection of microarray plates (also called 'slides' or 'chips') on which there are printed spots containing small amounts of standardized biological material (probes from probe libraries) with known biochemical properties, e.g., oligonucleotides;
- incubation of the plates in standard environmental conditions, during

[17] http://www.nature.com/omics/.

[18] http://www.omeomics.org/. Omeomics thus defined seems to overlap with Systems Biology; so far, there are no serious publications under the hood of omeomics, and this example of terminological creativity probably remains in the sphere of mere hype.

which cDNA molecules from the investigated material hybridize (bind specifically) with probes on the slides;

- scanning of the slides with a laser scanner, to assess the amount of cDNA molecules bound to each probe;
- analysis of the images obtained from the laser scanner, a process which involves the classification of single pixels in the images as corresponding to the spots or to their surrounding background;
- statistical correction of systematic errors due to variation in the efficiency and accuracy of printing, labelling, hybridization, and scanning;
- statistical analysis of the resulting data, e.g., selection of differentially expressed genes;
- further analysis of the data in the light of previously published evidence, data available from other experiments, and general knowledge.

For a comprehensive introduction to microarray technology, microarray data analysis, and related issues see, e.g., Knudsen [206], Kohane et al. [207], or Schena [308]. For a relatively recent overview of the developments in this field see Barrett [29] and other articles in The Chipping Forecast III issue of Nature Genetics, as well as previous editions in this series.

The Moore's law (Moore [259]) predicts an increase in the number of transistors in an integrated electronic circuit (a microchip) of the order of two every two years.[19] An analogous trend can be described in the case of the increasing amount of biomedical data.[20] For example, in 2002, to study growth factor-responsive genes in neuroendocrine gastrointestinal tumour cells, we used custom-made microarray plates manufactured by the Norwegian Microarray Consortium,[21] printed with over 5,000 probes[22] (Hof-

[19] http://www.intel.com/technology/mooreslaw/.

[20] In 1996, Hooft et al. [172] observed that "more than 4,700 data sets are available, and the number is expected to double every 18 months" (about the Protein Data Bank, PDB).

[21] http://www.mikromatrise.no/

[22] In the context of DNA microarrays, a probe is a relatively short molecule of DNA (an oligonucleotide) immobilized (printed onto) a microarray slide, to which molecules of RNA or complementary DNA (cDNA) from the tested sample bind with high degree of sequence specificity (Kohane [207]).

sli et al. [169]). In 2003, to study liver gene expression in rats in response to the peroxisome proliferator-activated receptor-α agonist ciprofibrate, we used microarrays printed with over 7,500 elements (Yadetie et al. [393]). As of 2007, newer, commercially available microarray chips, such as the Affymterix GeneChip Human Genome U133 Plus Array,[23] contain as many as 1,300,000 oligonucleotide features. (Printed (spotted) cDNA MAs differ from photolitographic oligonucleotide MAs such as the Affymetrix GeneChips in that in the latter case the oligonucleotides are shorter, and the microarrays contain a number (usually in the order of 10) of slightly different oligonucleotides per gene, so that the numbers of features on a single array are not directly comparable between these two MA types. Nevertheless, these numbers are illustrative of the trends in miniaturization and growing coverage of the microarray technology.) The use of microarrays has contributed substantially to the flood of biomedical data (Hoheisel [171]). It is said that no other methodological approach has transformed molecular biology more in recent years than the use of microarrays. Microarray technology has led the way from studies of the individual biological functions of a few related genes, proteins or, at best, pathways towards more global investigations of cellular activity. The development of this technology immediately yielded new and interesting information, and has produced more data than can be currently dealt with (Hoheisel [171]).

The development and accessibility of hardware and software technologies for experimentation in molecular biology has been parallelled by a rapid increase in the number and size of publicly available molecular biology databases,[24] ranging from comprehensive, omics-oriented, multispecies ones such as Ensembl (Hubbard et al. [174]) and NCBI database resources (Wheeler et al. [385]), to those dedicated to narrower domains, such as the tissue-specific human enhancers database VISTA (Visel et al. [376]) or the HIV positive selection mutation database (Pan et al. [275]). And, of course, there are quite a few databases providing microarray gene expression data, e.g., the Stanford Microarray Database storing data generated from more than 60,000 microarrays (Demeter et al. [90]), the EMBL-EBI ArrayEx-

[23] http://www.dnavision.be/pharmacogenomics/affymetrix_expression.php

[24] For evidence and history of this trend, see the annual summaries of publicly available databases, published in the Database Issues of Nucleic Acids Research (Galperin [119, 120, 121, 122]).

press, a rapidly growing database, which currently contains data from over 1,500,000 individual expression profiles (Parkinson et al. [276]), and the NCBI Gene Expression Omnibus, containing tens of millions of expressions profiles (Barrett et al. [30]). For orientation, Tab. 2.1 shows the number of entries in a few other databases.

Database	Domain	Approximate size
dbSNP	genetic variations	34,000,000
EMBL Nucleotide Sequence	DNA sequences	80,500,000
Entrez	DNA and protein sequences	91,000,000
GenBank	DNA sequences	61,000,000
IntAct	molecular interactions	126,000
NCBI Taxonomy	named organisms	240,000
PubMed	scientific publications	16,500,000
UniGene	gene-oriented sequence clusters	1,200,000
UniSTS	sequence-tagged sites	500,000

Table 2.1: Approximate sizes of selected molecular biology databases as of December 2006 (Benson et al. [37], Kerrien et al. [197], Kulikova et al. [214], Wheeler et al. [385]).

The amounts of data that are already available are, on the one hand, encouraging: it may seem that the more data we have, the more we know. On the other hand, it is no longer possible to perform analyses by hand, and we need tools — computer programs — that can find, download, process and integrate data automatically. Methods of computer science are essential for effective handling and making sense of the data, and for rendering them accessible to biologists working on a wide variety of problems. This is one of the major challenges of bioinformatics today; we are swimming in a rapidly rising sea of data — how do we keep from drowning? (Roos [300]). In the field of information processing, there arises what can be called the 'database Tower of Babel' problem: different groups of researchers use their own idiosyncratic terms and concepts to represent the information they produce and release to the public. To put this information together, methods must be found to resolve terminological and conceptual incompatibilities

(Smith [318]). Many attempts to solve this data integration problem have been made. One possibility is to implement tools that can present data from multiple databases in a uniform way, thereby hiding the implementational details from the user. We shall have a look at such solutions in the next section.

2.5 Data Integration

To become meaningful, data must be interpreted. For this to make practical sense, the way in which users interpret the data should closely corespond to the way in which the creators of the data intended them to be interpreted. To ensure this, data must be formatted and tagged in a way that leaves no doubt as to what the data are about.[25] Furthermore, data from various sources have to be integrated, and only then will they give us an all-covering image of the biological reality. The true value of the data that is being generated in omics experiments will become visible come when we turn from analyzing microarray and proteomic and metabolomic and all other data sets independently, and rather combine them to provide a single coherent view of the fundamental biological phenomena. Consequently, we need to consider methods that will place these diverse data into a common reference frame that can organize the information in a manner facilitating its interpretation (Quackenbush [288]).

Until not so long ago, molecular biology databases were organized into collections of so-called 'flat files', where data were formatted in separate, tagged lines of human-readable text.[26] Each database had its own idiosyncratic format, and to retrieve relevant pieces of information one had to use database-specific tools that could parse, filter, and present a selection of the content of flat files to the human user. (One could of course manually browse the files or use a simple text search tool, but for complex queries this

[25] There is a noteworthy element of circularity or regression in this requirement; any tag attached to a piece of data is itself a piece of data (and is thus sometimes called 'metadata'), and that data must be first interpreted as representing the tag.

[26] Some databases are still stored, if only partially, as collections of flat files.

approach is very inefficient.) While relatively convenient for manual curation, flat file formats severely restrict the usability of databases — to perform advanced, complicated queries, one had to master system tools such as *grep*, *sed* or *awk*,[27] or write programs in languages such as Fortran or Perl. The fact that a single database could employ a number of incompatible formats did not make the situation better. Ideally, the information should be presented in a way such that it can be parsed by a computer program correctly pulling out the relevant descriptions from the appropriate fields, and standard names should be used to describe common properties. In 2001, most sequence annotations did not meet these criteria. The difficulties in parsing annotations of most sequence databases were known to anybody who had tried to develop such parsers (Brazma [56]).

For example, consider the Protein Data Bank (PDB; Berman et al. [39]), a database that faced this problem:

> "One of the most difficult problems that PDB now faces is that the legacy files are not uniform. Historically, existing data ('legacy data') comply with several different PDB formats and variation exists in how the same features are described for different structures within each format." (Berman [40])

Figure 2.1 shows an excerpt from a PDB character-formatted flat data file, corresponding to an entry about the glutathione synthetase enzyme molecule. For comparison, Fig 2.2 shows an excerpt from a UniProt (The UniProt Consortum [368]) database flat file. While it is, arguably, easy to read for a human, it is not so easy to parse and search automatically. (It is certainly easy to parse once one has an appropriate parser. From the perspective of a computer scientist, implementing a parser for a clearly defined flat file format specific to a biological database is not a challenging task; however, most early software in computational biology has been created by biologists, and the obstacle was considerable.)

In their presentation of PDBFINDER, a database that "contains summary

[27] *grep*, a command line tool for searching patterns in text files using regular expressions; *sed*, a programming language for textual transformations on streams of data; *awk*, a pattern-matching programming language for working with text files [299].

```
COMPND      MOL_ID: 1;
COMPND    2 MOLECULE: GLUTATHIONE SYNTHETASE;
COMPND    3 CHAIN: A;
COMPND    4 SYNONYM: GAMMA-L-GLUTAMYL-L-CYSTEINE\:GLYCINE LIGASE
COMPND    5 (ADP-FORMING);
COMPND    6 EC: 6.3.2.3;
COMPND    7 ENGINEERED: YES
...
```

Figure 2.1: Excerpt from a Protein Data Bank flat file in the PDB Format. From PDB, http://www.wwpdb.org/.

information for all PDB data sets", Hooft et al. discuss a number of problems they met while writing Perl scripts for extracting information from PDB flat format files. For example,

> "For the 'Enzyme-Code' field the program contains seven different regular expressions (text patterns) that have to be tested sequentially. This is needed to recognize expressions like `(E.C. 3.2.1.27)` and `EC: 3.2.1.27;` as the same reference." (Hooft et al. [172])

One solution to this problem is to isolate the biologist from the implementational details by means of an interface that provides a transparent access to multiple sources of data in a uniform presentation layer. The Sequence Retrieval System (SRS; Etzold and Argos [104, 105]), for example, was designed as a system for indexing flat file libraries, intended to provide fast access to individual library entries via retrieval by keywords from various data fields; it included a "sophisticated parsing engine for information extraction" and effectively isolated the user — a biologist — from the specifics of flat file formats by means of a web-based graphical user interface (Etzold and Verde [106]). Another example is the MRS server, which allows for very rapid queries in a large number of flat-file data banks, such as EMBL, UniProt, OMIM, dbEST, PDB, KEGG, etc. (Hekkelman and Vriend [163]).

However, this solution works only if the biologist is interested in browsing or searching through the databases via a graphical, typically web-based

```
ID   GRAA_HUMAN              Reviewed;         262 AA.
AC   P12544;
DT   01-OCT-1989, integrated into UniProtKB/Swiss-Prot.
DT   01-OCT-1989, sequence version 1.
DT   07-FEB-2006, entry version 77.
DE   Granzyme A precursor (EC 3.4.21.78) (Cytotoxic T-lymphocyte proteinase
DE   1) (Hanukkah factor) (H factor) (HF) (Granzyme-1) (CTL tryptase)
DE   (Fragmentin-1).
GN   Name=GZMA; Synonyms=CTLA3, HFSP;
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Euarchontoglires; Primates; Catarrhini; Hominidae;
OC   Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [MRNA].
RC   TISSUE=T-cell;
RX   MEDLINE=88125000; PubMed=3257574;
RA   Gershenfeld H.K., Hershberger R.J., Shows T.B., Weissman I.L.;
RT   "Cloning and chromosomal assignment of a human cDNA encoding a T cell-
RT   and natural killer cell-specific trypsin-like serine protease.";
RL   Proc. Natl. Acad. Sci. U.S.A. 85:1184-1188(1988).
...
```

Figure 2.2: Excerpt from a UniProt flat file. Source: UniProt, http://www.ebi.uniprot.org/ .

interface. For queries that are not supported by the interface, and for programmatical analysis of large amounts of information — an activity essential for doing systems biology — this approach is unhelpful. Fortunately, the problem of having to parse idiosyncratically formatted flat files is largely historical; modern databases are stored using high-speed database management systems (DBMS),[28] though some still use flat files to store parts of their content. There are various types of database management systems: relational (RDBMS, such as MySQL[29]), object-oriented (ODBMS; for example, db4objects[30]), hybrid object-relational (ORDBMS; e.g., Caché[31]), and others. Many of them are available for a low fee, or at no cost at all, which is particularly welcome by database developers affiliated with publicly funded

[28] For the persistence of data a DBMS stores them in a collection of files, but these files are completely hidden behind an interface provided by a standard query language.

[29] http://www.mysql.com/.

[30] http://www.db4o.com/.

[31] http://www.intersystems.com/cache/.

research institutions. Furthermore, where data is accessible in files rather than through a database query language, the common solution now is to use XML as the basis for syntax definition, which virtually solves the problem of parsing.

Nevertheless, while data stored in DBMS-managed databases are accessible through a standardized data definition and query language, typically some flavour of SQL (Chamberlin and Boyce [73], Kline et al. [205]), and writing flat file parsers is not any more a prerequisite for programmatic access to the data, this does not solve the problem of data integration. In fact, one could argue that the availability of free to download and easy to install and run DBMSs made the problem even worse: everyone can now create and publish a database in minutes,[32] and as of 2007, there are almost 1,000 publicly available molecular biology databases (Galperin [122]). Despite being accessible through a standardized query language, the databases are still speaking in tongues: they are built using different and often incompatible schemas and nomenclatural conventions, and more often than not it is meaningless to run the same SQL query against two databases that provide data of the same sort.

Analogously as in the case of flat file-based databases, one solution is to build centralized services which allow a user to browse the content of multiple databases through a unified interface — for example, SOURCE (Diehn et al. [92]) or GeneCards (Safran et al. [304]). Datawarehousing is a solution based on bringing all the relevant data from multiple distributed sources into one integrated database; an example of biomedical datawarehouse is the Annotation Tool suite implemented by the Norwegian Microarray Consortium (Beisvåg et al. [34]). This approach, however, has not always been successful; one of the most ambitious attempts of this kind was the Integrated Genome Database (IGD; Ritter [297]) project, which aimed at combining human genome sequence data with multiple genetic and physical maps that were in the focus of human genomics at the time. At its peak, IGD integrated more than a dozen source databases, including GenBank, the Genome Database (GDB) and the databases of many human genetic-mapping projects. The IGD project survived for slightly longer than a year

[32] Populating the database may take much more time, though.

before collapsing (Stein [347]).

Some services offer programmatic access to integrated data via an application programming interface (API). For example, Ensembl (Hubbard et al. [174])[33] stores genome-centric data from a number of eukaryotic species, and provides a comprehensive set of APIs that serve as a middle layer[34] between the underlying database schemas and application programs. The APIs aim at encapsulating the database layout by providing efficient high-level access to the tables, and at isolating applications from changes in the underlying database schema. The Ensembl Perl API is based on the BioPerl Toolkit provided by the BioPerl community, an international open-source collaborative effort of biologists, bioinformaticians, and computer scientists.[35] BioPerl has evolved over the past decade into a comprehensive library of modules available for managing and manipulating life-science information (Stajich et al. [345]). While Bioperl enables a researcher to programmatically access various databases in a relatively uniform fashion, it still relies on database-specific implementational details provided in, e.g., separate modules supplied by the developer of the database. Figure 2.3 illustrates the issue with an excerpt from a Bioperl program that accesses the EMBL database to retrieve the sequence identified by the UniProt ID 'U14680'. Note that it is necessary to use an EMBL-specific module.

Other examples of code libraries that allow to programmatically access various molecular biology resources include BioPython (de Hoon et al. [89]), BioJava and BioJavax (Pocock et al. [285]), BioLingua and BioBike (Massar et al. [245]), etc. For a recent overview of libraries of molecular biology-related code written for a number of diverse programming languages, see, e.g., McGuffee [249] and the Open Bioinformatics Foundation website.[36] Stajich and Lapp [346] provide a survey of open source tools and toolkits for bioinformatics.

The challenge of integration, however, is not constrained to having to communicate with multiple sources of data. In parallel to the development of

[33] http://www.ensembl.org/.
[34] Some would use the term 'middleware' (Bernstein [42]) in this context.
[35] http://www.bioperl.org/.
[36] http://www.open-bio.org.

```
use Bio::DB::EMBL;
use Bio::SeqIO;

my $db = new Bio::DB:::EMBL();
my $seq = $db->get_seq_by_acc("U14680");
my $seqout = new Bio::SeqIO(-format => "genbank");
if (defined $seq) {
   $sequot->write_seq($seq);
}
```

Figure 2.3: A fragment of Bioperl code for retrieving a sequence from a remote database, using the Bio::DB::EMBL module (Stajich [345]).

databases, there is an even more rapid production of new tools for doing all sorts of analysis of the available data. The annual Database Issue of the Nucleic Acids Research journal was first released in 1993; since 2002, it has been accompanied by the Web Server Issue, which highlights many servers that are available on the internet to perform useful computations on DNA, RNA and protein sequences and structures; furthermore, it lists as a few servers that help to mine scientific literature or cover other aspects of biology (Nucleic Acids Research Database Issue 2006, Editorial [3]). As of 2006, there are over 1,000 web servers listed in the Bioinformatics Links Directory (Fox et al. [116]). Instead of implementing libraries of separate modules necessary to access every newly created database or analysis tool, enforcing those distributed resources to use standard representation schemas and communication protocols may, as one could expect, be a better solution to challenge of data integration. We shall discuss some of such standards in the next section.

2.6 Standardization

High-throughput technologies generate large amounts of complex data that have to be stored in databases, communicated to various data analysis tools

and interpreted by scientists. Data representation and communication standards are needed to implement these steps efficiently (Brazma [58]). The underlying idea is that once a representation or communication standard is established, tools may be implemented that would be able to access previously existing and newly created databases and other resources without knowing their implementational details. If new resources comply with the standards, there is no need for implementing new parsers, translators, etc. What is the status of such standardization efforts for information exchange in biomedicine? A recently published survey conducted by the World Technology Evaluation Center (WTEC)[37] compared activities of system biologists in the United States, Europe and Japan. The survey revealed absence of a suitable infrastructure for systems biology, particularly for data and software standardization, which is a major impediment to further progress (Cassman [69]).

Standardization efforts are aimed at solving the challenge of integration in a variety of ways. On the side of database architecture, the Generic Model Organism Database Toolkit (GMOD)[38] is a noteworthy example of an open source project providing a complete set of software components for creating and administering a model organism database. It includes a number of modules — genome visualization and editing tools, literature curation tools, a robust database schema, biological ontology tools, etc. Unfortunately, its design may impede integration with resources which are not based on GMOD. The design of the relational database schema Chado,[39] for example, includes a number of optimizations[40] which may make programmatic access to the database via object-relational mapping (ORM) frameworks (e.g., Hibernate for Java[41]) more difficult than necessary.

The extensible markup language (XML)[42] is one of the most successful attempts to standardize data exchange between applications; XML is a general-

[37] http://www.wtec.org/.
[38] http://www.gmod.org/.
[39] http://www.gmod.org/wiki/index.php/Schema/.
[40] The term 'optimization' was used by one of the developers of Chado to explain some of the peculiarities of the design (private conversation).
[41] http://www.hibernate.org/.
[42] http://www.w3.org/XML/; see also Harold and Means [160].

purpose, human- and machine-readable language[43] with a strict, easy to learn, write, and parse syntax. However, XML itself is not a data exchange language; it is a language which provides the basis for defining application-specific or domain-specific data (information, knowledge) representation languages. By imposing a strictly hierarchical structure of a well-formed (XML-valid) document, the standard greatly simplifies the implementation of parsers for languages based on XML. Any XML parser can parse a document written in any XML-based language. However, XML does not specify how elements of an XML-based language should be interpreted and processed; providing such specifications is the task of the developers of an XML-based representation standard. Similar comments apply to generic representation languages such as the Resource Description Framework (RDF).[44] While some explore the possibility of using RDF to represent, store and query both data and metadata across life sciences datasets (e.g., Cheung et al. [76]), others complain that "putting a bunch of RDF into a bucket is not the same as integration, just like exporting two databases as RDF doesn't necessarily mean you can link their content" (Goble [128]).

In bioinformatics, many standards for the exchange of experimental data have been based on XML.[45] For example, the Microarray Gene Expression Data Society (MGED Society) defined a standard for the storage and transfer of microarray data, Minimum Information About a Microarray Experiment (MIAME; Brazma et al. [57]). A MIAME-compliant record from a microarray experiment must contain, in addition to the raw data, information about the experimental design, array design, preparation of samples, hybridization procedure, scanning process, statistical analysis, etc. In addition, the UML-based MAGE Object Model (MAGE-OM) provides a formal model of the domain, and its XML-based counterpart, the Microarray Gene Expression Markup Language (MAGE-ML; Spellman et al. [343]) provides a means for data exchange. Some microarray gene expression databases, notably ArrayExpress (Parkinson et al. [276]) and GEO (Barrett et al. [30]),

[43] Arguably, XML is not a language, but rather a language template. The human-readability of XML-based documents is also subject to dispute.

[44] http://www.w3.org/RDF/.

[45] For some time, 'BioXML' used to be a trendy buzzword; there were a number of related web sites, e.g., bioxml.org, bioxml.com, bioxml.net, etc., but they are no longer under active development. A more recent bio-buzzword is 'BioRDF' (http://esw.w3.org/topic/BioRDF_Top_Level_Task).

support the MIAME standards. Similar standardization efforts are made in other biomedical domains; see, e.g., Strömback et al. [357] for a recent review.

Unfortunately, while the relative effortlessness with which XML-based languages can be defined greatly enhances the development of representation and exchange standards, the job may easily be overdone — everyone can now create a language for which parsers are already available. The Database Tower of Babel problem may thus be converted, rather than successfully solved, to what one could appropriately call the 'Standards Tower of Babel' problem. Consequently, standardization has not only become a popular topic in bioinformatics, but it has almost developed into a field of its own. New standards-related acronyms, such as MAGE, MO, MIAPE, MISFISHIE, MIRIAM and MIACA,[46] are appearing almost monthly. Most of these initiatives for developing standards are community-based and involve close collaboration of biologists, bioinformaticians and information technologists. Nevertheless, the sheer number of different standards and the pace at which the field is changing is making it difficult, even for professionals, to keep track of all the new developments (Brazma et al. [58]). But while Brazma complains about the emergence of innumerable 'standards',[47] he subscribes to the view that enforcing a top-down, centralized approach to the development of standards is not the right way to go (Quackenbush [289]). Indeed, the Gene Ontology, undeniably one of the most successful terminological standards in molecular biology, has emerged as a *de facto*, bottom-up, community-built rather than a *de jure*, top-down, authority-established standard.

Another line of standardization efforts in computational molecular biology is related to naming and identification[48] of biological entities. The challenge of data integration is grounded not only in incompatible programmatic database interfaces and data exchange formats; it is also the problem of in-

[46] 'Minimal Information on ...', as in MIRIAM = Minimal Information Required In the Annotation of biochemical Models. See http://mibbi.sourceforge.net/, the website of MIBBI, Minimum Information for Biological and Biomedical Investigations.

[47] Some of those specifications, in particular MAGE and MO, were created by Brazma's own team.

[48] Identification in the sense of assigning unique identifiers, not in the sense of recognizing and establishing the identities of observed entities.

compatible and incoherent identification schemes. The problems created by the lack of standards for gene names and their spellings are well known and have haunted the life sciences for more than a decade (Brazma [56]). Two notable examples of attempts to standardize the identification of biological entities are the Life Science Identifier scheme (LSID; Clark et al. [81]) and the Human Gene Nomenclature (Wain et al. [377, 378]) introduced and maintained by the Human Genome Nomenclature Committee (HGNC; Povey et al. [287]). The former effort aims at standardizing the form of identifiers, while the latter aims at controlling the actual collection of terms used to name genes — the HUGO Gene Nomenclature Database (Eyre et al. [109]) provides unique and approved (by HGNC itself) gene names and symbols. Ideally, each gene mentioned in the scientific literature and described by entries in various databases should be referred to by means of the official HUGO name or symbol. In practice, however, alternative names and symbols — aliases — are used more often than the official ones (Tamames and Valencia [363]). Some genes have had their symbols (or names) changed, and their old symbols serve as aliases for other genes; some symbols are aliases to more than one gene. For example, the gene with the HGNC identifier '4122' is assigned the official symbol 'GALNS' and the symbol 'GAS' as an alias, while the latter symbol had previously (until November 20., 2005) been used as the official symbol of the gene with the HGNC identifier '4164' and the (currently used) official symbol 'GAST'. While the use of symbols should alleviate the problems of identification, in practice it seems to produce even more ambiguities. For example, the comparison of microarray data from different sources requires exact mapping of the names used by different authors; this task is greatly complicated by ambiguous symbols, which in different publications identify different genes. The identifier 'PAP' can refer to five different human genes, and in the absence of additional information it is impossible to correctly judge which one is the intended target (Tamames and Valencia [363]). As of July 2007, there are seven human genes that can be referred to with 'PAP' as a symbol; see Table 2.2.

The other mentioned above example of an attempt to standardize identification in molecular biology is the initiative undertaken by the Interoperable

Official symbol	Official name
DDEF1	development and differentiation enhancing factor 1
DDEF2	development and differentiation enhancing factor 2
MRPS30	mitochondrial ribosomal protein S30
PAPOLA	poly(A) polymerase alpha
PDAP1	PDGFA associated protein 1
REG3A	regenerating islet-derived 3 alpha
TUSC2	tumor suppressor candidate 2

Table 2.2: Genes with 'PAP' registered as an alias or previous symbol. Source: HGNC, http://www.genenames.org/.

Informatics Infrastructure Consortium (I3C; no longer on the web):[49]

> "Current bioinformatics applications and databases each have unique formats for the identifiers that they generate and maintain. In order to integrate disparate applications it is necessary for bioinformatics developers to include code to parse and identify the identifiers. This problem is made thornier by the fact that identifiers can often not be recognized simply by looking at the identifier itself out of context, e.g., is GI000197 a GenBank accession number or a GI (GenInfo) number? If we were to write this identifier as an LSID instead:
>
> urn:lsid:genbank.ncbi.nih.gov:genbank.gi:000197
>
> then it becomes immediately recognizable to a program: that is looking at an identifier; what kind of identifier it is; and what authority to contact for resolution, methods queries and so forth." (Clark et al. [81])

Unfortunately, despite the initial enthusiasm and adoption of LSID by the Object Management Group (OMG)[50] and a few other institutions (Martin

[49] See http://xml.coverpages.org/lsid.html.
[50] http://www.omg.org/.

et al. [243]), the LSID identification scheme is not universally used and does not seem to be considered superior to other identification schemes, e.g., employing URIs.[51] However, problems with identification of resources available on the web are by no means specific to the domain of molecular biology. The International World Wide Web Conference[52] regularly features papers focused on what became called the 'Identity Crisis of URIs'.[53] While, for example, Parsia and Schneider [277] and Connolly [83] put forward proposals for how to solve the problem in technological terms, Halpin [156] and Ginsberg [127] call for a more careful philosophical analysis of the issue of identity, reference, and meaning in the context of the semantic web, and Gagnemi and Presutti [118] propose an ontology of web resources and their referencing kinds. LSIDs, as well as other biological identifiers, are all subject to the identity problem; indeed, much of the recent discussions on the mailing list of the W3C Semantic Web Health Care and Life Sciences Interest Group (HCLSIG)[54] have been devoted to questions such as whether a UniProt ID identifies a class of proteins or rather a database entry, etc. See also Good and Wilkinson [135] for more discussion.

Related to, and actually dependent on the efforts to standardize communication protocols (as well as on the development of biomedical and other ontologies) are attempts at data integration based on web services technologies.[55] The myGrid project (Goble et al. [129], Stevens et al. [353])[56] is probably the most prominent example of this approach. With a dedicated tool (Taverna; Hull et al. [175], Kawas et al. [194]), myGrid allows researchers in computational biology to dynamically combine various sources of data and services into so-called 'workflows', using a graphical user interface and without exposing implementational details of the resources used. Unfortunately, despite much effort devoted recently to the development of

[51] Uniform Resource Identifiers; Berners-Lee, http://tools.ietf.org/html/rfc1630/.

[52] WWW, http://www2006.org/.

[53] K.G. Clark's *Identity Crisis*, http://xml.com/pub/a/2002/09/11/deviant.html, is one of the earliest to point out the issue. There is a relevant and interesting discussion between, among others, Tim Berners-Lee and Pat Hayes available from the W3C mailing lists archives (http://lists.w3.org/Archives/Public/www-tag) under the title 'Resources and URIs'.

[54] http://www.w3.org/2001/sw/hcls/.

[55] http://www.w3.org/2002/ws/.

[56] http://www.mygrid.org.uk/.

methods for automated web service discovery and composition (see, e.g., Küngas [217]), myGrid workflows can only be constructed manually. Furthermore, while workflows can be stored and reused by others, there does not seem to be any way of systematically organizing existing workflows, based on a structured description rather than on keyword search.

2.7 Biomedical Ontologies

While the standardization of data representation and exchange formats improves the ability of automated agents to retrieve data of specific types, it hardly makes them understand what the data are about. Without an explicit representation of general knowledge about the domain, the agents have to be hard-coded with procedures for processing the data; this approach is not flexible enough — our knowledge about life is dynamic and changes rapidly,[57] and so do our needs for data representation and retrieval. Unfortunately, large parts of biological knowledge are available only in forms that require substantial processing effort to make them understandable for automated agents. Free-text descriptions are (ideally) informative to humans, but need to be parsed and converted to structured, computer-understandable forms using some sort of natural language processing technology (NLP)[58] — and NLP is considered one of the so-called 'AI-complete' tasks, a collection of problems such that solving one of them is equivalent to solving the entire AI problem: producing a generally intelligent computer program (Shapiro [314]).

As an example, consider the UniProt[59] entry for the protein *gastrin precursor* (name 'GAST_HUMAN', accession 'P01350'). This entry provides some highly structured information, mostly in the form of alternative names and symbols, organism identifiers, and cross-references to other databases that describe various aspects of the protein. However, information about its func-

[57] Undoubtedly, it is the case for what we think we know about complicated interactions between entities at low levels of granularity — at the level of genomes, proteomes, etc.

[58] See, e.g., Jurafsky and Martin [189] for an introduction.

[59] http://www.ebi.uniprot.org/.

tion is available mostly in a non-structured form; while the protein is annotated with the rather general GO terms 'hormone activity' (GO:0005179; molecular function ontology) and 'signal transduction' (GO:0007169, biological process ontology), further details are available only as free-text comments and references to publications reporting original experimental studies of the protein. The corresponding entry in the Online Mendelian Inheritance in Man database (OMIM; Hamosh et al. [157]) consists entirely of a free text description and a list of references.[60] Recently, Belleau et al. [35] have proposed Bio2RDF,[61] an integrative system for biomedical knowledge. It contains, among others, OMIM entries translated into a structured form, encoded in RDF. Figure 2.4 shows an excerpt from the OMIM entry on GAST, and Fig. 2.5 shows a fragment of this information encoded in Bio2RDF.

```
Database: OMIM
Entry: 137250

MIM Entry: 137250
Title:
  *137250 GASTRIN; GAS
Text:
  Gastrin, which is normally formed by mucosal cells in the gastric antrum
  and by the D cells of the pancreatic islets, is a hormone whose main
  function is to stimulate secretion of HCl by the gastric mucosa. HCl, in
  turn, inhibits gastrin formation. Human gastrin has a molecular weight
  of 2,117 and contains 17 amino acid residues. Gastrin I and gastrin II
  ...
```

Figure 2.4: An excerpt from the OMIM entry on the gene GAST. From http://www.ncbi.nlm.nih.gov/omim/.

It is often suggested that manual searching or browsing is no longer a reasonable approach to the exploration of vast resources of biomedical data. Typically, a user or a bioinformatics tool developer is left trying to deal with the following issues: which resources to use; how to use these resources; understanding the content of the resources and interpreting results; transferring data between resources and reconciling values. All these steps are

[60] Interestingly, the OMIM entry still uses the gene symbol 'GAS', even though it has been officially (by the HGNC) replaced with the symbol 'GAST'.

[61] http://www.bio2rdf.org/. Bio2RDF should not be confused with BioRDF.

```
<http://bio2rdf.org/omim:137250>
- <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://bio2rdf.org/omim#GeneticDisorder>
- <http://www.w3.org/2000/01/rdf-schema#label> <"GASTRIN; GAS [omim:137250]">
- <http://purl.org/dc/elements/1.1/identifier> <"omim:137250">
- <http://purl.org/dc/elements/1.1/title> <"GASTRIN; GAS">
- <http://bio2rdf.org/bio2rdf#lsid> <urn:lsid:bio2rdf.org:omim:137250>
- <http://bio2rdf.org/omim#TEXT> "TEXT Gastrin, which is normally formed by mucosal cells
                                 in the gastric antrum and by the D cells of the pancreatic
                                 ..."
  ...
```

Figure 2.5: An excerpt from the Bio2RDF entry on GAST. From http://bio2rdf.org/.

dependent on knowledge of the biological domain, as well as of the more technical specificities of communication protocols etc. It is no longer tenable for an individual biologist to acquire and retain this range and complexity of knowledge; bioinformatics needs computational support for storing, exploring, representing and exploiting knowledge (Stevens et al. [352]). It is claimed that a solution to the challenge of data integration and *interpretation* must involve explicit representation of domain knowledge, in the form of so-called 'ontologies'. There is also a need for systems that can apply the domain expert's knowledge to biological data, systems that can *reason* about what the data represent (Stevens et al. [351]).

What are ontologies? Despite (or perhaps because of) decades of research in ontological engineering, there seems to be no unique, commonly agreed definition of what an ontology is (Guarino [148, 146], Kuśnierczyk [218]).[62] Smith et al. propose the following definition, which attempts to clearly distinguish between the represented and the representation:

> *An ontology*: a representational artifact, comprising a taxonomy as proper part, whose representational units are intended to designate some combination of universals, defined classes, and certain relations between them (Smith et al. [331]).

[62]The problem is also reflected in titles of ontological engineering-related publications such as, e.g., *The Karlsruhe View on Ontologies* (Ehrig et al. [98]).

To fully apprehend this definition, one has to have an understanding of the terms 'taxonomy', 'universal', 'class', 'representation', etc., used in it. Unfortunately, as in the case of 'ontology', these terms have been used in a number of conflicting ways, which often reflect incompatible philosophical views on what (and how) there is in the reality, though sometimes it rather indicates a mere failure (or negligence) to note that there are such different views and that one of them should be consistently adhered to for an ontology to be both technically and philosophically coherent (see Ch. 3). Many other definitions have been proposed. Perhaps most often cited are various versions of the one provided by Gruber [143]: "an ontology is a specification of a conceptualization". For the purpose of this introductory chapter, it shall be enough to adopt the view that an ontology is a logically structured representation of some domain (a portion of reality), which reflects our general knowledge about that domain. The qualifier 'logically structured' refers to the technical side of the representation: it is expressed within a controlled syntax and is given semantics that can be used to perform logical inferences. The qualifier 'general' underlines the scope and purpose of the ontology: to provide a framework for characterizing individual entities, rather than to describe each individual separately.

During the past decade there has been a rapid growth of interest in ontological engineering (OE), the discipline concerned with designing, implementing and deploying ontologies; consult, e.g., Mizoguchi [258], Devedžić [91], and Gómez-Pérez et al. [133]. The first workshop dedicated to OE was held in 1996, in conjunction with the 12th European Conference on Artificial Intelligence (ECAI). Since then, many other conferences have been accompanied by workshops on OE. There are also events entirely dedicated to ontological issues in computational sciences, e.g., the International Conference on Formal Ontology in Information Systems (FOIS).[63] As a result of intense activity in the field, various methodologies for the development of ontologies have been designed; for an introduction, see, e.g., *Ontological Engineering* by Gómez-Pérez et al. [133].

One of the most visible testimonies of the trend is the ontological activity in biomedicine and bioinformatics, perhaps best represented by the Open

[63] http://www.formalontology.org/.

Biomedical Ontologies project (OBO)[64] and its successor OBO Foundry.[65] OBO ontologies can be browsed through the BioPortal of the National Center for Biomedical Ontology (NCBO),[66] or, alternatively, through the Ontology Lookup Service (OLS)[67] provided by the European Bioinformatics Institute (EBI), and at a few other websites. The Gene Ontology (GO), which served as the initial kernel of OBO and successfully continues to be its driving force, is the result of an effort aimed at providing a structured, precise, shared vocabulary for describing roles of genes and gene products in any organism (Ashburner et al. [19], The Gene Ontology Consortium [366, 367]). One of the major motivations for the development of the Gene Ontology was the observation that the terminology used by biologists tends to be ambiguous, which can severely impair the goal of data integration. In his *Integrating biological databases*, Stein comments:

> "A more subtle problem is the clash of concepts as users move from one database to another. An extreme example, first noted by Michael Ashburner, considers the use of the term 'pseudogene' by different researchers and research communities. To some, a pseudogene is a gene-like structure that contains in-frame stop codons or evidence of reverse transcription. To others, the definition of a pseudogene is expanded to include gene structures that contain full open reading frames (ORFs) but are not transcribed. Some members of the *Neisseria gonorrhea* research community, meanwhile, use 'pseudogene' to mean a transposable cassette that is rearranged in the course of antigenic variation. There are also more subtle disagreements. The human genetics community uses the term 'allele' to refer to any genomic variant, including silent nucleotide polymorphisms that lie outside of genes, whereas members of many model-organism communities prefer to reserve the term 'allele' to refer to variants that change genes. Even the concept of the gene itself can mean radically different things to different research communities. Some researchers treat the gene as the transcriptional unit itself, whereas others extend this definition to include up- and downstream regulatory el-

[64] http://obo.sourceforge.net/.
[65] http://obofoundry.org/.
[66] http://www.bioontology.org/.
[67] http://www.ebi.ac.uk/ontology-lookup/.

> ements, and still others use the classical definitions of cistron and genetic complementation." (Stein [347])

Questions such as "What is a gene?", or "What does 'gene' mean?", are not of marginal importance. Recently, it has been argued that the lack of a clear idea of what a gene is may hinder collaboration between researchers, while reaching a consensus over the definition may be impossible (Pearson [279]). The Gene (*sic*) Ontology avoids defining 'gene' altogether. But if the term 'gene' is used to describe database entries and form queries, how can we know — how can a computer know — what those genes are? How can an ontology help? Biomedical ontologies range from structurally simple controlled vocabularies and dictionaries to complex graph-like structures of linked terms, though typically they are not extensively axiomatized.[68] They have, roughly, the following few application scenarios (Stevens et al. [351, 352], Smith [323]):

- Ontologies can be used for defining database schemas. An ontology may provide a high-level view of the data stored in a database, a view that hides the implementational details of the database, an interface between the schema and an application that accesses the data. In the RiboWeb database (Altman et al. [14]), for example, four ontologies were used to specify different aspects of the covered domain: the physical-thing ontology was a specification of the ribosome's molecular components and cofactors, and specified the objects and relations critical for representing data about ribosomal structure; the data ontology specified the types of data that are gathered from biomedical experiments; the data and molecule ontologies interact heavily, because the data ontology's attributes are often constrained to be instances from the physical-thing ontology; the reference ontology specified the publication types; and the methods ontology specified the types of actions RiboWeb could perform.
- Ontologies can be used to provide a translation between the schemas of multiple databases. For example, in the Transparent Access to Mul-

[68] See, e.g., Lassila and McGuinness [232] for an attempt to classify representational artifacts according to their structural complexity.

tiple Bioinformatics Information Sources project (TAMBIS; Goble et al. [130], Stevens et al. [350]), an ontology — the TAMBIS global domain ontology, TaO, an ontology of biomedical terminology (Baker et al. [26]) — was intended as a mediator for accessing multiple biological information sources round the world. The Semantic Meta Database (SEMEDA; Köhler et al. [208]) is another example of this approach, though TAMBIS and SEMEDA differ in logical and implementational details. To my best knowledge, neither TAMBIS nor SEMEDA had ever been used in an actual biological investigation.[69]

– Ontologies can be used as controlled vocabularies for annotation of individual database entries. One of the best known examples of this approach are the so-called 'functional annotations' with terms from the Gene Ontology (Camon et al. [65], Lee et al. [233], Blake et al. [47]). Recently, many other ontologies from the OBO family have been used for such annotations. The Gene Ontology has been extensively used for this purpose, and to date, arguably, most efforts within the GO community have been dedicated to the annotation process (Blake and Bult [48]).

– Ontologies can be used for reasoning over sets of appropriately annotated data. For example, annotations with terms from the Gene Ontology can be used to analyze results from microarray gene expression experiments (Khatri and Draghici [198], Beisvåg et al. [33, 34], Alterovitz et al. [12]).

– Ontologies can be used to annotate and retrieve scientific publications. The Medical Subject Headings terminology (MeSH; Lowe and Barnett [240]) has been one of the most successful structured vocabularies used to annotate biomedical literature.[70] Other examples are given, e.g., by Bodenreider [49], Suomela and Andrade [359], etc.

[69] Confirmed for TAMBIS (private conversation with one of its developers). Its development, however, is continued as part of the ComparaGRID project (http://www.comparagrid.org/). My attempts to contact the authors of SEMEDA have not been successful.

[70] Arguably, MeSH is not an ontology, but rather a hierarchically structured controlled vocabulary with no underlying semantics.

For details on further application scenarios and development tools, see, e.g., Bodenreider et al. [50], Bodenreider and Stevens [53], Cimino and Zhu [78], Lambrix [227], Lambrix et al. [228], Smith [323], and others.

It is commonly recognized that ontologies should be developed based on well-defined, commonly agreed principles, and carefully evaluated before public release (Gómez-Pérez [132]). While there have been reported attempts to build bio-ontologies in a decentralized, non-curated approach (Good et al. [134]), these studies did not provide substantial evidence that ontologies developed in this way are useful for a wider community, and that they have desirable representational and computational properties. It should be noted, though, that folksonomies — subject taxonomies generated by users of various online services in a distributed and uncontrolled manner, based on tags (terms assigned by users to pages, images, etc.) — are often claimed to have been successful in facilitating search through vast amounts of online data at a cost much lower than that needed to develop an ontology in a top-down approach. Folksonomies, however, are structurally much simpler than logically consistent ontologies, and are sometimes called, somewhat pejoratively, 'mob-indexing' (Golder and Huberman [131], Kipp and Campbell [201]).

Ontologies are intended to be a solution to the challenge of data integration (Draghici et al. [94], Khatri et al. [199]), but when developed by separate teams adopting different design principles, they lead to what could be called the 'Ontology Tower of Babel' problem:

> "Ontologies tend to be everywhere. They are viewed as the silver bullet for many applications, such as database integration, peer-to-peer systems, e-commerce, semantic web services, or social networks. However, in open or evolving systems, such as the semantic web, different parties would, in general, adopt different ontologies. Thus, merely using ontologies, like using XML, does not reduce heterogeneity: it just raises heterogeneity problems to a higher level." (Euzenat and Shvaiko [107])[71]

[71] To appear.

During the past decade, much effort has been devoted to the definition and implementation of knowledge representation languages with precise logical semantics, often with an XML-based syntax, such as the Web Ontology Language (OWL; Antoniou and van Harmelen [17]).[72] OWL is available in three versions (sublanguages) with semantics based on a family of extensively studied and computationally well-characterized logical formalisms, Description Logics (DL; Baader et al. [22]).[73] OWL-DL, the most expressive and yet still decidable OWL dialect,[74] is one of the most commonly used knowledge representation languages within the Semantic Web. Its relative simplicity, as well as the availability of a number of tools that facilitate its use, have undoubtedly contributed to the recent explosion in the number of ontologies being developed and deployed — anyone, without much prior experience with knowledge representation, can now build an ontology in minutes (Noy and McGuinness [271]). But while the application of a common ontology language helps to remove syntactical and semantic barriers for communication between agents that use different ontologies,[75] it does not address problems arising from different and often incompatible views on the same reality, reflected in the ontologies.[76]

One way to assure the compatibility between different ontologies (and thus to improve the interoperability of software based on those ontologies) is to provide a unique *top-level ontology* (TLO; an *upper-level ontology*) and force domain ontologies to connect their top-level nodes to appropriate nodes in

[72] http://www.w3.org/2004/OWL/.

[73] For a detailed study of the complexity of various DL formalisms see, e.g., Tobies [370] and Donini [93].

[74] OWL-DL is one of the three sublanguages of OWL 1.0, officially acknowledged by W3C. Recently, a family of slightly more expressive languages has been proposed in the unofficial OWL 1.1 specification; the $\mathcal{SHOIN}$ logic underlying OWL-DL has been replaced with the more expressive logic $\mathcal{SROIQ}$. Several tractable fragments of OWL 1.1 have also been defined. See http://www.webont.org/owl/1.1/.

[75] If agents using different ontologies are to communicate successfully, their ontologies should be aligned first; see, e.g., Sampson [305] and Euzenat and Shvaiko [107] for a comprehensive review of ontology matching techniques and tools. Lambrix and Tan [229] describe an approach targetted specifically at aligning and merging biomedical ontologies.

[76] Focusing exclusively on the syntactic and model-theoretic correctness of ontologies may lead to what is sometimes called the 'nonsense in nonsense out' problem (Spear [342]): ontologies may be well-formed and logically valid, but still inaccurate content-wise. Unfortunately, many believe that for an ontology to be a good one it suffices to encode it in a logical formalism (e.g., a description logic) and no inconsistencies are found when the ontology is checked with a reasoner.

the top-level ontology. Unfortunately, it seems highly unlikely that a single, non-trivial top-level ontology would be agreed by all ontology developers, and the project of building such an ontology has largely been abandoned (Smith [318]). Nevertheless, a number of competitive TLOs have been built and adopted for use by distinct communities. Some of the most prominent examples of those are:

- the Basic Formal Ontology[77] (BFO; Grenon [136, 140], Smith and Grenon [328], Spear [342]; see also Ch. 3);
- the Descriptive Ontology for Linguistic and Cognitive Engineering[78] (DOLCE; Gangemi et al. [123], Masolo et al. [244]);
- the Sowa's top-level ontology[79] (Sowa [340]);
- the Cyc top-level ontology,[80] and others.

Attempts have been made to combine several top-level ontologies into one structure, as in the case of the Multi-Source Ontology (MSO);[81] however, the ontologies differ not only in terms of the formalisms used, but also in terms of the underlying philosophical doctrines, and mapping between them is neither obvious nor complete (see, e.g., Masolo et al. [244] for an attempt to provide such a mapping).[82]

In the case of OBO, one of the most recent and significant attempts at standardizing the development of its member ontologies is the adoption of BFO as the top-level ontology. Substantial efforts are being dedicated to systematize OBO ontologies in this and a number of other aspects. The goal of Ch. 3 is, among others, to present this effort in more detail, with focus on BFO as the common core.

[77] http://www.ifomis.uni-saarland.de/bfo/.
[78] http://www.loa-cnr.it/DOLCE.html.
[79] http://www.jfsowa.com/ontology/toplevel.htm.
[80] http://www.cyc.com/cycdoc/vocab/top-vocab.html.
[81] http://www.webkb.org/doc/MSO.html.
[82] Masolo compare DOLCE with BFO and OCHRE using a trivial example, and the example is already biased towards a particular theory of existence.

Chapter 3

A Standardization Effort in Biomedical Ontology

This chapter discusses the issue of incoherent terminology in literature on Ontological Engineering (OE). We provide evidence that there is a load of confusion in what ontological engineers say. The terminologies they use are often ambiguous and inconsistent, and these nomenclatural issues reflect either carelessness of expression, underspecification of the technical meaning of terms, or an unfortunate lack of a common, coherent philosophical foundation. Terminological issues lead to problems when ambiguous plain language characterizations are to be turned into formal definitions, as discussed by, e.g., Guarino and Giaretta [148] and Guarino [146].

The chapter is structured as follows:

- Section 3.1 provides a brief introduction into the problem.
- Section 3.2 discusses a number of definitions or characterizations of some of the terms most commonly used in Ontological Engineering.
- Section 3.3 describes a framework for ontological development that is

currently being adopted by the Open Biomedical Ontologies (OBO)[1] community.

3.1 Introduction

Ontology, a branch of philosophy occupied with the study of being, has had a long history. Questions such as *What is there?* and *What is existence?*, were disputed by ancient philosophers even before Plato and Aristotle laid the groundworks of ontology in its modern shape. On the other hand, ontologies, artifacts for expressing and exchanging knowledge about selected portions of reality with the rigor of a formal and computer-understandable language, are a relatively recent invention (Smith [318]). There is currently an explosion of efforts in development, publishing, merging and applying ontologies. This trend has been fuelled, among others, by the rapid increase, both in number and size, of publicly available online sources of data, information and knowledge; these sources employ diverse underlying structures and incompatible languages (the so-called 'Database Tower of Babel' problem, Sec. 2.4) that call for a principled approach to integration at the semantic level. This explosion, in turn, has created a broad niche for further research on philosophical ontology and ontological engineering (Guarino and Musen [149]).

Ontologies flourish in just about every imaginable corner of our scientific and non-scientific activity. This is especially visible in the natural sciences, e.g., under the umbrella of Open Biomedical Ontologies, where there is a strong need for semantic integration of otherwise highly distributed data. But while one of the principal goals behind the effort of constructing ontologies is to enable both a human user and an automated reasoner to access and comprehend data from various sources without being forced to investigate their terminological and implementational details, the cure seems often to be no better than the disease: the ontologies themselves are based on incompatible philosophical doctrines, represented in different languages and stored in custom-made databases whose schemas do not match with each

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

other. Consequently, the promised benefit may easily become outbalanced by the burden of having to parse, interpret and match multiple heterogeneous ontologies (what can be called the 'Ontology Tower of Babel' problem, Sec. 2.7). Despite a common syntactic commitment, OBO ontologies, for example, employ quite different methods and techniques in the modeling of their domains. (There are, however, intense ongoing efforts to standardize OBO ontologies also at the level of their ontological commitments.)

3.2 [N]ontological Engineering

Ideally, team of experts who set off to create an ontology of a particular domain should consist, at the very least, of domain experts, knowledge modelers (usually computer scientists), and those with expertise in philosophical ontology. Yet given that formal ontology[2] is just one of many branches of philosophy, and that the subject-matter of virtually any science and industry may be the object of ontological engineering, philosophical ontologists will likely be greatly outnumbered by ontological engineers. The scale and diversity of the attempts to fill up the ontology niche emphasizes the need for a sound, understandable and reusable basis for this discipline. An attempt to systematize the foundations should be based on consent rather than on competition, and must be thorough. An account of these foundations should, to best serve the ontology engineering world, be easily understandable, but by no means oversimplified or confused; it should allow for a shared understanding of how different ontological engineering paradigms may be used to model the same reality.

Two recently published books seem to be intended as a reference for those who seek such an account. One of them, the *Handbook on Ontologies* (Staab

[2]The phrase 'formal ontology' can be interpreted in at least two ways. It may denote the discipline that studies the most general *forms* of existence; it is formal because of what it aims to describe. On the other hand, it may also denote the activity aimed at producing structured, logically axiomatized descriptions of the reality; it is formal because of how it aims to describe. In this chapter, 'formal ontology' is used to capture both characteristics. The products of formal ontology are often called 'foundational ontologies' (Masolo et al. [244]). See, e.g., Smith [318] for further details.

and Studer [344]), "demonstrates standards that have been created recently; it surveys methods that have been developed and it shows how to bring both into practice of ontology infrastructures and applications that are the best of their kind". The other, *Ontological Engineering* (Gómez-Pérez et al. [133]), "presents the major issues of ontological engineering and describes the most outstanding ontologies currently available". Unfortunately, a closer look at the books reveals a number of problems, among others redundant and incoherent introduction of terms, which certainly interfere with a reader's understanding. Both books were written, by and large, by computer scientists for computer scientists to provide guidance through the world of ontological engineering; yet the result of the authors' effort may be no less confusing than what they attempt to disambiguate.

After having presented a number of conflicting definitions of what an *ontology* is (see below), the authors of *Ontological Engineering* conclude: "We can say that as there is consensus among the ontology community, no one can get confused about its usage [of the term 'ontology']." But the actual situation seems to be quite the opposite: there is much confusion as to what an ontology (a representational artifact) is, what it is that its components represent, and what the principles for building an ontology are. There has been much debate between the leading ontologists and ontological engineers[3] on what ontologies — and more specifically, natural, semi-formal, and formal ontologies — are, how (and whether) they differ from taxonomies, conceptual models, controlled vocabularies, thesauri, etc.

In his discussion of principles for the design of ontologies, Tom Gruber says that ontological commitment is based on consistent use of vocabulary (Gruber [143]). Ontologies are built to provide consistent vocabularies for different domains. But is there any commitment to a consistent vocabulary in the community of ontological engineers? In the following, we discuss evidence that the answer to this question is negative. *Quis custodiet ipsos custodes?*[4]

[3] See, e.g., archives of the Ontolog Forum, http://ontolog.cim3.net/ .

[4] "Who will guard the guards?", attributed to Decimus Iunius Iuvenalis (Juvenal), I–II AD.

3.2.1 Knowledge and Models

In philosophy, ontology is a systematic account of being (Hofweber [170]). It is the *existence in reality* which is the subject of study and description. What there exists, exactly, has always been a matter of fierce philosophical debates. Philosophers characteristically charge each other with improperly identifying what there is, and in the history of philosophy every kind of entities will at one time or another have been thought to be a fictitious result of an ontological mistake (Blackburn [46]). Nevertheless, the term 'ontology' seems to have been granted a commonly agreed meaning here.

In computer science, however, there is hardly any agreement as to what 'ontology' should mean. While an ontology is usually considered a representational artifact, it is not entirely clear what the necessary and sufficient conditions are for a representational artifact to be called an 'ontology' — *what* it is that an ontology must represent, and *how* it must represented. The meanings assigned to the term 'ontology' are often conflicting. In his *Towards Principles for the Design of Ontologies Used for Knowledge Sharing*, Gruber proposed the following definitions:

> "A specification of a representational vocabulary for a shared domain of discourse — definitions of classes, relations, functions, and other objects — is called an ontology. ... An ontology is an explicit specification of a conceptualization." (Gruber [144])

The second part has been one of the most often cited definitions, sometimes with 'specification' replaced by 'formal specification', and 'conceptualization' replaced by 'shared conceptualization'. However, it is not clear what it is that an ontology specifies: a conceptualization of the domain, the vocabulary used to represent the domain, or, perhaps, the vocabulary used to represent the conceptualization — all of which are usually not the same.[5] Furthermore, the term 'conceptualization' itself is unclear; as argued by Guarino and Giaretta [148], Gruber's use of this term is grounded in its incorrect (according to those authors) extensional interpretation by Genesereth and

[5] Except for some rather pathological (in this context) cases of homoiconicity or metacircularity.

Nilsson [124].

In his *Formal Ontology and Information Systems*, Guarino provides another definition:

> "An ontology is a set of logical axioms designed to account for the intended meaning of a vocabulary." (Guarino [147])

In this sense, an ontology is a theory of a domain — a portion of reality — expressed formally, i.e., using a logical language equipped with formal semantics. The axioms of the theory are built of terms that are used to describe the domain, and thus the theory specifies formally (by way of the semantics of the language) the meaning of those terms. This view is on par with those of some other authors:

> "Ontologies are quintessentially content theories, because their main contribution is to identify specific classes of objects and relations that exist in some domain." (Chandrasekaran et al. [74])
>
> "An ontology is a hierarchically structured set of terms for describing a domain that can be used as a skeletal foundation for a knowledge base." (Swartout et al. [360])

Gruber's ontologies are specifications of conceptualizations. However, Guarino's and Chandrasekaran's ontologies *are* conceptualizations; a conceptualization is an abstract model of some aspect of the world, taking the form of a definition of the properties of important concepts and relationships (Baader et al. [22]). Other authors seem to propose that ontologies are representations of one's imagination — a cognitive representation, conceptualization — of the domain, thereby increasing the distance of indirect representation between an ontology and the actual domain. Other authors suggest that

> "Ontologies can be used to provide a concrete specification of term names and term meanings. ... An ontology is a specification of the conceptualization of a term." (Lassila and McGuinness [232])
>
> "Ontology languages allow users to write explicit, formal conceptualizations of domain models." (Antoniou et al. [17])

> "Ontologies are explicit representations of agents' commitments to a model of the relevant world. ... Ontologies are specific, high-level models of knowledge underlying [*sic*] all things, concepts, and phenomena.[6] ... Generally, an ontology is a metamodel describing how to build models." (Devedžić [91])

> "Ontologies are agreements about shared conceptualizations. Shared conceptualizations include conceptual frameworks for modeling domain knowledge, ... and agreements about the representation of particular domain theories." (Uschold and Gruninger [374])

Many other definitions, more or less coherent with some of the above and conflicting with others, have been given. The issue is not merely nomenclatural; some biomedical terminological systems — e.g., the Medical Subject Headings (MeSH),[7] or the Unified Medical Language System (UMLS)[8] and its relatives — are sometimes used as if they were ontologies,[9] although it has been shown that logical reasoning with such systems is problematic and may lead to erroneous conclusions (Burgun and Bodenreider [63], Schulze-Kremer et al. [312]). Ontologies are commonly built for the purpose of drawing inferences about the part of reality they are intended to represent. Any misuse of the term 'ontology' may lead to confusion with potentially disastrous consequences — particularly in the biomedical domain.

Some of the definitions above employ the term 'term', as in "an ontology is a specification of the conceptualization of a term" (Lassila and McGuinness [232]). There is, however, a substantial difference between a term and what the term represents, between the term as an element of a statement and the term as an object of the proposition expressed by that statement; unfortunately, this difference is blurred in definitions such as the above. The distinction has been christened the 'use-mention' distinction (Kenyon [196], Spear [342]).

[6] Perhaps the best illustration to the claim that knowledge about reality also underlies that reality is the graphic *Drawing Hands* by M. C. Escher (1948).

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

[8] http://umlsinfo.nlm.nih.gov/ .

[9] Librelotto et al. [235], for example, use the 'MeSH ontology' to build an 'ontological index' for PubMed.

Three issues are in focus in the above characterizations of an ontology as a representational artifact: *what* the domain of an ontology is; *how* an ontology represents that domain; and *why* the ontology has been built — the purpose of the representation. Although the issue of representational formalisms has been researched by the knowledge representation community for decades, there is little consensus on what structure is essential for a representational artifact to be called an 'ontology'. See, e.g., Lassila and McGuinness [232] or Studer et al. [358] for more discussion. As to the 'what?' question, it has only recently been pointed out that often it is not sufficiently clear what the subject matter of an ontology is: specific individuals, or rather general patterns in a domain; our knowledge about them — or the *lack* of such knowledge — or even terms in that very same ontology (Bodenreider et al. [51], Smith [319]). The few example definitions cited above clearly illustrate this uncertainty. Is an ontology an abstract model of a domain, a conceptualization of such a model, or a specification of such a conceptualization? What is the domain represented by a specification of a conceptualization of a model?

In the context of biomedicine — specifically, within OBO Foundry community — we propose to focus on what it is that an ontology represents, irrespectively of the expressivity of the underlying representation language or the actual structural complexity of the artifact (Smith et al. [331], Kuśnierczyk [218], Schober et al. [311]). The issue may seem, superficially, to be of purely terminological nature. However, as it is argued in, e.g., Smith and colleagues' *Wüsteria* [327], dire consequences follow if it is not made perfectly clear what the represented domains are: whether a term in an ontology represents an entity in reality, a belief about an entity, an act of observation of an entity, a documentation of an observation or belief, a belief about an observation, etc. Where communities of ontology developers do not share a single coherent view on these matters, the result is a confusion. But even if the problem were purely terminological, one motivation for building and using ontologies is to standardize vocabularies — why should we not speak in standard terms about these very standardization efforts themselves?

3.2.2 Concepts and Classes

The problem of imprecise definitions of the term 'ontology' is accompanied by the equally untidy use of terms such as 'concept', 'class', 'type', 'kind', 'instance', 'property', 'relation', 'role', etc. to refer both to the elements of an ontology, and to the corresponding elements of the domain, which are represented by the former. These terms are commonly used in the literature on ontological engineering. What do they mean? What are they *supposed* to mean?

In some cases, typically in the tradition of description logics (DL; Nardi and Brachman [262]), the term 'concept' is used to speak of those elements of an ontology that represent what is called 'classes' in the reality. Likewise, in conceptual graphs concepts are representational elements:

> "A class is a set of entities." (Chaudhri et al. [75])
>
> "Concepts are terminological descriptions of classes of individuals." (Welty [381])
>
> "In a conceptual graph, the boxes are called *concepts*, and the circles are called *conceptual relations*." (Sowa [340, p. 476])

Others, however, use these terms in quite a different way:

> "Classes represent concepts, which are taken in a broad sense." (Gómez-Pérez et al. [133])
>
> "A class has an intensional meaning (the underlying concept) which is related but not equal to its class extension." (Bechhofer et al. [32])
>
> "The most basic concepts in a domain should correspond to classes that are the roots of various taxonomic trees." (Smith et al. [336])

In many cases, the terminology is even less clear:

> "A concept is a meaning. There are major groupings of semantic

> types for organisms, anatomical structures, biologic function, chemicals, events, physical objects, and concepts or ideas."[10]
>
> "[Concepts] can be concrete (like a patient) or abstract (a prototypic patient)." (Plaza and Arcos [284])
>
> "Just as in the object-oriented paradigm, there are two fundamental types of concepts in KM: instances (individuals) and classes (types of individuals)." (Clark and Porter [80])

The terms 'instance' and 'individual' enjoy a similarly casual treatment:

> "Individuals are assertional, and are considered instances of concepts." (Welty [381])
>
> "Each of the entities in a class is said to be an instance of the class." (Chaudhri et al. [75])
>
> "Instances are used to represent elements or individuals in an ontology. ... Individuals represent instances of classes. ... Individuals represent instances of concepts." (Gómez-Pérez et al. [133])

Both in actual ontologies and in their documentation, there can be found statements such as "living subject is a code system", "the term 'house' is not a month in the real world, though 'January' is", or "normal cell is a subclass of microanatomy". Behind this terminological diversity there also hides the problem of what it is that classes (or concepts) are instantiated by, and whether there are (or can be) classes (concepts) whose instances are not individuals but rather other classes (concepts), and what the criteria are for deciding whether to represent an entity as an individual or as a class (concept).

Some representation languages — including Semantic Web languages such as RDF/RDFS[11] and OWL[12] in its unconstrained version (OWL-Full) — allow classes to have classes (including themselves) as instances:

[10] UMLS Semantic Network, http://www.nlm.nih.gov/research/umls/meta3.html.

[11] The Resource Description Framework (RDF) and the RDF Vocabulary Description Language (RDF Schema, RDFS), http://www.w3.org/RDF/.

[12] The Web Ontology Language, http://www.w3.org/TR/owl-ref/.

> "A class can be an instance of a class." (Chaudhri et al. [75])
>
> "The class `rdfs:Class` defines the class of all classes." (Gómez-Pérez et al. [133])

This freedom of expression comes at a high cost: the languages are intractable, and some serious foundational philosophical questions arise (e.g., the Russell's paradox; Irvine [181]). The so-called 'intensional semantics' of KIF are an example of an attempt to escape such paradoxes (Hayes and Menzel [162]). But some authors suggest that higher-order instantiation (classes as instances of classes) is indeed necessary:

> "The canonical example ... is species/animal. While most introductory courses teach the difference between classes, such as *Mammal* or *Human*, and instances, such as *Chris*, they stop short of explaining how second-order classes, such as *Species*, would fit into the picture. *Human* is a subclass of *Mammal*, and *Chris* is a *Human* and therefore a *Mammal*. Is *Human* also a subclass of *Species*? ... In fact, *Human* turns out to be an instance of *Species*." (Guarino and Welty [150], original emphasis)

(We return to this example in Appendix B to show that there can be a reasonable solution given which does not require admitting higher-order classes. Such solutions are essential, at least if a tractable language, e.g., OWL-DL, is to be used.)

3.2.3 Classes and Individuals

What are the criteria for considering, and correspondingly representing, an entity as a class or as an individual? In other words, how to decide whether an entity should be represented with a class-representing element or with an individual-representing element? Some authors suggest that the choice depends on the particular *application* of the ontology:

> "Deciding whether a particular concept is a class in an ontology or an individual instance depends on what the potential applications

> of the ontology are. Deciding where classes end and individual instances begin starts with deciding what is the lowest level of granularity in the representation. The level of granularity is in turn determined by a potential application of the ontology. ... Individual instances are the most specific concepts represented in a knowledge base. ... If concepts form a natural hierarchy, then we should represent them as classes." (Noy and McGuinness, *Ontology Development 101: A Guide to Creating Your First Ontology*[13])
>
> "In certain contexts something that is obviously a class can itself be considered an instance of something else. For example, in the wine ontology we have the notion of a `Grape`, which is intended to denote the set of all *grape varietals*. `CabernetSauvingonGrape` is an example instance of this class, as it denotes the actual grape varietal called Cabernet Sauvignon. However, `CabernetSauvignonGrape` could itself be considered a class, the set of all actual Cabernet Sauvignon grapes.
>
> It is very easy to confuse the instance-of relationship with the subclass relationship. For example, it may seem arbitrary to choose to make `CabernetSauvignonGrape` an individual that is an instance of `Grape`, as opposed to a subclass of `Grape`. This is not an arbitrary decision. The `Grape` class denotes the set of all *grape varietals*, and therefore any subclass of `Grape` should denote a subset of these varietals. Thus, `CabernetSauvignonGrape` should be considered an *instance* of `Grape`, and not a subclass. It does not describe a subset of Grape varietals, it *is* a grape varietal." (Smith et al., *OWL Web Ontology Language Guide*[14])

Here, apparently, representational elements are conflated with what they represent (compare: "the `Grape` class denotes", "`CabernetSauvignonGrape` ... is an instance of `Grape`", and "`CabernetSauvignonGrape` does not describe a subset of Grape varietals, it *is* a grape varietal"). Furthermore, the naming convention is rather confusing: while `CabernetSauvignonGrape` is the class (called 'Cabernet Sauvignon') of all actual grapes — individual fruits — of the varietal Cabernet Sauvignon, `Grape` is not the class of all

[13] http://protege.stanford.edu/publications/ontology_development/ontology101.pdf.
[14] http://www.w3.org/TR/owl-guide/.

actual grapes of any varietal, but rather the class of varietals. (The examples given by Smith et al. come from the Wine Ontology,[15] available online.) The *Ontology Development 101* provides a similar discussion:

> "Consider the wine regions. Initially, we may define main wine regions, such as France, United States, Germany, and so on, as classes and specific wine regions within these large regions as instances. For example, `Bourgogne region` is an instance of the `French region` class. However, we would also like to say that the `Cotes d'Or region` is a `Bourgogne region`. Therefore, `Bourgogne region` must be a class (in order to have subclasses or instances). However, making `Bourgogne region` a class and `Cotes d'Or region` an instance of `Bourgogne region` seems arbitrary: it is very hard to clearly distinguish which regions are classes and which are instances. Therefore, we define all wine regions as classes." (Noy and McGuinness, *Ontology Development 101: A Guide to Creating Your First Ontology*)

One problem here, again, is that the representation is conflated with the represented. We can, of course, represent a wine region with either a class (if 'class' is used to speak of representational elements) or an instance (if, likewise, 'instance' is taken to be a representational rather than ontological term); the choice may be application dependent, or may be enforced by the underlying formalism (OWL in this case).[16] However, a wine region — a *particular* wine region, such as the Bourgogne region — is not, and cannot possibly be, a class — irrespectively of the corresponding representational element used in some ontology. No French region, including the Bourgogne region, is a class, and there is no arbitrariness in making it such — a region just can't be made a class, irrespectively of the intended application of an ontology. With some dose of uncritical creativity on the side of the ontological engineer, a French region can be represented as a class. But whether this makes sense should be carefully judged before publishing the ontology.[17] (Furthermore, note that Côte-d'Or is, administratively, *not* a region — it is a

[15] http://www.w3.org/TR/2004/REC-owl-guide-20040210/wine.rdf.

[16] None of which really justifies the ontologically inaccurate use of a class-representing element to represent a non-class entity.

[17] These and other problems inherent in the examples above were discussed at FOIS'2006 [218]. As a result, Chris Welty expressed the will to re-edit the OWL Guide (private conversation).

department, and as such, it is a *part* of the Bourgogne region. Côte-d'Or is certainly an instance — it is a *particular* department, an instance of, among others, the class of all French departments; but it is *not* an instance of Bourgogne — or of what Noy and McGuinness would wish to call the 'Bourgogne region-class'. If there is a class of Bourgogne regions, the Bourgogne region in France is surely its unique instance.)

Interestingly, this example also provides an illustration of a related, often misunderstood issue — the difference between levels of specificity and levels of granularity. France (an individual, an instance of, e.g., the class of all countries) is divided into non-overlapping regions; each of these regions is, in turn, divided into departments, and the departments into arrondisements. Both the regions and the departments are parts of France; at different levels of granularity, France is divided into 26 regions, 100 departments, 342 arrondissements, etc. Each of these administrative units is an instance of some class. For example, 22 of 26 French regions are metropolitan regions, 21 of those are mainland regions, etc. These three classes of regions — that of French regions, that of metropolitan French regions, and that of mainland French regions — are defined at different levels of *specificity*, but at the same level of *granularity* (they are all classes of first-order administrative divisions of the territory of France). On the other hand, the class of French regions and the class of French departments are defined at the same level of specificity, but at different levels of granularity (they are classes of first-order and second-order divisions, respectively). For more on the distinction between specificity and granularity, see, e.g., Spear [342].

3.2.4 Further Notes

There seems to be no coherent interpretation of the most commonly used terms in Ontological Engineering; unfortunately, the issue is not merely terminological, and the underlying philosophical disagreements render the problem difficult to solve generally. Similar complaints have been raised earlier, e.g., about incoherent views on what an ontology is (Guarino [146]), about misuses of the term 'concept' (Smith [319, 320, 322]), etc. Neverthe-

less, this 'state of the art' persists; the terminology proposed by Guarino and Giaretta [148], in which they suggest to replace the ambiguous 'ontology' with more the precise (according to those authors) 'conceptualization' and 'ontological theory', has not been widely adopted. On the one hand, it is certainly a property of natural language that the meaning of expressions is context-dependent — it is specified not only by the semantics, but also by the pragmatics (Korta and Perry [212], Leech and Weisser [234], Stevenson and Wilks [354]). On the other hand, this is usually undesirable in automated knowledge-based systems, whether in the case of biomedical ontologies (Tuason et al. [371]) or in the case of the Semantic Web in general (Booth [54], Ginsberg [127], Parsia and Patel-Schneider [277]).

While global disambiguation of terms may be unachievable,[18] it may well be worth heading for in well-defined communities, such as the community of Open Biomedical Ontologies. In the next section I present an initiative which I had the opportunity to participate in, undertaken within the community of biomedical ontologists and aimed at resolving the problems of ambiguity by delineating a coherent philosophical and terminological framework for the development of biomedical ontologies. These the efforts are supported by organizations such as the National Center for Biomedical Ontology (NCBO),[19] the Open Biomedical Ontologies (OBO) Consortium, the OBO Foundry, and others (see, e.g., Rosse and Mejino [302]).

3.3 A Philosophical Framework for the Integration of Biomedical Ontologies[20]

As argued in Sec. 3.2 (and also in, e.g., Smith et al. [319] and Kuśnierczyk [218]), the term 'concept' is somewhat problematic in ontological engineering. But also other terms, such as 'class', 'object', 'instance', 'individ-

[18] See, e.g., *In Defense of Ambiguity* (Hayes [161]) for an interesting discussion of this issue in the context of the Semantic Web and the URI crisis of identity.

[19] http://www.bioontology.org/.

[20] This section is in large parts based directly on the content of Smith et al. [331], without further discussion.

ual', 'property', 'relation', etc., have been non-uniformly adopted by ontology developers, and have been used with multiple, often conflicting meanings. What is needed, then, is a set of terms referring unambiguously to the different kinds of entities in reality; such a set of terms should serve as common target for mappings between various knowledge representation systems, thereby mediating translations between ontologies built in those systems. It has been found essential that various biomedical ontologies, to be fully interoperable, must follow the same philosophical, representational, and terminological conventions (see, e.g., Spackman and Reynoso [341], Johansson [184], Klein and Smith[21]). In Schober et al. [311], we provide details of a recently introduced collection of terminological, nomenclatural, and typographical conventions based on the results of our research on ontology development practices in the MSI,[22] PSI,[23] FuGE,[24] and other related communities, and suggested for adoption by OBO.

The framework introduced in further text corresponds to and complements the Basic Formal Ontology (BFO),[25] a realist theory of existence,[26] adopted recently by the OBO community as their standard top-level ontology. See, e.g., Spear [342] and Grenon [137, 138, 136] for more details and a formal axiomatization of BFO, and Grenon et al. [140] for more on its application in biomedical ontology. See also Masolo et al. [244] for a comparison of an early version of BFO with, among others, the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE).[27] BFO is under continuous development; a number of interesting discussions related to its adoption by the biomedical ontology community can be found in the archives of the bfo-discuss mailing list.[28]

[21] *Concept Systems and Ontologies*, http://ontology.buffalo.edu/concepts/.

[22] The Metabolomics Standards Initiative, http://msi-workgroups.sourceforge.net/, a committee appointed by the Metabolomics Society.

[23] The Proteomics Standards Initiative, http://www.psidev.info/, a committee appointed by the Human Proteome Organization (HUPO; [190]).

[24] The Functional Genomics Experiment Community, http://fuge.sourceforge.net/.

[25] http://www.ifomis.uni-saarland.de/bfo/.

[26] Very roughly, *realists* deny dependence of the existence of entities in reality on human cognition. See, e.g., Miller [255], Boyd [55], or Khlentzos [200].

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

[28] http://groups.google.com/group/bfo-discuss/.

3.3.1 Reality and Representation

One of the most important distinctions we insist on in what follows (and in the context of OBO ontologies in general) is the one based on the relation of *representation* (or *reference*) that holds between two entities, the represented (the *referent*) and the representing (the *reference*).[29] An entity represents (or refers to) another entity just in case there is an agent, whether alive or artificial, who recognizes the former as standing in some respect for the latter. This informal definition goes in line with that given by Peirce, and also reflected in the so-called 'semiotic triangle' (or 'triangle of meaning') of Ogden and Richard's [272]:

> "A sign, or *representamen*, is something which stands to somebody for something in some respect or capacity. It addresses somebody, that is, it creates in the mind of that person an equivalent sign, or perhaps a more developed sign. That sign which it creates I call the *interpretant* of the first sign. The sign stands for something, its *object*, not in all respects, but is reference to some sort of idea, which I have sometimes called the *ground* of the representation." (Peirce, as quoted in *Knowledge Representation*, Sowa [340, p. 192])

The relation of representation that holds between the representing and the represented is thus dependent on the observer, rather than fixed for the two relata.[30] One of the goals of ontology development is to provide representations that are understood by all users in a like manner. Note that Peirce's definition may be seen as suggestive of that it is not the object, but rather the mental representation (the interpretant) which is represented by the sign. This view is further discussed — and argued against — in Sec. 3.3.3. (The issue of the nature of mental representations has been of

[29] Representation and reference are not necessarily the same; for the sake of simplicity we proceed here as if they were. See, e.g., Reimer [295].

[30] It is thus a ternary rather than a binary relation: it holds between the sign, the object, and the agent who interprets the sign as representing the object. If the mental sign (interpretant, concept) is to be acknowledged, representation should be seen as a quaternary relation. The role of the observer is explicitly acknowledged in an extension to the semiotic triangle — the semiotic tetrahedron of Falkenberg et al. [110].

interest to cognitive science and artificial intelligence for over half a century (Thagard [364]). For an overview of theories of cognitive representations see, e.g., Aydede [21], Bermúdez [41], Horst [173], Pitt [283], and Thomas [369].)

In principle, there can be self-referential entities — entities which represent themselves. Such entities are not uncommon in computer science: the Scheme [9] statement

```
(define symbol 'symbol)
```

and the IKL[31] sentence

```
(= symbol 'symbol')
```

are typical examples of definitions or assertions about self-referential entities that can be written in a programming or a knowledge representation language. Again, in biomedical ontology self-referential representations are of rather little importance; but see, e.g., Smith and Ceusters [325] for examples of how self-referential representations have actually been included — confusedly — in an international standard for healthcare records.[32]

3.3.2 Three Levels of Reality

Grounded in the above, we distinguish three levels of entities which have a role to play wherever biomedical ontologies are used:

Level 1. The entities — physical objects, the processes they participate in, their qualities, states, etc. — existing independently of the cognitive activities of an observer (for example, on the side of a patient examined by a clinician or a researcher).

[31] http://www.ihmc.us:16080/users/phayes/IKL/GUIDE/GUIDE.html.
[32] Health Level 7 (HL7), http://www.hl7.org/.

Level 2. The cognitive representations of those entities made by an observer (for example, on the part of a clinician or a researcher examining a patient).

Level 3. The textual, graphical, and other representations of the objective reality, reflecting (but *not* representing) the cognitive representations.

Level 1 is indispensable even in situations where the representations are 'data models' rather than models of patients and their diseases — here, real data gathered by the clinician are the objects of the representation. Level 2 reflects the fact that a crucial role is played in ontology and terminology development by the cognitive representations made within the minds of human subjects. Level 3 reflects the fact that cognitive representations can be shared, and serve scientific ends, only when they are made communicable in a form whereby they can also be subjected to criticism and correction, and also to implementation in software. Level 3 representations represent the same portion of reality as Level 2 representations — namely, they represent Level 1; the first are not representations of the second.[33]

Note that the distinction into three levels is relative to the particular goals of an ontology development project. The textual and graphical artifacts, for example, distinguished in Level 3 are, of course, themselves objects on Level 1 in situations where such artefactual representations should themselves be represented — as in the so-called 'knowledge representation' ontologies such as Gruber's Frame Ontology (Gruber [144]), the Open Knowledge Base Connectivity framework (OKBC; Chaudhri et al. [75]), etc. Note also that while Level 1 might be called the 'objective' reality, and Level 2 might be called the 'subjective' reality, this would not always be accurate. There may be, on the side of the patient, perceptual experiences, feelings, delusions, hallucinations, etc., which are, of course, subjective, and as such cannot be observed by the physician, but may nevertheless need to be represented in specific cases. The physician may actually want to (artefactually)

[33] In cases where one does intend to represent cognitive representations, these representations are themselves Level 1 entities, but then there are also cognitive representations of those representations, which are Level 2 entities and are not represented at Level 3.

represent his own subjective experiences in relation with the patient, which does not make these experiences objective, even if treated as Level 1 entities. However, most biomedical ontologies deal with objective reality, not with subjective experiences.[34]

3.3.3 (Against) The Concept Orientation

According to the realist approach to ontological engineering,[35] ontologies should consist exclusively of representational units which are intended to designate entities in the target domain, and not conceptualizations thereof. Defenders of the concept orientation in medical terminology development (whom we shall refer to as 'terminological conceptualists') have offered a series of arguments against this view, to the effect that such terminologies should include also (or even exclusively) representational units referring to what are called 'concepts' (Cimino [77]). These arguments can, roughly, be classified as follows.

1. The argument from *intellectual modesty*: domain experts disagree, and thus a terminology should embrace no claims as to what the world is like, but rather make efforts to represent the *concepts* that different experts have.

2. The argument from *non-existence*: there is a need for terms that would refer to what does not exist, or what does not exist yet. For example, a patient may hallucinate that she hears voices when there actually are none; drug designers may talk about substances that haven't been synthesized yet. Terms corresponding to such non-existent entities are taken to represent concepts.

3. The argument from *history of medicine*: some terms are historically grounded in beliefs, which were only later proved to be false. Such

[34] Subjective experiences are the subject of study of phenomenology (Smith [335]). Recently, I have participated (together with Arild Faxvaag and Barry Smith) in an effort to develop an ontology of subjective experiences.

[35] See, e.g., Smith et al. [331].

terms are often needed in biomedical ontologies, even if they do not in fact represent real phenomena, but rather (false) concepts.

4. The argument from *syndromes*: many entities in the biomedical domain are abstractions, and as such they are mere concepts. Syndromes, for example, are not real entities, but rather clinicians' concepts corresponding to a number of signs and symptoms occurring together.

5. The argument from *error*: in medical records, it is often necessary to record statements that are known to be incorrect. Such statements do not speak of the reality, but rather about concepts about it.

6. The argument from *multiple perspectives*: patients, physicians, and scientists look at the reality from different perspectives, and may make mutually exclusive statements about it. Since there is just one objective reality, such conflicting statements cannot be about that reality, but must be about the persons' concepts of it.

In what follows, we briefly address the last two arguments; for a more involved discussion, the reader is refer to Smith et al. [331].

When erroneous statements are made in a clinical record, and are subsequently interpreted as being about Level 1 entities, then logical conflicts can arise. If both *p* and *not p* are asserted, there is an apparent contradiction, since in logic (at least in classical logics) these two cannot be simultaneously true. However, such conflicting assertions are often needed to record what has been stated by different people. It has been proposed (e.g., Rector et al. [293]) that this implies that the use of a meta-language should be made compulsory for *all* statements in electronic health records (EHR). That is, these records should be not about entities in reality, but rather about what are called 'findings'. Instead of the contradictory 'it is that *p*' and 'it is that *not p*', records should state that 'X observed that *p*' and 'Y observed that *not p*', or 'X stated that *p*' and 'Y stated that *not p*', so that logical contradiction is avoided. The terms in terminologies devised to serve such EHRs would then all have to refer not to patients, diseases, drugs, etc., but rather to X's

and Y's concepts of them. This, however, blurs the distinction between entities in the patient-side reality (Level 1) and the corresponding cognitive representations on the clinicians' side (Level 2), and opens the door to the inclusion in a terminology of problematic findings-related expressions such as 'absent nipple', 'absent leg', etc.[36] Certainly clinicians need to record such findings. But then their findings are precisely that a leg is absent; not that an *absent leg* is present.

Different patients, clinicians and biologists have their own perspectives on one and the same reality. To do justice to these differences, it is argued, we must hold that their respective representations point, not to this common reality, but rather to their different 'concepts' thereof. This argument has its roots in the work of Ogden and Richards, and specifically in their discussion of the so-called 'semiotic triangle'. The latter is of importance not least because it embodies a view of meaning and reference that still plays a fateful role in the terminology standardization work of ISO (see, e.g., Smith et al. [327]). As Fig. 3.1 makes clear, the triangle in fact refers not to concepts, but rather to what its authors call 'thought or reference' (Ogden and Richards [272]). To understand what this means, we note that Ogden and Richards' account is rooted in a theory of psychological causality.[37] When we experience a certain object in association with a certain sign, then memory traces are laid down in our brains in virtue of which the mere appearance of the same sign in the future will, they hold, *evoke a thought or reference* directed towards this object through the reactivation of impressions stored in memory. The two solid edges of the triangle are intended to represent what are held to be causal relations of 'symbolization' (evocation) and 'reference' (perception or memory) on the part of a symbol-using subject. The dashed edge, in contrast, signifies that the relation between the symbol and the referent — the relation that is most important for the discussion of terminology — is merely imputed.

The background assumption here is that multiple perspectives are both ubiquitous and (at best) only locally and transiently resolvable. The meanings

[36] From the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), a "universal healthcare terminology", http://www.snomed.org/index.html.

[37] See, e.g., Robb and Heil [298] for an overview of this subject.

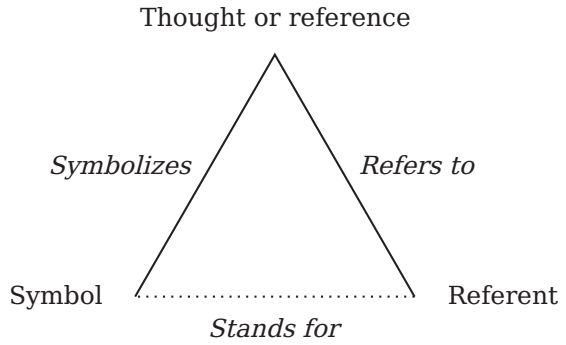

Figure 3.1: The semiotic triangle (Ogden and Richards [272]).

words have for different people depend on their past experiences of uses of these words in different contexts. Ambiguity must be resolved anew (and a new 'imputed' relation of reference spawned) on each successive occasion of use. From this, Ogden and Richards infer that a symbolic representation can never refer directly to an object, but rather only indirectly, via a 'thought or reference' within the mind. It is a depsychologized version of this latter thesis which forms the basis of the concept orientation in contemporary terminology research. The terms in terminologies do not refer to entities in reality (via the Ogden and Richards' imputed relation). Rather, it is held that they refer to concepts occupying a special concept realm. On this view, concepts are not transparent mediators of reference; rather they are its targets, and the job of the ontology developer is to build a list of terms representing not the reality, but rather this special realm of concepts (Smith et al. [327]). The relation between terms in a terminology and the reality beyond becomes hereby obscured. Reality exists, if at all, only behind a conceptual veil — and hence apparently confused statements according to which, for example, the concept of bacteria would cause an experimental model of disease, or the concept of vitamin would be essential in the diet of man.[38]

[38]From the Unified Medical Language System (UMLS), http://semanticnetwork.nlm.nih.gov/.

3.3.4 BFO — A Basic Formal Ontology

A top-level ontology (an upper-level ontology, a formal ontology) is a domain-independent ontology that addresses the most general categories of entities common to all domains. While a top-level ontology does not, by definition, describe any particular domain with a satisfactory level of detail, it provides (ideally) a set of precise, logically formalized and axiomatized categories that can be further extended ('subclassed') by a number of domain ontologies. The purpose of a top-level ontology in this context is to provide a standard system of categories and relations between them, which, if shared by the domain ontologies, should ensure (or at least contribute to) their interoperability.

Unfortunately, there appears to be no chance for a single top-level ontology agreed by all ontology designers — and there are various reasons (see, e.g., Smith [318]). There have been quite a few top-level ontologies proposed, none of which is in complete agreement with the others in the underlying philosophical perspective, categorization of entities, and terminology. In the world of biomedicine, however, some successful attempts have been made to organize the development of domain-specific ontologies in compliance with a unique upper-level ontology, the Basic Formal Ontology (BFO; Smith and Grenon [328], Grenon [136, 137, 140]).[39] The OBO Foundry initiative[40] has collected a number of biomedical ontologies which are now being developed according to a set of organizing principles.[41] These include:

1. The ontologies must be expressed in a commonly shared syntax. Currently, this includes OWL as well as an OBO-specific ontology language.[42]

2. The ontologies have clearly specified and clearly delineated content, and should cover different domains or different perspectives on the same domain, e.g., human anatomy, physiology, and pathology. See

[39] http://www.ifomis.uni-saarland.de/bfo/.

[40] http://www.obofoundry.org

[41] For the complete list of all 10 principles, see http://www.obofoundry.org/crit.shtml.

[42] The OBO Syntax, serialized into a flat file format and a number of XML-based formats (OBO-XML, RDF-XML, OboInOwl); see http://www.geneontology.org/GO.format.shtml.

Fig. 3.2 for a systematic overview of a selection of OBO ontologies, organized according to the level of granularity covered and the top-most categories in BFO (see below for further details).

3. The ontologies use relations defined in the Relations Ontology (Smith et al. [326]), a member of the OBO family.

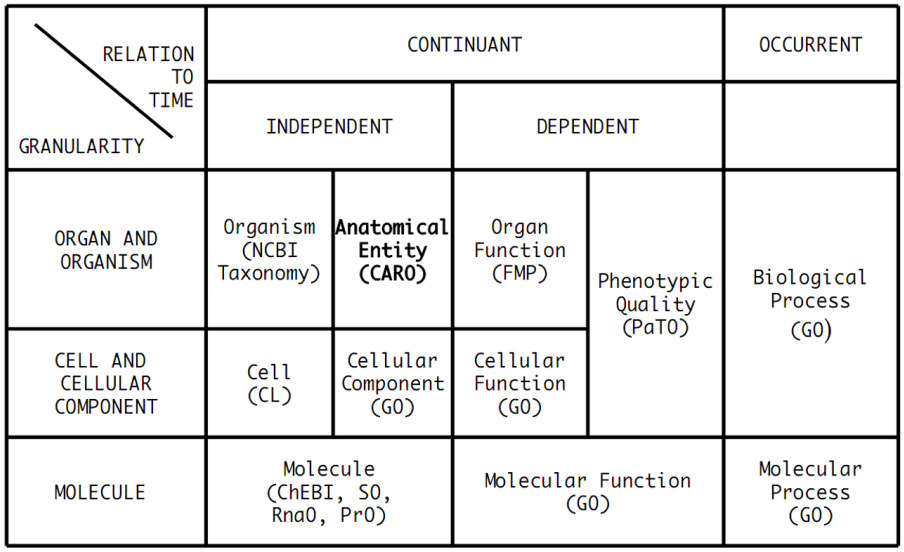

Figure 3.2: Selected members of the OBO family of ontologies, systematized according to levels of granularity (rows) and top-level distinctions in BFO (columns). From Haendel et al. [155].

While not an explicit requirement, the use of BFO as a unique formal ontology has been adopted by the OBO community as another guiding principle. It should be noted that OBO does not encompass all existing biomedical ontologies, with most prominent examples being the UMLS and SNOMED.[43]

[43] Both UMLS and SNOMED CT have been criticized for a number of mistakes and poor design

Entities

BFO is a relatively small, hierarchically organized ontology with single inheritance (see Fig. 3.3). Its top-level category is called 'entity'. An *entity* is anything which exists, including objects, processes, qualities and states on all three levels of reality (see Sec. 3.3.2) — thus including representations, models, beliefs, utterances, documents, observations, etc. Entities need not be material, or even physical. For example, a spatial region (but not what resides in it) is an immaterial physical entity; the concept of a unicorn is an abstract (non-physical) entity.[44] 'Entity' understood as above is a well-established term in philosophy, though this does not imply any sort of agreement on what there actually exists. The term 'existence' itself is quite problematic, and an attempt to define it inevitably engages one in disputes on the opposition between actualism and possibilism, presentism and eternalism, etc. In some formalisms — for example, in the possibilist version of intensional logics (Fitting [114]) — physical existence is expressed explicitly, e.g., by the predicate '$\mathsf{E}!$', and is opposed to mere logical existence expressed with the quantifier '$\exists$'. The axiomatization of DOLCE employs the predicate 'PRE' for asserting presence (roughly corresponding to physical existence). For an overview of various doctrines on what there exists and of different modes of existence, see, e.g., Bacon [24], Menzel [252], Miller [255], Miller [256], Reicher [294], Woleński [389], Yagisawa [394].

principles (Kumar and Smith [215], Ceusters et al. [71, 72], Bodenreider et al. [52], and others).

[44] Whether the concept of a unicorn is a representation or not, is perhaps a matter of taste. If for a mental image to be a representation there must exist — or must have existed, at least — an actual entity recognized by the agent as the object of the representation, then the concept of a unicorn is not a representation. Barry Smith argues (private conversation) that an intention to represent, on the side of the agent, is enough to call a concept a representation.

The term 'abstract' is by no means an unambiguous one. The abstract/concrete distinction has a curious status in contemporary philosophy: it is widely agreed that the distinction is of fundamental importance, but there is no standard account of how the distinction is to be explained (Rosen [301]).

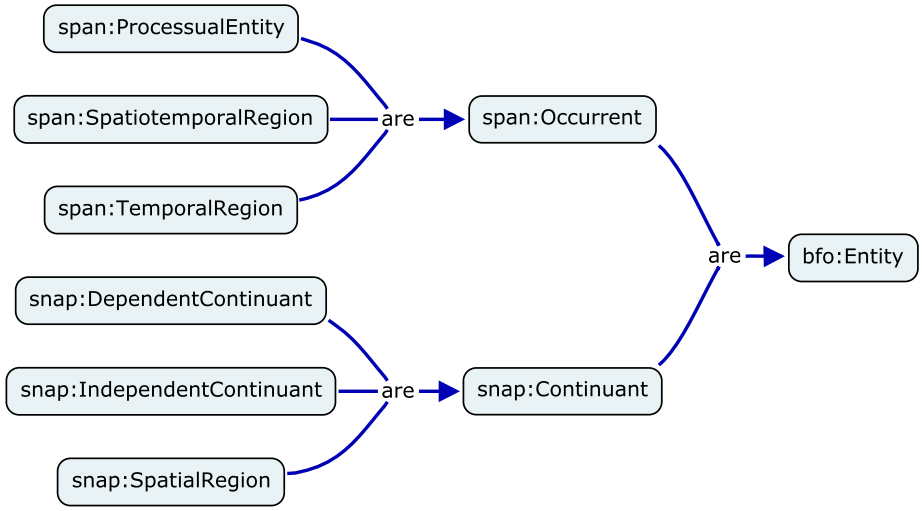

Figure 3.3: Top-most levels of BFO. From http://www.ifomis.org/bfo/1.1.

Universals and Particulars

Among the central distinctions made in the theory of existence underlying BFO is the one made between universals and particulars.[45] A number of attempts have been made to characterize this distinction both formally and informally; see, e.g., Bealer [31], Bittner et al. [45], and Neuhaus et al. [264]. One possibility is to take as primitive the relation of *instantiation*; *universals* are entities that can have instances, while *particulars* are entities that cannot have instances. Again, there are definitions to be found that blurr this distinction, e.g.,

> "A particular is a single thing, thought of in contrast to qualities or universals, or in contrast to an aggregate of things. Universals

[45] The term 'particular' is sometimes taken to be synonymous with 'individual' or 'token', though the latter is often considered to have a more specific meaning (Wetzel [383]).

> themselves can be regarded as particulars [*sic*], themselves having higher-order properties and relations. However, a universal can be instanced by particular things, whereas a genuine particular cannot." (Blackburn [46])

BFO allows no higher-order instantiation: universals have only particulars as instances, and no particular is a universal. In the Aristotelian version of realist ontology endorsed by BFO, universals exist only in virtue of their being instantiated by particulars, i.e., there are no universals of which there are no instances. In this view, then, universals are entities that do have instances, and particulars are those that don't (thus we are dispensed with the possibilist 'can' or 'may'). If time is considered, universals may exist at some times — precisely, when some particulars instantiating them exist; universals do not exist those times when there are no particulars which would instantiate them (Neuhaus et a. [264]). Interestingly, in DOLCE all top-level universals[46] are asserted non-empty, i.e., they are instantiated at all times, but it does not have to hold for more specific universals.

The distinction between universals and particulars is sometimes explained in terms of the way natural languages are used to talk about reality. It is said that universals are referred to by *general terms*, such as 'dog', 'fracture', or 'biosynthesis'. In contrast, particulars are referred to by means of *proper names*, such as 'Osama Bin Laden' or 'Fido', or by means of complex expressions involving general terms, indexicals, etc., as in 'this dog here' or 'your intervention yesterday'. This, however, does not mean that every general term denotes a universal, and that every proper name denotes a particular. Likewise, in a logical formalism predicates (corresponding to general terms in natural language) may be seen as denoting universals, and constants and variables (corresponding to proper names) may be considered as denoting (or ranging over) particulars. This does not have to be true, however:

– it is held (by realist ontologists endorsing the so-called 'sparse' theory of universals) that composite predicates, such as 'P-or-Q', do not

[46]DOLCE has its own notion of top-level universals, called 'basic categories'. See Masolo et al. [244, p. 14].

denote universals, even if their component predicates 'P' and 'Q' do;[47]

- there are general terms purported to denote universals, for which there has never happend to exist an instance, e.g., 'unicorn';
- there are general terms purported to denote universals, for which there can never exist an instance, e.g., 'cubic sphere' or 'square circle'.

Furthermore, it is (obviously) possible to use, in a logical formalism, constants to denote not only particulars, but also universals. This is what, for example, Neuhaus et al. [264] do in their formal theory of substances, qualities, and universals, and Bittner et al. [45] in their formal theory of individuals, universals, and collections. There, however, care is taken not to confuse particulars and universals — referred to by means of constants and variables — by the use of multisorted logic. But in some biomedical ontologies care is not taken to keep these distinct. For example, in their Cell Component Ontology (CCO),[48] Karp et al. suggest that it is a matter of taste whether an entity is represented as a class (roughly corresponding to a universal in that context) or as an individual; thus they have the class-terms 'organelle', 'vacuole', etc., but 'nucleus' as an individual-term:

> "We could have chosen to make "nucleus" be a class instead of an instance, however, for the purposes for which this ontology was designed, we wished to emphasize in our model the notion of cellular structures as the base-level objects of discourse. ... Our instance "nucleus" refers to whatever cell or population of cells the user of our ontology is choosing to treat as a single entity (an instance) in their model."[49]

In the OBO community, it has been agreed[50] to adopt the policy that it is exclusively universals that should be represented in bioontologies. This, however, is by no means a comfortable decision. Numerous, sometimes

[47] The (Boolean) composition in this case is such that the extension of P-or-Q is the union of the extensions of P and Q.
[48] http://brg.ai.sri.com/CCO/.
[49] Peter Karp, private communication.
[50] It may, at least, seem so.

fierce debates on BFO- and OBO-related discussion lists[51] provide evidence that the criteria for distinguishing general terms that denote universals from those that do not are still far from obvious, and that further clarifications are needed.

The problem of universals has been debated for centuries (Klima [204]), and the recent literature is confusing in terminology, incompatible standards for evaluating theories, and "philosophers talking past one another" (Swoyer [361]). The view that universals exist as entities in the objective reality is by no means the only possible one. Against this *realist* stance various arguments have been proposed; very roughly, *conceptualists* claim that general terms refer not to universals (because, according to them, there are no such entities), but rather to concepts within the minds of cognitive agents.[52] *Nominalists* claim that general terms do not refer to anything else than to (multiple) particulars (and, similarly to conceptualists, that there are no such entities which would correspond to the name 'universal').[53]

Continuants and Occurrents

Another top-level distinction essential for all OBO ontologies is the one between continuants and occurrents (Smith et al. [326, 331]; see Fig. 3.3), corresponding to the division of BFO into two forms, the time-instant form (the snapshot form, SNAP) and the time-interval form (the span form, SPAN). *Continuants* are those entities that exist wholly or fully at any time at which they exist at all, even if they may have different parts at different times of their existence. *Occurrents* are those entities whose existence extends in time, and at a particular time only some of their temporal parts are present. (More accurately, BFO takes an eternalist view with respect to occurrents; occurrents exist throughout the whole of time (which is an occur-

[51] For example, bfo-discuss@googlegroups.com and obo-cell-type@lists.sourceforge.net

[52] Conceptualism is the theory of universals that sees them as shadows of our grasp of concepts. (. . .) A concept is that which is understood by a term, particularly a predicate. (Blackburn [46])

[53] Nominalism is the view that things denominated by the same term share nothing except that fact: what all chairs have in common is that they are called 'chairs'. The doctrine is usually associated with the thought that everything that exists is a particular individual, and therefore there are no such things as universals." (Blackburn [46])

rent itself), and are *located* (rather than *present*) in various temporal and spatio-temporal regions.) Continuants and occurrents are sometimes called 'endurants' and 'perdurants', respectively, though these terms are not necessarily taken by all philosophers as synonymous with the other. Continuants are further divided into independent and dependent continuants, and spatial regions (Fig. 3.4).

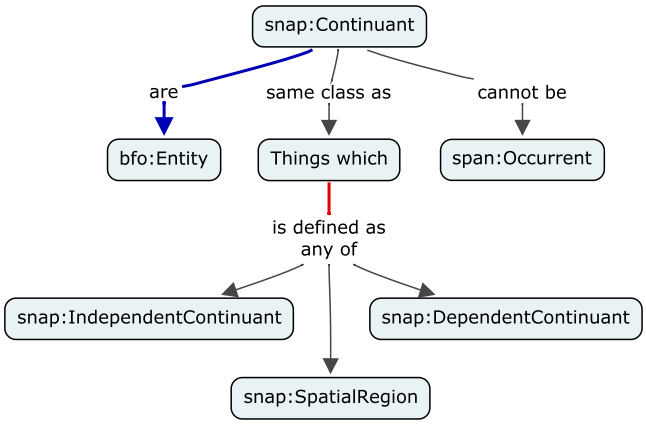

Figure 3.4: Continuants in BFO. From http://www.ifomis.org/bfo/1.1.

The continuant-occurrent distinction is not a universally accepted one; four-dimensionalists, for example, hold that both continuants and occurrents are only partially present at time instants, and that entities of both these kinds have temporal parts (temporal slices). The nature of time — its density, continuity, and the possibility of perception of time instants (indeed, the very existence of time instants) — is also far from enjoying a uniform view among philosophers (Markosian [242], Poidevin [286]). For a recent defense of the continuant-occurrent distinction endorsed by BFO, see Grenon and Smith [139].

Dependent and Independent Entities

The third major distinction essential to BFO is that between dependent and independent entities. More accurately, the distinction is between dependent and independent continuants, as BFO does not apply the dependent-independent criterion to occurrents. A *dependent entity* is an entity that for its existence requires the existence of another entity, its *bearer*, in which the former *inheres*. Functions, such as those represented in the Molecular Function branch of the GO, are dependent entities; they inhere in other entities, such as molecules. Other dependent continuants are *qualities* (e.g., colors, shapes, weights), *roles* (e.g., the role of a person as a surgeon), and *dispositions* (e.g., of the blood to coagulate).[54] Further distinctions are made, e.g., between specifically dependent entities (entities that depend specifically on some other particular entities) and generically dependent entities (entities that depend on instances of some universals, but not necessarily on the same instances at all times).[55] For the purpose of biomedical ontology, qualities are specified in much more detail in the Ontology of Phenotypic Qualities (PATO).[56]

Relations

The term 'relation' is pervasively used in ontological engineering to speak of both what holds between two universals and what holds between two particulars; instantiation of a universal by a particular is sometimes seen as a relation as well. However, it is useful to keep these two types of relations distinct, and to provide a consistent rule of quantification for relations between universals. Early versions of the Gene Ontology, for example, did not provide precise specifications of what the links 'is a' and 'part of' mean, and how the corresponding relations are to be quantified (Smith et al. [318]). Recently, the OBO community has adopted and continuously develops a

[54] Recent discussions on BFO-related mailing lists prove that the distinction between functions, roles, and dispositions is, in the context of biomedical ontologies, still unclear. There is an ongoing effort aimed at clarifying this branch of BFO.

[55] Again, this distinction has been recently subject to hot debates among OBO developers.

[56] http://www.bioontology.org/wiki/index.php/PATO:Main_Page.

specification of formal relations to be used in biomedical ontologies (Smith et al. [326]). One of the leading principles is that universal-universal relations have the all-some quantification: a universal U_1 is R-related to another universal U_2 just in case *all* instances of U_1 are r-related to *some* instances of U_2 (r is the instance-instance counterpart of R; see Smith et al. [326] and Kuśnierczyk [218] for more details).

Chapter 4 presents the specific problem of defining relations between terms in the GO and terms in the Taxonomy of Species, and chapters 5–A show how they can be defined so as to enable various patterns of inference. One of the core ideas in the framework presented in those chapters is that it may be useful to depart from the default all-some quantification pattern and provide a means for explicitly acknowledging patterns some of which are not available in common description logics-based formalisms.

3.3.5 Discussion

This chapter discusses the need for a unique, coherent top-level ontology and terminology that could be used to guide and bridge the development of different domain-specific ontologies. A possible candidate is the Basic Formal Ontology, BFO, which is briefly covered here; for more details the reader is referred to the numerous publications related to BFO mentioned in this chapter, and to BFO's official website for most recent news. While BFO itself is in a relatively stable form, there has recently been much effort dedicated to its adoption as the basis for the development of Open Biomedical Ontologies (Smith et al. [324]). The work is in progress, however, and many issues remain to be resolved. Despite the sound philosophical foundation that BFO provides for ontologies developed according to its principles, it is not entirely clear whether it would or would not be possible to achieve the same or better performance using a different philosophical basis; it is not clear whether the distinctions made in BFO, e.g., between bona fide and fiat entities, between dependent and independent entities, or between relational, dispositional, and functional entities are helpful for the purposes of biomedical ontologies. Discussions on BFO-related mailing lists provide

evidence that often it is not clear to domain experts how to classify an entity using BFO's top-level categories. It is also unclear whether enforcing a single view on how to represent the biological domain helps or rather suppresses the development of biomedical ontologies; in general, the idea of having a unique top-level ontology for all purposes and all communities has been largely abandoned in ontological engineering. While one of the goals of the activities reported here is to establish a commonly agreed top-level ontology in the rather limited context of biomedicine, it remains to be proved that such an approach is both feasible and desirable, the whole community taken into consideration.

The OBO Foundry approach has been successful in unifying multiple ontologies that cover the same domain but were developed by different teams; for example, the cell type ontology of Bard, Rhee and Ashbuner [28], the cell type ontology of Kelso et al. [195], the cell type-related terms in the Gene Ontology, and the cell type-related terms in the Foundational Model of Anatomy (FMA; Rosse and Mejino [302]) have all been merged into the OBO cell type ontology (CL; Mabee et al. [241]). The role of so-called 'reference ontologies',[57] such as the FMA and a few other OBO ontologies, in the integration of the Semantic Web has also been established (Brinkley et al. [59]). Furthermore, there are ongoing efforts to adopt the OBO Foundry principles and reuse its member ontologies in projects such as the NeuronDB database,[58] BIRNLex, the controlled vocabulary designed to annotate data sources of the Biomedical Informatics Research Network (BIRN),[59] or the Minimum Information for Biological and Biomedical Investigations initiative (MIBBI).[60] Nevertheless, the correctness, usefulness, or superiority of the OBO Foundry approach has not been the subject of any benchmark test yet.

[57] Ontologies intended to be reused in multiple application contexts rather than designed for a single, specific application.

[58] http://senselab.med.yale.edu/neurondb/.

[59] http://www.nbirn.net/.

[60] http://mibbi.sourceforge.net/.

Chapter 4

Subsetting the Gene Ontology with GO Slims

In this chapter we introduce the idea of partitioning the Gene Ontology based on criteria corresponding to various taxonomic contexts. We analyze how suitable for this task are the so-called 'GO slims', manually created subsets of terms from the three branches of the Gene Ontology, the molecular function, biological process, and cellular component ontologies. The chapter is structured as follows:

- Section 4.1 motivates the need for subsetting the Gene Ontology.
- Section 4.2 discusses the issue of generality and species-specificity of GO terms.
- Section 4.3 introduces GO slims, the current approach to providing subsets of the GO. We discuss a number of issues related to how the slims are constructed and maintained, and assess their reliability as a source of species-specific subsets of GO terms.

Chapter 5 and Appendix A provide a detailed description of a framework designed specifically to address the issue of aligning[1] the Gene Ontology with the Taxonomy of Species.

4.1 Introduction

During the past decade there has been a rapid growth of interest in ontological engineering, i.e., in designing, implementing and deploying structured representations of various real world domains (Devedžić [91], Gómez-Pérez [133], Mizoguchi [258]). One of the most visible testimonies of this trend is the ontological activity in biomedicine and bioinformatics, perhaps best represented by the Open Biomedical Ontologies project (OBO) and its successor OBO Foundry (Smith et al. [324]).[2] The Gene Ontology (GO; Ashburner et al. [367]), which served as the initial kernel of OBO and successfully continues to be its driving force, is the result of an effort aimed at providing a structured, precise, shared vocabulary for describing roles of genes and gene products in any organism (the Gene Ontology Consortium [366, 367]).

Ideally, ontologies should be built to enable automated agents to communicate and process information as if they had the understanding of the domain that human experts possess. On the other hand, ontologies are valuable sources of explicit, systematized knowledge for human users, especially those who are not experts in the domain represented by a particular ontology (Mizoguchi [258]). Automated agents use an ontology to retrieve relevant pieces of information (e.g., entries in a database) or whole documents (e.g., scientific publications) *annotated* with its terms, but it is human users that must understand the ontology to provide such annotations. Even if annotations are suggested by automated services, such as text mining of scientific literature, it is still human experts that must curate — accept, modify,

[1] Ontology alignment is typically defined as the (semi)automatic discovery of correspondences between ontologies that are distinct representations of the same domain (Davies et al. [86]). Recently, the OBO community has adopted the term 'bridge' for describing connections between modularized ontologies, as sketched in Rector [292].

[2] http://obo.sourceforge.net/; see also Sec. 2.7.

or reject — those suggestions. This dual role of ontologies is reflected in, e.g., the principle of intelligible definitions: definitions within an ontology should be both humanly intelligible and formally specifiable (Smith [321]).

Biomedical ontologies grow considerably in number, size and coverage; this trend is clearly illustrated by the expanding list of bioontologies available on the OBO Foundry website.[3] It has long been recognized that large ontologies and knowledge bases need to be partitionable into subsets tailored for the convenience of both a human and an automated agent (see, e.g., Wouters et al. [390], Bhatt et al. [43]). On the one hand, focusing on a narrow partition improves the efficiency of inference; on the other hand, human users may be presented with a constrained and tersely expressed portion of knowledge, relevant to the particular problem at hand. In the case of OBO, partitioning of the represented biomedical domain is realized in two ways:

- Firstly, the whole of our biomedical knowledge is divided into a number of separate ontologies covering different subdomains, such as gross anatomy of *Drosophila*, human diseases, cellular components, protein-protein interactions, etc. This partitioning reflects the perspective on reality taken by the authors of the ontologies.
- Secondly, terms within a single ontology may be selected or hidden to provide a partial view of the partition of reality represented by the ontology. This subpartitioning reflects a user-defined perspective on reality (its part captured within the scope of the chosen ontology).

The Gene Ontology was originally designed to be a vocabulary that could be applied to all eukaryotes (Ashburner et al. [19]), but now includes a large number of terms that cover gene products found in prokaryotes and viruses as well. The vocabulary of the Gene Ontology is comprehensive, in that its taxonomic scope ranges from species-specific terms to the most general terms representing features found in organisms of all sorts. One consequence of this development is that not all GO terms are necessarily of interest to a researcher focused on organisms of a particular kind, such as

[3] http://www.obofoundry.org/.

viruses, flowering plants, or fruit flies; not all terms need to be visible to every user, and many terms will be irrelevant to the inferences a particular experiment may require to be carried out. In Ch. 5, we show that by relating terms in the Gene Ontology to terms in the Linnaean Taxonomy of Species (TS; see, e.g., Ereshefsky and Matthen [103]) one is able to partition the GO according to various taxon-related criteria, and answer taxon-specific queries such as '*What biological processes take place in vertebrates but not in humans?*', '*What are the most specific classes of molecular functions that can be found in all vertebrates?*', or '*What are the most general classes of cellular components that cannot be found in organisms other than yeasts?*'. Furthermore, the system should also be able to process inversed queries, e.g., "*What are the organisms in which syncytia can be found?*", etc. The approach is novel, in that it substantially departs from how species-specificity is currently addressed with the GO slims technology.

Material The analysis of the structure and content of GO slims is based on data obtained from the Gene Ontology downloads site.[4] Where taxonomical information is involved, my reference is data obtained from the NCBI Taxonomy downloads site.[5] Both databases were most recently accessed in January 2007; both sources are also available for online browsing. It should be noted that many of the problems with species-specificity of GO terms mentioned in this chapter have been recently repaired, or at least addressed in some way by the developers of the Gene Ontology, partially following the criticism and suggestions included in theis and the next chapters.

4.2 The Generality and Specificity of GO Terms

From its very beginning, the Gene Ontology project has been a collaborative effort to construct and use ontologies to facilitate the biologically meaningful annotation of genes and their products in a wide variety of organisms (the Gene Ontology Consortium [367]), initially in eukaryotes. Currently,

[4] http://www.geneontology.org/.
[5] http://www.ncbi.nlm.nih.gov/Taxonomy/.

the GO aims at providing a controlled vocabulary that can be used to describe *any* organism. How should the stress on 'wide variety of organisms' and 'any organism' be understood? Clearly, a number of functions, processes and components are not common to all life forms. There are many such non-universal features[6] — features that are *not* found in some kinds of organisms — represented within the GO; more or less trivial examples are discussed later throughout this chapter.

In fact, many high-level GO terms represent features that do not appear in all life forms — features that have never been, and are not supposed to be, found in organisms of some specific types. (In this chapter, the terms 'high-level GO term' and 'general GO term' are used to speak of those GO terms that are placed relatively close to the root term of the ontology they belong to, irrespectively of how well the term-wise distance from the root reflects what a domain expert would consider a term's specificity.) As an example, consider structures of the type represented by the term cell — i.e., cells, the basic structural and functional units of all organisms; clearly, cells are *not* components of viruses. Likewise, structures of the type represented by the term virion — i.e., virions, complete viral particles — are *not* (canonically, at least) components of cellular organisms. I constrain myself here from the discussion of whether viruses — acellular structures — should be considered live organisms, or even simply organisms, or not; it suffices to note that the Gene Ontology does provide a vocabulary that describes features found only in viruses. The terms bioluminescence, photosynthesis, locomotion, and hatching are other relevant examples, and many other can be easily found.

4.3 The GO Slims

How is the dependence between GO terms and types of organisms addressed currently in the Gene Ontology? Can this approach be improved or ex-

[6] A terminological note: henceforth, molecular functions, biological processes and cellular components will collectively be referred to as *features of organisms*, or just *features*. This shorthand allows for an easier to read text; my use of the term 'feature' here should not be confused with its various other uses in biological terminology.

tended? The observation that there are terms which do not represent universal features — features found in organisms of all sorts — has, among many other reasons, led to the implementation of a simple approach to creating constrained views of the GO, the so-called 'GO slims'. Slims are versions of the GO ontologies in which the more specific terms (and therefore their annotations) have been collapsed up into the more general parent terms; for example, style development can be collapsed into its ancestor flower development (Clark et al. [79]). In general, if g_1 and g_2 are two GO terms such that g_1 is an ancestor of g_2, then in a slim that would include g_1 but not g_2 all annotations made with g_2 would be shifted up to the term g_1 (within the slim). That is, effectively, every gene product annotated with g_2 is annotated with g_1 in the slim. In the example given by Clark et al., the hypothetical slim contains the term flower development but not the term style developement. Since, according to the Gene Ontology, style developement is a part of flower development, the latter inherits all anotations of the former; specifically, the annotations of the *A. thaliana* genes STY1 and STY2 with style development are inherited[7] by floral organ development.

According to the online documentation of the Gene Ontology, slims are cut-down versions of GO ontologies containing a subset of the terms in the whole GO; they give a broad overview of the ontology content without the detail of the specific fine grained terms. Slims are created by users according to their needs, and may be specific to species or to particular areas of the ontologies.[8] Slims are not constrained to providing only general views (views including only general terms). One may thus create any view of the ontology, in particular one that contains only 'fine-grained' terms. There are a few officially supported slims: the Plant slim, the Yeast slim, the Generic slim (a slim that is not species specific, and should be suitable for most purposes), and the Gene Ontology Annotation Database (GOA) slim,[9] all of which can be found in the official OBO-format file distribution of the GO. Many other slims, e.g., the Honey bee ESTs slim, the *Drosophila* slim, the Rice slim, etc., have been archived and are available for download from the

[7] Note the upward direction of this inheritance pattern.

[8] GO slim guide, http://geneontology.org/GO.slims.shtml.

[9] The GOA slim contains 'high-level terms' that cover the main aspects of each of the three GO ontologies (http://www.ebi.ac.uk/GOA/).

GO website.

To analyze data annotated with a subset of GO terms, one may use any of the existing slims, or create one anew. To create a slim with some particular intended coverage, all GO terms deemed relevant for that purpose have to be manually selected. In practice, within an OBO format file, the subsetdef tag is used to introduce the slim's identifier and name, and every term to be included in that slim is marked with the subset tag together with the slim's identifier; see the GO File Format Guide[10] for further details and examples. Figure 4.1 shows an excerpt from the Generic GO slim, including a fragment of the preamble and a modified entry for the term reproduction.

To constrain the view of the whole of the Gene Ontology to what is covered in a particular slim, one simply needs to select all those terms that are tagged correspondingly. This can be easily done with an OBO-compliant tool, such as OBO-Edit (Day-Richter et al. [88]). For example, in the standard OBO-format file the term bud is tagged as belonging to the Yeast slim, while the term synapse is not; therefore, the former will appear in the yeast-related view of the Gene Ontology based on that file, while the latter will not.

This approach to contextualization of the Gene Ontology is fairly straightforward. A number of publications can be found in which GO slims are reported to have been used, e.g., to provide an overview of the functional composition of proteomes (Kanapin et al. [191]), categorize proteins according to the processes they participate in (Wohlschlegel et al. [388]), or to improve estimation of missing values in microarray experiments (Tuikkala et al. [372]). However, as the following observations make clear, one should be careful in making assumptions about a GO slim's generality or species-specificity, which would have impact on the correctness or relevance of an analysis based on the slim.

[10] http://geneontology.org/GO.format.shtml.

```
format-version: 1.2
date: 21:06:2007 16:45
saved-by: jane
auto-generated-by: OBO-Edit 1.101
subsetdef: goslim_generic "Generic GO slim"
subsetdef: goslim_goa "GOA and proteome slim"
subsetdef: goslim_plant "Plant GO slim"
subsetdef: goslim_yeast "Yeast GO slim"
subsetdef: gosubset_prok "Prokaryotic GO subset"

[Term]
id: GO:0000003
name: reproduction
namespace: biological_process
def: "The production by an organism of new individuals that contain
      some portion of their genetic material inherited from that organism."
     [GOC:go_curators, GOC:isa_complete, ISBN:0198506732
      "Oxford Dictionary of Biochemistry and Molecular Biology"]
subset: goslim_generic
subset: goslim_plant
subset: gosubset_prok
synonym: "reproductive physiological process" EXACT []
is_a: GO:0008150
```

Figure 4.1: A modified fragment of the Generic GO slim. The preamble includes a few subset definitions (note the subsetdef tags). The term reproduction is assigned to three of theses subsets (note the subset tags). From http://www.geneontology.org/GO_slims/goslim_generic.obo.

4.3.1 GO Slims Have Imprecisely Defined Scope

Slims are created by users according to their needs, but, due to this very fact, they do not necessarily satisfy the needs of a broader community. The existing slims lack definitions that would precisely describe their content. Slims were introduced early on in the life of the Gene Ontology (Adams et al. [10]), and have since remained *ad hoc* customized views. For research purposes, it is not advisable to guess a slim's scope and goals from its name alone; e.g., one should not make assumptions about yeast-specificity of the Yeast slim. Unfortunately, the documentation of GO slims is not very helpful

in this respect.

GO slims have no explicit, precise criteria for the inclusion of terms. Specifically, it is not clear what it means that slims give a broad overview of the ontology content without the detail of the specific fine-grained terms; the notion of 'fine-grainedness' of GO terms is rather intuitive and imprecise. As illustrated in Fig. 4.2, there are terms excluded from the Generic slim despite their ancestors and successors being included in this slim. Whatever intuitive understanding of the specificity of GO terms one may have, it seems to be contradicted by this example.

[GO:0008150 biological process]
    GO:0009987 cellular process
        [GO:0007154 cell communication]
            [GO:0007267 cell-cell signalling]
        [GO:0008037 cell recognition]
            GO:0009988 cell-cell recognition
        GO:0050875 cellular physiological process
            [GO:0007049 cell cycle]
    [GO:0007582 physiological process]
        GO:0050875 cellular physiological process
            [GO:0007049 cell cycle]

Figure 4.2: A partial view of the biological process branch of the Gene Ontology. Indentation reflects subsumption of terms. Terms included in the Generic slim are in square brackets. The term cellular process is excluded from the slim, although both the term biological process (its ancestor) and the term cell recognition (its successor) are included in it.

A simple approach to estimate a term's specificity (or its inverse, generality) is to use the count of ancestors of the term. The terms biological process and physiological process (see Fig. 4.2) would have, following this line, equal generality; the term cell-cell signalling would be more specific than the term cellular physiological process, though the former is not actually a successor of the

latter.[11] An alternative approach, implemented in the Gene Ontology Partition Database (Alterovitz et al. [12]), is to employ information-theoretic calculations, based on the count of annotations associated with GO terms. However, in general, such count of still incomplete annotations seems to reflect the interest and activity of particular research communities rather than any sort of term generality. This approach will not necessarily be reliable until all genes and gene products in all organisms have been fully annotated.

These two approaches are based on the static structure of the GO and on its so-called 'information content', respectively. In either case, the result may not correspond well to the specificity of terms as it could be understood by domain experts. Computational assessment of the specificity of GO terms is not an infrequent topic of discussion on the GO-friends[12] mailing list.

4.3.2 'Species-Specificity' Has Imprecise Meaning

Until only recently, many terms in the Gene Ontology were said to be species-specific, and marked as such by a 'sensu ...' inclusion in the name and an 'as in, but not restricted to, ...' inclusion in the definition (where the ellipses stand for a taxon name and a taxon description, respectively). As an example, Fig. 4.3 shows the term proteasome complex (sensu Eukaryota), encoded in the OBO file format. In an effort to clarify the intentions, 'sensu ...' has been replaced by 'sensu ... research community', but this change was later reversed; currently, 'sensu' terms are being modified so that their names reflect the actual differentiating criteria rather than the use of a term by a particular research community. For example, vacuolar lumen (sensu Magnoliophyta) has been replaced with lumen of vacuole with cell cycle-independent morphology, etc.

'Species-specificity' is also invoked in the description of slims. For exam-

[11] This simple approach is somewhat complicated by the fact that the Gene Ontology allows for multiple inheritance, i.e., a term may have more than one path to the root of the ontology, and the paths may be of different lengths. *Ad hoc* solutions to this problem include considering the maximal, minimal, or average distance from the root as a measure of specificity.

[12] http://www.geneontology.org/go.list.gofriends.shtml.

```
[Term]
id: GO:0000502
name: proteasome complex (sensu Eukaryota)
namespace: cellular_component
def: "A large multisubunit complex which catalyzes protein degradation.
      This complex consists of the barrel shaped proteasome core complex
      and the regulatory particle that caps the proteasome core complex.
      As in, but not restricted to, the eukaryotes (Eukaryota,
      ncbi_taxonomy_id:2759)."
     [GOC:rb]
synonym: "26S proteasome" NARROW []
is_a: GO:0043234 ! protein complex
is_a: GO:0044424 ! intracellular part
```

Figure 4.3: The OBO-format entry for the term proteasome complex (sensu Eukaryota).

ple, the Prokaryotic GO subset is "a prokaryote-specific subset of GO terms, [which] contains only terms that are applicable to prokaryotes".[13] The subset may thus include terms that are applicable *not only* to prokaryotes (the definition does not state that it contains only terms that are applicable *only* to prokaryotes); it may also exclude some terms that *are* applicable to prokaryotes (the definition does not state that it contains *all* terms that are applicable to prokaryotes). The Plant slim contains the term biological process, which clearly is not plant-specific (plants are not the only organisms in which biological prcesses take place). It also contains the term extracellular matrix (sensu Metazoa), whose name suggests specificity to animals; it should rather include a term such as extracellular matrix (sensu Viridiplantae).[14]

4.3.3 Relations Between Taxa Are Neglected

Although some of the slims seem to have been intended as collections of terms corresponding (in a rather underspecified sense) to particular taxa,

[13] From http://geneontology.org/GO.slims.shtml.
[14] Not an actual GO term, I made this up for the sake of the example.

the taxonomic relations which hold between the classes of organisms referred to by the slims have not been taken into consideration. Thus, for example, the Plant slim contains only some of the terms included in the Rice slim, and the latter contains only some of the terms included in the former. For most of the taxa that are explicitly associated with the so-called 'species-specific' terms (terms with a 'sensu' clause in their names), there are no slims defined (see Fig. 4.4).

Archaea (superkingdom) ...
Bacteria (superkingdom) ...
Eukaryota (superkingdom)
Apicomplexa (phylum/division)
Viridiplantae (kingdom) ...
Fungi (kingdom)
Candida albicans (species)
Schizosaccaromyces (genus)
Saccharomyces (genus)
Metazoa (kingdom)
Protostomia ...
Deuterostomia
Vertebrata (subphylum)
Mammalia (class)
Actinopterygii (class)
Amphibia (class)
Nematoda (phylum)

Figure 4.4: Partial hierarchy of the Taxonomy of Species, based on the NCBI Taxonomy. Indentation reflects taxonomical ranks and subsumption. Only taxa that are explicitly referred to by GO terms (by means of 'sensu' clauses in their names) are shown. Ellipses signallize subtaxa referred to by GO terms, which we omitted in the figure. Where available, taxon ranks are given in parentheses.

4.3.4 Slims Are Built Manually

Terms are assigned to slims manually, and each term to be included in a slim must be explicitly tagged as such. The criteria for adding terms to a slim are unclear, and thus it is not obvious how to implement measures for automatically controlling the slim's completeness (in the sense of the slim's including all those terms that are relevant for its purpose) and consistency (in the sense of the slim's not including those terms that are not relevant for its purpose). Due to incomplete documentation, even a human expert may not be able to assess the coherence of and adherence to a slim's policy concerning inclusion and exclusion of terms — because the policy is not carefully specified. Some of the examples given above — e.g., the term extracellular matrix (sensu Metazoa) found in the Plant slim — are likely the result of a mistake. This is not entirely clear, however, since extracellular matrix (sensu Metazoa) represents extracellular matrix as it is in, but not necessarily *only in* animals; the term could thus reasonably be seen as describing a feature found also in plants.[15]

4.3.5 Slims Are Updated Manually

GO slims do not automatically reflect changes done to the GO ontologies, other than those involving terms already included in the slims. Consider an insertion of a new term between two other terms already included in a slim. The new term becomes an ancestor of one of the old terms, and a successor of the other. However, it will not be automatically included in the slim — the slim has to be updated manually for this change to be visible in it. The case of the term cellular process excluded from the Generic slim (see Fig. 4.2 again) might be a valid example here.

[15] Recently, extracellular matrix (sensu Metazoa) has been replaced with proteinaceous extracellular matrix, and the definition no longer refers to the taxon *Metazoa*, and the inclusion of this new term in the Plant slim is much more reasonable.

4.4 Discussion

This chapter reviews GO slims as an approach to subsetting the structure of the Gene Ontology. Since my interest lies in connecting the Gene Ontology and the Taxonomy of Species, here we focused on how the slims address the issue of species-specificity of GO terms. It appears that the issue of applicability of terms in various taxonomic contexts is addressed only very loosely in the Gene Ontology. The so-called 'species-specificity' of GO terms is not defined precisely, and cannot be used for any inference regarding the relations between terms and taxa. Likewise, nothing should be assumed or inferred about the genericity of the Generic slim, or the species-specificity of the Plant or Yeast slims. Slims lack precise documentation and clear term inclusion and exclusion criteria, and should only be used with care, if by anyone besides their original authors.

The observations made in this chapter motivate the development of a framework for connecting the Gene Ontology and the Taxonomy of Species in a precise and effective way. The following chapters present such a framework.

Chapter 5

Connecting the Gene Ontology and the Taxonomy of Species

This chapter informally introduces a framework that allows for a substantial improvement in performance, consistency, and flexibility of creating taxon-based partitions of the Gene Ontology, as compared to what can be obtained using the traditional GO slims. The framework is not intended to be a replacement of slims, however, as its scope is limited to the very specific problem of selecting terms based on their relations with taxa, while slims may be created according to criteria of any other sort.

The chapter is structured as follows:

- Section 5.1 introduces the idea of explicitly linking GO terms with taxonomic terms from the Taxonomy of Species.
- Following this observation, Sec. 5.2 presents, in the form of informal but precise definitions, a framework that allows one to take advantage of the hierarchical structure of both the Gene Ontology and the

Taxonomy, and to partition the GO dynamically, based on taxonomic selection criteria.

- Section 5.3 Shows how the links between the Gene Ontology and the Taxonomy can be automatically propagated along both structures, to decrease the need for manual annotation.
- The framework is then illustrated with an example in Sec. 5.4.
- Section 5.5 discusses related philosophical, terminological, logical, and other practical issues.

5.1 Introduction

The three ontologies collected under the common title 'Gene Ontology' — the ontology of molecular functions, the ontology of biological processes, and the cellular components ontology — are hierarchical structures composed of linked terms. Terms represent types of biological entities, and term-term links represent relations between those types. Currently, links between GO terms represent two kinds of relations: subsumption (represented by the so-called 'is a' links) and meronymy (represented by the so-called 'part of' links). Both relations are partial orders; they are transitive and antisymmetric, and do not form cyclic structures — the GO ontologies are (representable as) directed acyclic graphs (DAGs). The following presentation focuses on the 'is a' relation; issues relevant for the 'part of' relation are briefly discussed in Sec. 5.5. Moreover, it should be noted that the GO is likely to undergo substantial changes in the nearest future, changes that would introduce a number of other relations (see, e.g., Wroe et al. [391], Mungall [260], Smith et al. [326], Myhre et al. [261]). Further study would thus be needed if this solution were to be adapted to such an extended Gene Ontology.

GO terms may be associated with *taxa* — classes (roughly corresponding to types, kinds) of organisms. The classes are related by subsumption and represented as a hierarchical structure called the Taxonomy of Species (TS).

The ontological nature of species, and of taxa in general, is a matter of philosophical debate; see, e.g., Ereshefsky [101, 102] for a detailed discussion of species and of the Linnaean and other taxonomies. In what follows, taxa of sub-species ranks will be ignored, with no loss of generality of the discussion. The NCBI Taxonomy of Species database is referred to as a concrete implementation of TS; all of the information on taxa mentioned in this chapter comes from that database. Figure 4.4 presents a small portion of the Taxonomy of Species, including the taxa explicitly referred to by GO terms. This chapter shows how the hierarchical structure of the Taxonomy can be used to produce constrained views of the GO faster and more flexibly than it is currently possible with the manual approach of GO slims.

Material The data used in this chapter come from the Gene Ontology website and from the NCBI Taxonomy database as of January 2007 (see Sec. 5.1).

5.2 Relations Between the GO and the TS

As argued in Ch. 4, phrases such as 'prokaryote-specific' are used in the Gene Ontology rather vaguely, with an underspecified meaning. Here, we propose to systematize the dependencies between GO terms and taxa by creating links with precise meanings, and to classify them into three types, designated as the *validity*, *specificity*, and *relevance* of a GO term with respect to a taxon.[1] The framework presented in this and the next sections is illustrated with an example in Fig. 5.1.

[1] The terms 'validity', 'specificity', and 'relevance' will be used extensively in this chapter. Alternative terminologies have been discussed with members of the Gene Ontology Consortium. The terms proposed here are only tentative and may be replaced by other terms in the future.

5.2.1 Validity

A term g in the Gene Ontology can be put in the relation of *validity* with a term t in the Taxonomy of Species if, and only if, the feature (function, process, or component) represented by g can be found in (can be attributed to) organisms of *all* species subsumed by the taxon represented by t. It can also be said that g is *valid* for the taxon represented by t. Note that the definition does not require that the feature can be found in *all* organisms of the involved species, but rather that it must be present in *some* organisms of *every* species subsumed by the specified taxon. It is possible that the feature is present only in organisms of a particular gender, at some time during their life, etc.

For example, the GO term suckling behaviour stands in the relation of validity with the taxonomic term Mammalia (mammals; alternatively, suckling behaviour is valid for the taxon *Mammalia*), since all mammals suckle, at some time during their life. Likewise, the term biological process is valid for the taxon *Viridiplantae* (green plants), because some biological processes take place in all green plants. Note that the fact that a GO term is valid for a particular taxon does not prevent it from being valid for other taxa as well; indeed, biological process is valid not only for green plants. On the other hand, the term conjugation is not valid for the taxon *Mammalia*, because there are species of mammals that do not conjugate. (Indeed, no mammals conjugate, in the sense of 'conjugation' adopted by the Gene Ontology.) Likewise, locomotion during locomotory behaviour is not valid for *Viridiplantae*, because there are species of green plants that never locomote[2] during locomotory behaviour. In fact, most green plants never locomote, but there seem to be plants that do locomote, at least in some sense: in the case of *Oxalis*, a plant's roots can pull the plant 60 cm around in the soil.[3]

[2]The Gene Ontology defines locomotion as a "self-propelled movement of a cell or organism from one location to another".

[3]See R. Koning, *Morphology and Anatomy*, Plant Physiology Information Website, http://koning.ecsu.ctstateu.edu/Plants_Human/MorpAnat.html.

5.2.2 Specificity

A GO term g is in the relation of *specificity* with a TS term t if, and only if, the feature represented by g can be found only in organisms of some of the species subsumed by the taxon represented by t. In other words, the feature cannot be found in an organism of any species outside of (not subsumed by) that taxon. It can also be said that g is *specific* to the taxon.

For example, the term suckling behaviour is specific to the taxon *Mammalia*, because no organisms other than mammals suckle.[4] The term is also trivially specific to the taxon *Metazoa*, because no organisms other than animals suckle. Likewise, apoplast is specific to *Viridiplantae*, because only plant cells have apoplasts (it seems). On the other hand, the term maternal behavior is not specific to the taxon *Mammalia*, because many animals other than mammals demonstrate maternal behaviour. Likewise, extracellular region, of which apoplast is a subterm, is not specific to *Viridiplantae*, because many organisms other than green plants have extracellular regions within their bodies.

5.2.3 Relevance

A GO term g is in the relation of *relevance* with a TS term t if, and only if, the feature represented by g can be found in organisms of some of the species subsumed by the respective taxon. The feature may be absent in organisms of some (but not all) species subsumed by that taxon; the feature may also be present in organisms of any species not subsumed by that taxon. Relevance may thus be seen as a 'weak' form of validity.[5] It can also be said that g is *relevant* for the taxon represented by t.

For example, the term hatching is relevant for the taxon *Mammalia*, because there are mammals that hatch — monotremes (*Monotremata*, an order-

[4] By definition: the Gene Ontology defines suckling as "specific actions of a newborn or infant mammal that result in the derivation of nourishment from the breast".

[5] 'Relevance' may not be the most relevant name for this relation; 'applicability' is another applicable term.

ranked taxon under *Mammalia*) lay eggs, and presumably hatch as well. Hatching is neither valid for nor specific to *Mammalia*, because mammals of some (most) mammal species do not hatch, and animals of some (many) non-mammal species do hatch. Likewise, cell wall is relevant for *Viridiplantae*, because some green plants have cells surrounded by a cell wall. Cell wall is not specific to *Viridiplantae*, since organisms of many non-plant species (e.g., fungi) have cell walls as well (albeit with a different structure). Furthermore, cell wall does not seem to be valid for *Viridiplantae* — plants of some species seem not to have cell walls.[6] On the other hand, the term cell wall is not relevant for the taxon *Mammalia*, and hatching is not relevant for *Viridiplantae*, for reasons obvious in both cases.

5.2.4 Additional Notes

With the Gene Ontology as one of their examples, Smith et al. [326] discuss the all-some pattern recommended for relations represented in an ontology. In the GO, the part of link between the terms cell wall and cell means that every instance of the type *cell wall* (i.e., every individual cell wall) is a part of an instance of the type *cell* (i.e., of an individual cell). This does not mean, however, that every individual cell has a cell wall as its part — which is not the case, in fact. Analogously, the link specific to between the GO term apoplast and the taxonomic term Viridiplantae means that every individual apoplast is found in (in this case, it is a component of a cell of) an individual green plant. It does not mean, however, that there are in every species of green plants some individuals in which apoplasts can be found (even if it were true).

Relations between features and organisms can be different in nature. For example, when GO terms come from the cellular component ontology, the relations may be partonomic, or partonomic-like: *part of*, *component of a cell of*, etc; other relations may hold between functions and organisms, and between processes and organisms. Relations between terms from the Gene

[6] Some algae, which are green plants, seem not to have cell walls; see, e.g., http://www.biologie.uni-hamburg.de/b-online/e26/26d.htm.

Ontology and terms from the Taxonomy of Species may be of practical importance for the developers of the GO: whenever one wants to make the taxonomical characteristic of a GO term explicit, the first step may be to state which taxa the term is valid, specific, or relevant for. This will then serve as an indication that there is some sort of 'found in' relation between the corresponding feature and the taxa, without the developer being forced to precisely define this relation in the very first place. The nature of such a relation may not yet be known precisely, and to correctly define ontologically valid relations is not a trivial task (Bittner [44], Smith and Grenon [328], Smith and Rosse [332], Johansson et al. [185]).

5.3 Inference Patterns — Rules of Propagation

The advantage of the GO-TS term-term relations introduced in Sec. 5.2 does not follow merely from the fact that their meaning is precisely defined. The idea is not simply to replace each manual annotation of a term as belonging to a slim with a manual annotation of that term as being valid, specific, or relevant to a taxon. In fact, these relations can be propagated along term-term links both within the Taxonomy of Species (along its is a links) and within the Gene Ontology (along its is a and part of[7] links). This leads to a substantial reduction in the amount of work necessary to manually associate GO terms with TS terms, and allows one to automatically discover certain types of inconsistency in manual assertions, and to form taxon-dependent views of the Gene Ontology on-demand. GO-TS term-term relations can be propagated along term-term links in both directions, i.e., both from a parent term to a child term, and from a child term to a parent term. We will speak of *down-propagation* in the former case, and of *up-propagation* in the latter case. There are thus six different patterns of propagation, two for each relation defined in Sec. 5.2. The patterns are introduced here only informally; a logical formalization is given in Ch. A.

Propagation of validity within the Gene Ontology Validity up-propagates along the links within the Gene Ontology: if a GO term g is valid for

[7] See Sec. 5.5 for more discussion of the 'part of' relation in this context.

a particular taxon t, then every ancestor of g is valid for t. For example, the term suckling behavior is valid for the taxon *Mammalia*, and so are all its ancestors, i.e., the terms behavioral interaction between organisms, interaction between organisms, etc. In practice, if there is a valid for link between suckling behaviour and Mammalia, then all terms of which suckling behaviour is a successor can be automatically included in a 'valid for *Mammalia*' view of the GO — even if none of those terms is explicitly marked as valid for *Mammalia*.

Propagation of validity within the Taxonomy of Species Validity down-propagates along the links within the Taxonomy of Species: if a GO term g is valid for a particular taxon t, then g is valid for every successor of t. For example, the term suckling behavior is valid for the taxon *Mammalia*, and thus it is valid for all taxa subsumed by *Mammalia*, i.e., *Prototheria*, *Theria*, *Eutheria*, etc. In practice, if there is a valid for link between suckling behaviour and Mammalia, then suckling behaviour can be automatically included in every 'valid for *T*' view of the GO where *T* stands for any taxon subsumed by *Mammalia* — even if suckling behaviour is not explicitly marked as valid for *T*.

Propagation of specificity within the Gene Ontology Specificity down-propagates along the links within the GO. For example, the term virion is specific to the taxon *Viruses*, and thus are also all its successors, i.e., the terms viral capsid, viral genome, etc. (The taxonomic status of viruses is not clear. NCBI Taxonomy contains a term for the (unranked) taxon *Viruses*, and thus we use it in the example, despite that, arguably, viruses are not organisms.)

Propagation of specificity within the Taxonomy Specificity up-propagates along the links within the TS. For example, the term suckling behavior is specific to the taxon *Mammalia*, and thus it is specific to all ancestors of this taxon, i.e., the taxa *Amniota*, *Tetrapoda*, etc.

Propagation of relevance within the Gene Ontology Relevance up-propagates along the GO. For example, the term hatching is relevant for *Mammalia*, and thus are all its ancestors, i.e., development and biological process.

Propagation of relevance within the Taxonomy Relevance up-propagates along the TS. For example, the term hatching is relevant for *Mammalia*, and thus it is relevant for all ancestors of this taxon, i.e., *Amniota*, *Tetrapoda*, etc.

5.3.1 Logical Properties of the Rules of Propagation

The rules above are sound — any inference from correct assumptions may lead only to correct conclusions (see Ch. A for more details). However, one may not be able to infer all correct assertions. For example, from the validity of suckling behavior for the taxon *Homo* we cannot infer, by propagation, that suckling behavior is valid for *Mammalia*. A separate rule may be added to the effect that if a GO term is valid for all taxa subsumed by a particular taxon, then the term is valid also for that taxon (note that such inference would be based on the closed world assumption; see Sec. 5.5 for further discussion). One may also want to design other rules.

Figure 5.1 illustrates the idea of propagation in a generic case. The asserted and inferred GO-TS relations are also shown in Tab. 5.1. Section 5.4 presents a simple example based on actual terms from the Gene Ontology and the Taxonomy of Species.

5.3.2 Consequences of Propagation

One practical importance of the rules of propagation is that one does not need to explicitly link *all* terms in the Gene Ontology with *all* the terms in the Taxonomy of Species which they are valid, specific, or relevant for. Rather, it suffices to relate some of them and use the rules to automatically propagate validity, specificity and relevance within both the Gene Ontology and the Taxonomy. The following observations show how to minimize the effort of manually associating GO terms with TS terms:

– To completely express validity for a particular taxon, only the most specific GO terms (i.e., the terms most distant from the root) which

are valid for that taxon need to be explicitly marked as such. Similarly, to express the validity of a particular GO term, it suffices to explicitly link the term with only the most general TS terms for which it is valid. For example, if the term suckling behaviour is explicitly valid for the term Mammalia, then it is not necessary to explicitly assert any such link between suckling behaviour (or any of its ancestors) and Mammalia (or any of its successors).

– To express the specificity to a particular taxon, only the most general GO terms specific to that taxon need to be explicitly marked as such. Similarly, to express the specificity of a particular GO term, it suffices to explicitly link the term with only the most specific (in the sense of their position within the hierarchy) TS terms to which the GO term is specific.

– To express relevance, only the most specific GO terms need to be explicitly linked with the most specific TS terms for which the GO terms are relevant.

Note also that validity implies relevance, but not vice versa. Likewise, specificity implies relevance as well, under the assumption that each GO term represents a feature that can be found in organisms of at least one species.

	t_1	t_2	t_3	t_4	t_5	t_6	t_7
g_1	R	V, R	—	V, R	V, R	—	—
g_3	$R, \overline{V}$	$\mathbf{V}, \mathbf{S}, R$	$\overline{V}, \overline{R}$	V, R	V, R	$\overline{V}, \overline{R}$	$\overline{V}, \overline{R}$
g_6	$S, R, \overline{V}$	S, R	$\overline{V}, \overline{R}$	—	—	$\overline{V}, \overline{R}$	$\overline{V}, \overline{R}$
g_7	$S, R, \overline{V}$	S, R	$\overline{V}, \overline{R}$	—	—	$\overline{V}, \overline{R}$	$\overline{V}, \overline{R}$

Table 5.1: Asserted and inferred GO-TS relations in the generic case presented in Fig. 5.1. All symbols as in that figure. GO terms for which no relations are asserted or inferred are omitted.

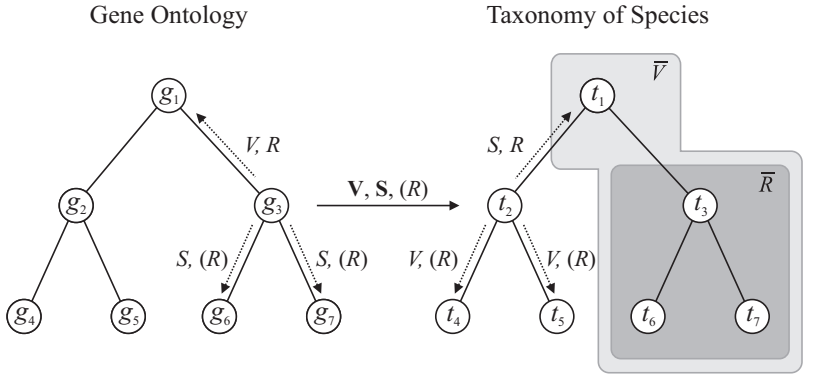

Figure 5.1: An abstract example of propagation of GO-TS term-term relations based on inference from manual assertions. GO terms, g_i, on the left; TS terms, t_i, on the right. V, validity, S, specificity, R, relevance in boldface (manually asserted) and in italics (inferred). Relevance implied from validity or specificity appears in parentheses. Dotted arrows symbolize the directions of inference. Validity and specificity between g_3 and t_2 are manually asserted. $\overline{V}$, inferred non-validity between GO terms specific to t_2 (i.e., g_3, g_6, and g_6) and those TS terms for which they cannot be valid (i.e., t_1, t_3, t_6, and t_7, light grey area). $\overline{R}$, inferred non-relevance between these GO terms and the TS terms for which they cannot be relevant (t_3, t_6, and t_7, dark grey area). See Sec. 5.5 for further explanation.

5.4 Dynamic Partitioning of the Gene Ontology — an Illustrative Example

GO-TS term-term relations can be used to generate taxon-dependent (e.g., species-specific) views of the Gene Ontology on demand. The idea is illustrated with a simple example involving three GO terms: cell wall, cell wall (sensu Magnoliophyta), and secondary cell wall, and three taxa: *Viridiplantae* (green plants, or simply plants, a kingdom), *Magnoliophyta* (flowering

plants, a division), and *Magnolia* (magnolias, a genus). Both the GO terms and the taxa are listed in the order of decreasing generality. It is reasonable to assume that (some) cells of all flowering plants have cell walls, and that cell walls as found in flowering plants can also be found in other plants — but not in organisms other than plants. Under this assumption, the GO term cell wall (sensu Magnoliophyta) can be explicitly linked with the the taxonomic term Magnoliophyta, acknowledging the former's validity for *Magnoliophyta*. Cell wall (sensu Magnoliophyta) can be also linked with Viridiplantae, acknowledging the specificity of the term to plants.

Table 5.2 shows the two explicit, manual links, and also all implicit links, inferred according to the rules of propagation. Note that if secondary cell wall is relevant (or valid) for *Magnoliophyta* (or for *Magnolia*), additional links must be added manually, as they cannot be inferred from those made earlier.

	Viridiplantae	Magnoliophyta	Magnolia
cell wall	*R*	*V*, *R*	*V*, *R*
cell wall (sensu Magnoliophyta)	**S**, *R*	**V**, *R*	*V*, *R*
secondary cell wall	*S*, *R*		

Table 5.2: Relations between three GO terms (rows) and three TS terms (columns). The generality of GO terms decreases from top to bottom, generality of TS terms decreases from left to right. Letters in table cells indicate associations: V — validity; S — specificity; R — relevance. Explicit (manually asserted) relations appear in boldface, inferred relations appear in italics. Validity propagates towards the top and right of the table, specificity propagates towards the bottom and left of the table, relevance propagates top-leftwards and also follows from validity and specificity.

It is now possible to constrain the GO to taxon-dependent views which contain not only those terms that were explicitly associated with the corresponding taxonomic term, but also those with inferred associations. One may want to select, say, those GO terms that are specific to green plants (in our example, the selection would include cell wall (sensu Magnoliophyta) and

secondary cell wall), or all terms valid for the taxon *Magnolia* (the selection would include the terms cell wall and cell wall (sensu Magnoliophyta)). It is also possible to make inverse queries, i.e., ask for taxa that correspond to some GO term-based criteria. For example, one may want to select the most inclusive (general) taxa for which the term cell wall is valid — the selection would include Magnoliophyta.

Furthermore, if, at a later time, a new GO term is added into the hierarchy, e.g., between cell wall and cell wall (sensu Magnoliophyta), its validity for flowering plants and magnolias would be inferred automatically, as would its relevance for all three taxa in the example — without the need for a manual update. An analogous observation can be made in the case when a new term is added to the hierarchy of the Taxonomy of Species. In the case of the Generic slim (see Fig. 4.2 again), if the terms cell communication and cell cycle were appropriately marked as valid for all (cellular) organisms, then it would be possible to automatically add to this slim the terms cellular process and cellular physiological process at the time they were inserted into the biological process branch of the Gene Ontology.

Observe that the hierarchy of taxonomic ranks can be used to constrain the view of the Taxonomy of Species in a manner analogous to the way the Taxonomy's terms can be used to constrain the view of the Gene Ontology. One may choose, for example, those GO terms that are valid for taxa which are of a rank not lower than *family* — thus building a 'family-level generic' GO slim. In this way one may build slims with the level of genericity (or specificity) precisely defined in terms of taxonomic criteria.

5.5 Discussion

Chapter 4 shows that GO slims, the current approach to building constrained views of the Gene Ontology, have a number of properties that may render them inconvenient as a means of subsetting the GO based on the relation of its terms to taxa. As an alternative, this chapter proposes a framework in which it suffices to manually assert some relations between terms from the

Gene Ontology and the Taxonomy of Species, while others can be inferred automatically. The framework attempts to solve a number of issues.

1. It enables one to link GO terms with taxa in an unambiguous way.
2. It reduces the effort needed to manually select all GO terms appropriate for a particular taxonomic context.
3. The framework provides a means for automatically creating views of the Gene Ontology based on criteria involving taxa that are not explicitly linked with GO terms.
4. The effort needed to maintain a consistent set of GO-TS term-term links can thus be reduced.

In the Gene Ontology, there are a number of terms which represent features that have not been found in organisms of some taxa (including features that *by definition* cannot be found in some organisms). These terms are an essential part of the GO: since the intention of its inventors was, among others, to provide a vocabulary for cross-annotation of entries in molecular biology databases, taxon-specific terms are unavoidable; differences in taxonomic coverage among various databases are well reflected in, e.g., the yearly database issue of *Nucleic Acids Research* (Galperin [121] and other articles in that volume). The current convention of the GO is to include any term that can apply to more than one taxonomic class of organisms.[8] However, until only recently, the only attempt to explicitly address the relations between terms in the Gene Ontology and terms in the Taxonomy has been the 'sensu' tagging of GO terms. In a newly published work, Schlicker et al. [310] show that queries such as "*Which biological processes are present in* Saccharomyces cerevisiae *but not in human?*" can be answered according to measures of semantic similarity between GO terms. However, while undeniably useful, this data-driven approach, based on existing annotations, may miss some of the knowledge which ontology curators already have, but which has not yet been explicitly accounted for in the annotation databases. Until the annotation data completely cover the whole of the Taxonomy

[8] See GO Editorial Style Guide, http://www.geneontology.org/GO.usage.shtml#sensu.

rather than a relatively small number of model organisms, it is desirable to be able to make assertions about the validity, specificity, and relevance of GO terms where they cannot be inferred from the data.

5.5.1 Implementation of the Framework

The framework sketched in this chapter (and further specified in App. A) has been informally discussed with members of the GO Consortium (including scientific curators), most recently during an online meeting on April 27., 2007. The following conclusions have been reached:

- It is important to address the problem of connecting the Gene Ontology (as well as other OBO ontologies, in particular ontologies of canonical and pathological anatomy) as soon as possible.
- The solution proposed here is a plausible candidate, although it may need further clarifications and modifications. The framework is not intended as a replacement, but rather as a partially overlapping alternative to GO slims. The two technologies are compatible: none of them excludes the other, and custom GO slims may of course be defined on both the current as well as a taxonomically enhanced version of the Gene Ontology.
- The framework is a candidate for implementation within OBO-Edit, the tool used by scientific curators of OBO ontologies (Day-Richter et al. [88]). There is also an ongoing discussion on use cases.

5.5.2 Manually Created and Inferred Assertions

For practical reasons, linking terms from the Gene Ontology with terms from the Taxonomy of Species may not be a trivial task. Ideally, for each GO term and each TS term (specifically, each species term) a curator would decide whether the feature represented by the former may be found in organisms of the taxon represented by the latter. However, to examine over 20,000 GO

terms and over 300,000 TS terms in this way would hardly be a plausible task, and automated support is therefore highly desirable. In the case of roughly 500 GO terms, hints are provided by the 'sensu' clauses.[9] Unfortunately, the meaning of such clauses is imprecise (as discussed earlier in Ch. 4), and only relevance could be inferred in this way.[10] In other cases, inspection of the lexical structure of terms and search for qualifiers such as 'viral', 'bacterial', 'fungal', 'microbial', may provide some more hints. Again, manual curation would be indispensable, since such qualifiers may be misleading. For example, bacterial binding: interacting selectively with any part of a bacterial cell is not necessarily specific to or valid, though clearly relevant for bacteria.

It is also possible to draw on the existing GO annotations of protein, sequence, and other data, since such annotations are typically species-specific. Yet another possibility is to employ text mining techniques to retrieve tentative links from scientific literature, in a manner similar to how suggestions for annotations of gene products with GO terms are found (see, e.g., Camon et al. [66] and Couto et al. [84]). Plans for such a study have been made in our local research group, though no practical work has been started yet.

5.5.3 Epistemological issues

In the field of biomedical ontology, it is not uncommon to confuse ontological claims with those of epistemological nature (Bodenreider et al. [51]). The framework proposed here is intended to capture claims about the relations that hold between types of features — cellular structures, molecular functions, and biological processes — on the one hand, and types of organisms on the other hand. As in the case of any other representational artifact, such ontological claims are made to reflect the best of one's knowledge, and are thus subject to further revisions, both due to the progress of science and

[9] While the number of GO terms with 'sensu' in the preferred name may have decreased due to the recent modifications, the previous names are kept as synonyms, and may still be used as hints.

[10] Under the — rather reasonable — assumption that 'as in, but not restricted to t' may be interpreted as claiming that the respective feature has been in fact been observed in organisms of the taxon t.

due to changes in the underlying portion of reality.

While the proposed solution does not include any mechanism for explicitly expressing epistemological claims, it is possible to add, e.g., evidence codes such as those used in the Gene Ontology Annotation Database (Camon et al. [65]). We could thus say, for example, that the assertion 'suckling behaviour is a feature of all mammals' was made by a curator, or that it was automatically inferred from other assertions or from annotation data. Epistemological claims may also be used to implement a more advanced ontology change-tracking system, such as the one proposed by Ceusters and Smith [70].

5.5.4 Logical Implications

In the discussion throughout the chapter, we have implicitly adopted the open world assumption (OWA): What is not explicitly stated, is assumed to be unknown rather than false. Under the complementary closed world assumption (CWA;[11] Reiter [296]), the lack of any explicit link between a particular GO term and any taxon would mean that the term is not relevant for any taxon — which would contradict the essential assumption that *every* GO term represents a feature found in *some* organisms.[12] Note that under the OWA it is not possible, with the relations defined above, to state that a GO term is *not* relevant or that it is *not* valid for a particular taxon.[13] Should such statements be desirable, explicit negation may be added, or the framework can be extended with 'negative' relations such as 'non-valid' (not valid, though possibly relevant), 'non-relevant' (not relevant, and thus also not valid), or 'non-specific' (not specific, though possibly relevant or even valid).

Another obvious logical consequence of the definitions of validity, relevance, and specificity is that it is possible to make contradictory assertions. For

[11] If neither p nor $\neg p$ can be proven from a knowledge base K, add $\neg p$ to K.

[12] The Gene Ontology does not represent features of unicorns — every feature represented there must have been observed in some organisms.

[13] Except for when the term can be asserted to be specific to another taxon disjoint with the one in question.

example, one may state both that a GO term g is specific to a taxon t_1 and that g is relevant for a taxon t_2, which would be a contradiction if neither t_1 nor t_2 are subsumed by the other. While this may seem a drawback, it is, in fact, a virtue: such contradictions can be detected by a reasoner, and either reported to the curator, or fixed according to some default rules. This property might also be used to detect contradictory annotations.

5.5.5 Propagation along 'Part of' Relations

The rules of propagation introduced above focus on the 'is a' relation between terms within the Gene Ontology (as well as between terms within the Taxonomy of Species). However, the GO contains also 'part of' relations between its terms. With the all-some quantification of type-level relations in the GO,[14] the patterns of propagation of validity, specificity, and relevance are the same in the case of 'part of' relations as those in the case of 'is a' relations. Specifically:

1. Validity up-propagates along 'part of': if a GO term g_1 is valid for a taxonomic term t, and there is another GO term, g_2, such that g_1 is part of g_2, then g_2 is valid for t. Let F_1, F_2, and O be the types represented by g_1, g_2, and t, respectively; then every instance of O has some instance of F_1 as a feature (assumed validity), every instance of F_1 is a part of some instance of F_2 (assumed part of), and thus every instance of O must have some instance of F_2 as a feature (inferred validity) — it is not possible for an instance of O to have as a feature an instance of F_1, but no instance of F_2.

2. Specificity down-propagates along 'part of': if a GO term g_1 is specific for a taxonomic term t, and there is another GO term, g_2, such that g_2 is part of g_1, then g_2 is specific for t. Let F_1, F_2, and O be as above (but now F_2 is part of F_1); then every instance of F_1 is a feature of some instance of O (assumed specificity), every instance of F_2 is a part of some instance of F_1 (assumed part of), and thus every instance of F_2

[14] $T_1 R T_2 \equiv$ *all* instances of T_1 are R-related to *some* instances of T_2.

must be a feature of some instance of O (inferred specificity) — it is not possible for an instance of F_2 to be a part of some instance of F_1 but *not* be a feature of some instance of O.

3. Relevance up-propagates along 'part of': if a GO term g_1 is relevant for a taxonomic term t, and there is another GO term, g_2, such that g_1 is part of g_2, then g_2 is relevant for t. (The reasoning behind this pattern is analogous to those above.)

5.5.6 A Note on the Terminology

The careful reader should have noted that our treatment of terms such as 'relation' and 'link' is somewhat relaxed. While it may add to the terminological inconsistence observed in the literature on ontological enginering (Kuśnierczyk [218]), it was desirable to simplify the text and avoid philosophical discussions here. For purity, the term 'relation' may be reserved for referring to that in which two or more entities stand to each other, and the term 'link' may be used only to denote a representational unit in an ontology used to represent a relation. See Smith et al. [331] for a more detailed discussion on related terminological issues.

In this context, *links* such as part of or is a between GO terms represent *relations* that hold between what the terms represent; the *links* valid for, specific to, and relevant for between GO terms and TS terms represent different *quantification characteristics* of relations such as *component of*, *function of*, etc. that hold between the types represented by the GO terms and the TS terms. For example, a valid for link between the GO term suckling behavior and the TS term Mammalia amounts to the ontological statement that *some* organisms of *all* mammal species suckle. It should also be noted that what 'found in' means is context dependent, and that precisely defining the scope of a context may not be trivial task. For even if viral capsids are not present in uninfected cells, they may be found in infected cells. However, viral components are not elements of canonical (normal, prototypical) cells, and it is the context of *physiological* cellular components, functions, and processes that we implicitly adopt here. With some effort, the framework may

be adopted to cover other contexts. See, e.g., Neuhaus and Smith [265] for a discussion and logical account of canonicity.

Chapter 6

Concluding Remarks

This chapter briefly summarizes the work done so far, revisits the goals and questions presented in Ch. 1, and discusses some possibilities for further research.

6.1 Goals and Questions Revisited

Phase I Section 1.1.4 outlines the goals, questions, methods, and contributions related to the research underlying this dissertation. Among the three major goals, goal G1 has been achieved to a large extent, though we are still far from having a complete understanding of the molecular basis of gastric acid secretion and pathogenesis of gastric and liver carcinoma. For example, one of our activities focused on *protein-coding* genes responding to activation of the the peroxisome proliferator-activated receptor α (PPAR-α); however, it has been recently shown that *miRNA-mediated* signalling is critical for PPAR-α agonist-induced liver proliferation and tumorigenesis (Shah et al. [313]). Clearly, much further work will have to be done in order to explain these and other interactions and regulatory pathways.

Related to G1 were three questions. In our publications we show that mi-

croarray gene expression experiments can provide substantial evidence for confirming or rejecting hypotheses about the functioning of organs in the gastrointestinal system (Q1). We have also experimented with various approaches to the design and analysis of microarray experiments (Q2); however, microarray experiments are typically very costly, and optimal designs (e.g., full factorial designs)[1] are seldom feasible in practice, especially with so many variables to be investigated. This also implies difficulties in analyzing the data, with further complications due to various sources of systematic bias and random noise. Quite often, we were (unfortunately) forced to make *ad hoc* decisions as to the filering and normalization of the data; there has been some progress made in this domain, but experts are far from an ultimate agreement. Multiple testing is another example of a problem yet to be fully addressed (see, e.g., Storey and Tibshirani [355, 356], and also Langaas et al. [230]).

Regarding question Q3, preliminary study aimed at selecting a variety of machine learning and data mining technologies has been done, but as the primary research interest of the group was to develop classifiers based on the rough set theory, no more detailed reviews have been made. In principle, any technique well-grounded in statistical theory should be appropriate for the purpose of classification of gene expression patterns. We have provided evidence that rough-set based classifiers are a reasonable choice (Q4); however, the sets of rules produced during learning were usually too large to be examined manually by an expert (several thousand rules in a typical experiment), and thus one of the benefits of using the rough set approach (classifiers in the form of human-readable rules) was not really achieved. It is possible, in general, to design sophisticated rule-pruning algorithms, yet in our studies performance was the most important feature of the classifiers.

If I were to continue this line of research in the future, I'd focus on using external sources of domain knowledge for the *selection of features* to be used for the purposes of classification, in addition to using them as sources of preliminary classifications, as it was done in our studies. Furthermore, while rough set-based classifiers seemed a reasonable alternative, any established data mining technique should, in principle, be good enough, although

[1] See, e.g., Cox and Reid [85, Chh. 5 and 6], Hinkelmann and Kempthorne [167, Chh. 7–13].

the high-dimensional low-sample nature of the data makes some techniques more appropriate than others.[2]

Phase II The goal G2 has not been, unfortunately, reached, and the project remained in the design phase, except for some initial prototype components. One of the reasons was that the available structured ('computer-readable') sources of general domain knowledge were (and still are) mostly at a rather preliminary and instable stage (Q5, Q9). While the Gene Ontology, for example, has been extensively used for the purpose of indexing data in online databases[3] and, more recently, for extensive annotation of free-text scientific publications,[4] the loose specification of relations between its terms and the lack of formal semantics for the underlying representation language[5] made it difficult to perform any sort of reasoning with the GO, except for (and even here with care) navigating from more specific to more general terms, and vice versa. It is plausible that, given the recent improvements in OBO ontologies, the project would enjoy more successful progress if turned back to life now. Furthermore, the network-building tools we were examining were mostly commercial; the recent development of the Cytoscape suite for visualizing biological networks[6] make it a good candidate for further research on knowledge-guided network construction. Other considerable sources of general domain knowledge include free-text publications, but text-mining of scientific literature is notoriously difficult due to a number of reasons.[7]

[2] Specifically, linear discriminant analysis (LDA) is an often used technique, with nearest shrunken centroids (also called 'predictive analysis of microarrays', PAM) and shrunken centroids regularized discriminant analysis (SCRDA) among the most popular variants. Recently, Tai and Pan [362] show how LDA can be augmented with prior knowledge extracted from the Gene Ontology.

[3] Indeed, the need for such consistent indexing was one of the motivations for the development of the GO.

[4] One of the most notable and successful examples of employing the GO and other ontologies and controlled vocabularies for indexing scientific literature (the 'bibliome') is the GOPubMed (http://www.gopubmed.org/), a knowledge-based search engine tuned for the exploration of biomedical literature.

[5] See, e.g., Köhler et al. [209], Kumar and Smith [215, 216], Smith and Kumar [330], Smith et al. [318, 329], and other.

[6] http://www.cytoscape.org/. Cytoscape was first publicly released in 2002, and the substantially improved version 2.0 was released in 2004.

[7] See, e.g., Sætre [303] for an overview.

Another obstacle met in the project was achieving task-specific, episodic knowledge necessary for the design of the case-based component (Q6, Q7). We have experienced the 'curse of dimensionality':[8] with few experts willing to make an effort to provide a detailed account of how they build biological association networks, and a large number of usually not-so-obvious clues used by them, it was difficult to build a representative library of cases and thus a robust representation of case-specific knowledge, with accurate measures of similarity and case retrieval criteria. An initial analysis of the available literature was of little help, since scientific publications usually include a presentation and discussion of the final results, rather than a description of the stepwise process of exploration and integration of publicly available and local experimental and other data. While knowledge-acquisition techniques more advanced than mere observation and discussion with the experts[9] might provide additional hints, no technique is likely to compensate for the lack of collaborating experts.

As discussed in Midelfart [253], Gunther et al. [152, 153], etc., the hierarchical structure of the GO — and thus dependencies between its terms — pose difficulties for testing the overexpression of genes annotated with specific GO terms (Q8), and likewise for GO annotation-based machine learning approaches for classificationn of gene expression profiles. With our eGOn tool, we showed that it was possible to use the Gene Ontology — not only its terms and annotations, but also its structure — to analyze gene expression data; however, it should be noted that the results of our tests are strongly dependent on the (imperfect) structure of the GO, and mistakes in the ontology may easily lead to incorrect results, despite sound testing procedures. Recently, a number of new methods for assessing the overrepresentation of specific GO terms in lists of genes have been proposed, e.g., the parent-child analysis (Grossmann et al. [142]), the ontology-based pattern identification

[8] A term introduced by Richard Bellman [36], relates (roughly) to the fact that in higher-dimensional spaces the number of points needed to sample the space with a density corresponding to that of sampling a lower-dimensional space grows exponentially in the number of dimensions. This means that one needs a huge number of observations to achieve a a reasonable accuracy in sampling a higher-dimensional space. The problem of dimensionality is particularly visible in microarray gene expression experiments, where there typically are tens of thousands of variables and only tens of samples.

[9] See, e.g., Awad and Ghaziri [20] for an introduction.

(Zhou et al. [395]), the categorization approach implemented in the Gene Ontology Categorizer (Joslyn et al. [188], Verspoor et al. [375]), and many others. The idea of using the structure of the GO for classificationn purposes is definitely of interest to a broad community; in addition, the recent developments in the GO and other biomedical ontologies — perhaps most significantly the introduction of multiple, well-defined relations (Smith et al. [326]) — require further studies on how such enhanced ontologies can be used in, e.g., the analysis of microarray gene expression data. This line of research is an interesting option for future my work.

Phase III Essentially, the main goal of this phase (G3) has been reached, although further improvements are certainly possible. I contributed to improvements in the structure and contents of a few OBO ontologies, and provided a detailed study and suggested a solution to the problem of connecting the GO to the Taxonomy of Species in a consistent, structured manner. As already mentioned earlier in this chapter, it appears that many of the biomedical ontologies, including those built by the OBO community, had not been built in a particularly consistent and coherent way. However, the situation seems to have been substantially improved; I have participated in discussions on a number of issues related to OBO ontologies, hopefully having made contributions to both adoption of the OBO Foundry principles and to further development of the principles themselves. While a few years ago the question Q10 would have to be answered rather negatively — biomedical ontologies hardly shared any design principles beyond that they consisted of hierarchically organized terms linked by a handful of vaguely defined relations, encoded in an underspecified format[10] — the current situation is much better. Not only are OBO ontologies based, for the most part, on common syntactic *and* semantic principles; rather than remaining separate, completely independent and unrelated artifacts, OBO ontologies are now

[10]Terminologies such as the SNOMED CT (http://www.ihtsdo.org/) and the multi-vocabulary UMLS (http://www.nlm.nih.gov/research/umls/) do have a richer structure than OBO ontologies in their initial form, yet they have hardly been considered ontologies — representations of the reality — and, despite having been formalized in a description logic (in the case of SNOMED), have been shown to suffer from a number of flaws that make reasoning based on these resources difficult and unreliable (Schulze-Kremer et al. [312], Kumar and Smith [215]).

being organized into a framework of tightly connected (in the form of so-called 'cross-products').[11]

Arguably, much of this unification and integration has been achieved following the adoption of BFO and the OBO Foundry ontology design principles (Q11). The Gene Ontology, for example, has been substantially revised to meet the requirements imposed by BFO, which could be seen as a proof of concept (Q12). But while using a common top-level ontology seems a reasonable way to go (at least, in a limited context such as the biomedical ontologies initiative), it is by no means clear that BFO is the best solution possible. For one thing, there exist a number of top-level ontologies, and the BFO team has not been able, despite considerable efforts, to convincingly show the superiority of the BFO approach. Furthermore, it appears that the community of biologists who develop ontologies in compliance with BFO do not do so without substantial criticisms towards it; it has been argued by some[12] that enforcing a wide adoption of BFO may in fact slow down the development of biomedical ontologies without any obvious benefit (beyond mere integration, see above) to counterweight the effect. It is not clear what modifications to BFO would make it more suitable as a top-level ontology for the biomedical community (Q14); it is also unclear how willing the BFO team is to give up on some of its philosophical opinions that shape BFO in order to better satisfy the needs of the domain experts.

It is clear that the way associations between terms in the GO (and, more generally, in OBO ontologies) and terms in the Taxonomy of Species are addressed is not well-structured and does not provide a good basis for inference (Q15); a detailed discussion of this issue is one of the contributions central to my dissertation. A plausible explanation for this is that the GO was initially intended to be a standalone resource rather than a part of a larger framework of cross-linked ontologies (Ashburner et al. [365]),[13] and

[11] See http://www.berkeleybop.org/obol/, the OBO cross-product website, and Mungall [260] for an introduction to Obol, the tool used to create cross-products.

[12] The discussion here is based on public and private communication on various online fora and mailing lists; currently, there are no publications criticizing the OBO Foundry approach, which can, arguably, be attributed to the fact that among the biologists who develop biomedical ontologies there is little understanding or interest for the philosophical issues central to BFO. The discussions typically invlove up to ten different people, including the BFO team.

[13] The cross-product initiative mentioned above is relatively recent, with Mungall [260] being one

explicit references to taxa, in the form of 'sensu' clauses, have been used to syntactically differentiate terms that have different meanings for different research communities. Only recently has it been recognized that the sparse, underspecified 'sensu' associations can be replaced by a more comprehensive and logically structured framework. I believe that the solution proposed in Chh. 5–A is a step towards such a framework (Q16); however, final evaluation has to be postponed until the first trial version of the GO including explicit, well-defined links to the Taxonomy of Species is constructed and subjected to critique and practical tests. Such experimental version is under development and is planned to be announced in June 2008.[14] Since the usefulness of the framework partially depends on the quality of the taxonomic resource, an exploration of how resources other than the NCBI Taxonomy Database (Q17) might or might not give better results is an interesting subject for future research.

of the earliest steps towards integration of OBO ontologies.

[14]Private communication with Jennifer Deegan (http://www.ebi.ac.uk/Information/Staff/).

Appendix A

A Logical Formalization of the GO-TS Framework

Chapter 5 explored how the relations between terms in the Gene Ontology and terms in the Taxonomy of Species can be used to select GO terms based on various taxon-related criteria. The framework described there informally provides basic inference mechanisms to propagate such assertions along the hierarchies of both the GO and the TS, and thus to reduce the effort needed for manual annotation and allow a user to query for biological features relevant for taxa even in the absence of appropriate explicit assertions. This appendix provides a logical formalization of the framework; its purpose is to provide an unambiguous semantics for the relations and rules of propagation, and thus provide a sound basis for an implementation of the framework.

The appendix is structured as follows:

- Section A.1 provides a concise recapitulation of the essential elements of the framework.
- In Sec. A.2 we introduce a very simple representation language that allows to declaratively define the semantics of the framework.

– Section A.3 provides formal definitions of the all-some, some-some, recursive some-some, and only-some patterns.

– Section A.5 provides further discussion.

A.1 Introduction

Terms in Gene Ontology represent features of organisms. Some of those features are commonly found in organisms of all taxa; others are specific to a particular taxon or group of taxa. Some features are found in all organisms of a particular taxon; others are found only in some organisms of that taxon. Based on these observations, the following patterns of organism-feature dependencies may be distinguished:

1. The *all-some* pattern. For example, cells are structures present in all vertebrates — *all* vertebrates have *some* cells as parts.

2. The *some-some* pattern. For example, wings are structures present in some vertebrates — *some* vertebrates have *some* wings as parts.

3. The *recursive some-some* pattern. For example, vertebrae are structures present in some organisms of every vertebrate species — *all* vertebrate species include *some* organisms that have *some* vertebrae as parts.

4. The *only-some* pattern. For example, plant cell walls are structures present only in plants — *only* plants have *some* plant cells walls as parts.

In a system operating under the open world assumption (OWA, Reiter [296]), there is also a need for negative assertions, i.e., assertions to the effect that it is *not* that all (some, only, etc.) organisms of a particular taxon have features of a particular type. Under OWA, the lack of a statement that some organisms of the type O have some features of the type F does not imply that no O's have F's; likewise, the lack of a statement that only O's have F's

as features does not imply that there in fact are any organisms that are not O's, yet have F's as features. Therefore, the following negative patterns may be useful as well:

1. The *negative all-some* pattern. For example, it is *not* the case that *all* formicidae (ants) have some wings as parts — some ants do not have wings, though some ants of all ant species have wings.

2. The *negative some-some* pattern. For example, it is *not* the case that *some* vertebrates have plant cell walls as parts — no vertebrate features a plant cell wall.

3. The *negative recursive some-some* pattern. For example, it is *not* the case that some organisms of *all* vertebrate species have wings as features — there are no wings to be found in any organism of some vertebrate species.

4. The *negative only-some* pattern. For example, it is *not* the case that cell walls are found *only* in plants — fungi and bacteria have cell walls as well (though not plant cell walls).

In many cases, the relation between organisms and features exemplifies more than one pattern: for some type O of organisms and some type F of features, it may be *both* that all O's have F's as features *and* that only O's have F's as features. Some of the patterns imply the others: *if* all O's have some F's, *then* some O's have some F's. This chapter provides a formal account of such rules (Sec. A.3).

A.2 $\mathcal{L}_{\mathrm{OF}}$ — A Formalism for the GO-TS Framework

To provide a formal account of the framework introduced informally in Ch. 5 and summarized in Sec. A.1, a simple knowledge representation language,

$\mathcal{L}_{OF}$,[1] is defined in this chapter. $\mathcal{L}_{OF}$ has limited expressivity, designed specifically for the current purposes.

A.2.1 The Vocabulary of $\mathcal{L}_{OF}$

The vocabulary V of $\mathcal{L}_{OF}$ contains two disjoint sets, the set V_G of GO terms and the set V_T of TS terms.

$$V = V_G \cup V_T, \tag{A.1}$$
$$V_G = \{g \mid g \text{ is a term in the Gene Ontology}\}, \tag{A.2}$$
$$V_T = \{t \mid t \text{ is a term in the Taxonomy of Species}\}. \tag{A.3}$$

A.2.2 The Syntax of $\mathcal{L}_{OF}$

The syntax of $\mathcal{L}_{OF}$ includes the following sets of valid sentences:

$$\Phi_{AS} = \{\phi \mid \phi \text{ is of the form } (\textsf{all-some } t\ g)\}, \tag{A.4}$$
$$\Phi_{SS} = \{\phi \mid \phi \text{ is of the form } (\textsf{some-some } t\ g)\}, \tag{A.5}$$
$$\Phi_{RSS} = \{\phi \mid \phi \text{ is of the form } (\textsf{recursively-some-some } t\ g)\}, \tag{A.6}$$
$$\Phi_{OS} = \{\phi \mid \phi \text{ is of the form } (\textsf{only-some } t\ g)\}, \tag{A.7}$$

where $t \in V_T$ and $g \in V_G$.

In addition, $\mathcal{L}_{OF}$ includes the set $\Phi_{\subseteq}$ of subsumption sentences:

$$\Phi_{\subseteq} = \{\phi \mid \phi \text{ is of the form } (\textsf{subsumed-by } c_1\ c_2)\}, \tag{A.8}$$

where c_1 and c_2 are such that either $c_1, c_2 \in V_T$ or $c_1, c_2 \in V_G$.

Negative Patterns $\mathcal{L}_{OF}$ may be extended to accommodate for negative patterns. Negation can be realized by explicitly negating sentences, e.g.,

$$\Phi_{NAS} = \{\phi \mid \phi \text{ is of the form } (\textsf{not } \phi_{AS})\},$$

[1] For the lack of a better name, we call the language '$\mathcal{L}_{OF}$', a language for expressing relations between classes of **o**rganisms and their **f**eatures.

where $\phi_{AS} \in \Phi_{AS}$, or by introducing sentence form with embedded negation, e.g.,

$$\Phi_{NAS} \quad = \quad \{\phi \mid \phi \text{ is of the form } (\text{not-all-some } t\ g)\}\,.$$

In this text, negative patterns will not be discussed in further detail.

Links between the GO and the TS versus sentences in $\mathcal{L}_{OF}$ The correspondence between the types of statements involving terms from the Gene Ontology and terms from the Taxonomy of Species, introduced previously in Ch. 5, and the sentence forms of $\mathcal{L}_{OF}$ is as follows:

- Validity, as defined in Ch. 5, corresponds to the recursive some-some pattern, expressable with $\mathcal{L}_{OF}$ sentences of the form Φ_{RSS}.
- Relevance, as defined in Ch. 5, corresponds to the some-some pattern, expressable with $\mathcal{L}_{OF}$ sentences of the form Φ_{SS}.
- Specificity, as defined in Ch. 5, corresponds to the only-some pattern, expressable with $\mathcal{L}_{OF}$ sentences of the form Φ_{OS}.

Sentences from Φ_{AS} do not have a corresponding pattern in the framework of Ch. 5.

A.2.3 The Semantics of $\mathcal{L}_{OF}$

The language $\mathcal{L}_{OF}$ is given a declarative, model-theoretic semantics. An interpretation $\mathcal{I}$ is a tuple $\langle U, I\rangle$:

$$\begin{aligned}
\mathcal{I} &= \langle U, I\rangle && \text{(A.9)}\\
U &= U_O \cup U_F \cup U_o \cup U_f && \text{(A.10)}\\
U_O &= \{O \mid O \text{ is a class of organisms}\} && \text{(A.11)}\\
U_F &= \{F \mid F \text{ is a class of features}\} && \text{(A.12)}\\
U_o &= \{o \mid o \text{ is an organism}\} && \text{(A.13)}\\
U_f &= \{f \mid f \text{ is a feature}\} && \text{(A.14)}
\end{aligned}$$

where the sets U_O, U_F, U_o, and U_f are disjoint:

$$U_O \cap U_F \cap U_o \cap U_f = \emptyset \tag{A.15}$$

I is a mapping from the symbols of $\mathcal{L}_{OF}$ to the elements of U:

$$I : V_T \rightarrow U_O \tag{A.16}$$

$$I : V_G \rightarrow U_F \tag{A.17}$$

Thus, for any symbol $s \in V_T \cup V_G$, $s^{\mathcal{I}} = I(s)$. For simplicity, $\mathcal{L}_{OF}$ is given an extensional semantics, in which classes[2] are treated as sets of objects. Alternatively, intensional semantics might be used; see, e.g., Hayes and Menzel [162] and Guha and Hayes [151]. Choosing extensional semantics allows to simplify the notation (no additional symbol for the extension function is needed) without interfering with the essential properties of the framework. (Identifying classes with their extensions — timeless sets of instances — as in the extensional semantics adopted here is acceptable if temporal issues can be ignored, e.g., if classes can be assumed to have the same instances at all times. If classes are assumed to gain and lose instances within time, intensional semantics is more appropriate.)

The interpretation of a sentence ϕ of the form $(\mathsf{subsumed\text{-}by}\ c_1\ c_2)$ (i.e., $\phi \in \Phi_{\subseteq}$) is defined in the usual way: ϕ is true if, and only if, the class represented by c_1 is a subclass of the class represented by c_2:

$$(\mathsf{subsumed\text{-}by}\ c_1\ c_2)^{\mathcal{I}} = \begin{cases} T & \text{if } \forall x \in U : x \in I(c_1) \Rightarrow x \in I(c_2), \\ F & \text{otherwise,} \end{cases} \tag{A.18}$$

with T and F being special symbols denoting the binary logical truth values in an obvious way.

For convenience, the symbol '$\triangleright$' will be used to denote the relation between an organism and a feature possessed by that organism. For example, '$o \triangleright f$' will mean that the organism o possesses the feature f.

[2] As discussed in Kuśnierczyk [218] and Smith et al. [331], the term 'class' is used in the literature on ontological engineering with various intentions; for the purposes of this article, 'class' and 'type' are treated as synonyms.

Definition A.2.1 (Semantics of Φ_{AS}-sentences) *Let $\mathfrak{I}$ be an interpretation of $\mathcal{L}_{\mathrm{OF}}$. Let ϕ be a $\mathcal{L}_{\mathrm{OF}}$ sentence of the form* $(\mathsf{all\text{-}some}\ t\ g)$*, with $t \in V_T$ and $g \in V_G$. The sentence ϕ is interpreted under $\mathfrak{I}$ as true if all instances of the class of organisms represented by the TS term t possess as a feature at least one instance of the class of features represented by the GO term g; otherwise, ϕ is interpreted as false:*

$$(\mathsf{all\text{-}some}\ t\ g)^{\mathfrak{I}} = \begin{cases} T & \text{if } \forall o \in I(t)\ \exists f \in I(g) : o \triangleright f, \\ F & \text{otherwise.} \end{cases} \tag{A.19}$$

The all-some pattern is by far the most common semantics of formulas expressing relations between classes in logic-based knowledge representation languages. For example, the $\mathcal{L}_{\mathrm{OF}}$ sentence of the form $(\mathsf{all\text{-}some}\ t\ g)$ is semantically equivalent to the following expression in description logics, given equivalent interpretation of t and g:

$$t \sqsubseteq \exists \mathsf{hasFeature}.g$$

where the symbol '$\mathsf{hasFeature}$' denotes the relation $\triangleright$ that holds between an organism and its features. (See Sec. 5.5.6 for a brief discussion of the corresponding term 'found in'.) Note that the description logic must allow for full existential quantification.[3]

Definition A.2.2 (Semantics of Φ_{SS}-sentences) *Let $\mathfrak{I}$ be an interpretation of $\mathcal{L}_{\mathrm{OF}}$. Let ϕ be a $\mathcal{L}_{\mathrm{OF}}$ sentence of the form* $(\mathsf{some\text{-}some}\ t\ g)$*, with t and g as in Def. A.2.1. The sentence ϕ is interpreted under $\mathfrak{I}$ as true if some instances of the class of organisms represented by the TS term t possess as a feature an instance of the class of features represented by the GO term g; otherwise, ϕ is interpreted as false:*

$$(\mathsf{some\text{-}some}\ t\ g)^{\mathfrak{I}} = \begin{cases} T & \text{if } \exists o \in I(t)\ \exists f \in I(g) : o \triangleright f, \\ F & \text{otherwise.} \end{cases} \tag{A.20}$$

[3] Full existential quantification is a concept description of the form $\exists R.C$, where R is a role and C is a concept description. A description logic with full existential quantification is marked by the letter '$\mathcal{E}$' in the name (Baader and Nutt [23]).

Unlike the all-some pattern, the some-some pattern involves an existential claim: there *is* at least one instance of $t^{\mathcal{I}}$ (an organism of the taxon $O = I(t)$ denoted by t), and at least one instance of $t^{\mathcal{I}}$ is such that it has as a feature an instance of $g^{\mathcal{I}}$ (an instance of the class $F = I(g)$ of features denoted by g). The ontological consequences of this semantics are discussed in Sec. A.5.1. In description logic-based formalisms, existential claims about individuals are made only by asserting, in the ABox,[4] the existence of a particular instance of a class. (Thus, one cannot say that there is an instance of a particular class without actually naming the instance.)

Definition A.2.3 (Semantics of Φ_{RSS}-sentences) *Let $\mathcal{I}$ be an interpretation of $\mathcal{L}_{OF}$. Let ϕ be a $\mathcal{L}_{OF}$ sentence of the form* (recursive-some-some t g)*, with* t *and* g *as in Def. A.2.1. The sentence ϕ is interpreted under $\mathcal{I}$ as true if every class of the rank species or above subsumed by the class of organisms represented by the TS term* t *includes at least one instance which possesses as a feature an instance of the class of features represented by the GO term* g*; otherwise, ϕ is interpreted as false:*

$$(\text{recursive-some-some t g})^{\mathcal{I}} = \begin{cases} T & \text{if } \forall O \in U_O : \\ & \quad (O \subseteq I(t) \ \wedge \ c(O)) \Rightarrow \\ & \quad \exists o \in O \ \exists f \in I(g) : o \triangleright f, \\ F & \text{otherwise,} \end{cases} \tag{A.21}$$

where O *ranges over classes of organisms in* U_O *and* c*, the* condition of recursion*, is a boolean function* $c : U_O \rightarrow \{T, F\}$ *such that* $c(O) = T$ *iff* O *is of the taxonomic rank* species *or above.*

This pattern is called 'recursive some-some' because of its recursive nature: a sentence of the form Φ_{RSS} implies any other sentence of this form in which the TS term t is replaced by another TS term t' such that both (subsumed-by t' t) and $c(I(t')) = T$ are true. This is obviously not the case for sentences of the form Φ_{SS}.

[4] An ABox is the *assertional* component of a description logic system, where individuals are described, as opposed to a TBox, the *theory* component of a DL system, where classes (known as 'concepts' in DLs) are described (Nardi and Brachman [262], Nebel [263]).

Similarly to the some-some pattern Φ_{SS}, the recursive some-some pattern Φ_{RSS} is based on an existential claim.[5] The condition c is essential for the semantics to correspond to the definition of validity given in Ch. 5. In the general case, c may be replaced with any other condition that circumscribes the set of classes in U_O to be considered. If the most specific classes for which c is true are singleton classes (one-instance classes), then, effectively, the recursive some-some pattern is equivalent to the Φ_{AS} pattern. In the case of the Taxonomy of Species, however, the most specific classes (taxa of the rank *subspecies*, in the case of the zoological part of the TS, or of the rank *subforma*, in the case of the botanical part of the TS) are not singletons, and thus can all be considered without the consequence of equating Φ_{RSS}-formulas with Φ_{AS} formulas. That is, c can be chosen so as to be true for any class covered by the Taxonomy of Species, not only those that are of the rank *species* or above. The reason for choosing the rank *species* for the condition c is explained further in Sec. A.5.2.

As in the case of the some-some pattern, Φ_{RSS} formulas cannot be translated to a conventional DL language. Here, we need to quantify over classes; class terms (corresponding, roughly speaking, to what is called in DLs 'concepts') play in DLs the role of predicates, and these languages typically do not allow for such higher-order quantification.

Definition A.2.4 (Semantics of Φ_{OS}-sentences) *Let $\mathcal{I}$ be an interpretation of $\mathcal{L}_{OF}$. Let ϕ be a $\mathcal{L}_{OF}$ sentence of the form* (only-some t g)*, with* t *and* g *as in Def. A.2.1. The sentence ϕ is interpreted under $\mathcal{I}$ as true if only instances of the class of organisms represented by the TS term* t *can possess as a feature an instance of the class of features represented by the GO term* g*; otherwise, ϕ is false:*

$$(\text{only-some t g})^{\mathcal{I}} = \begin{cases} \text{T} & \text{if } \forall f \in I(g)\ \forall o \in U_o : o \triangleright f \Rightarrow o \in I(t), \\ \text{F} & \text{otherwise,} \end{cases} \tag{A.22}$$

where o *ranges over all organisms in U_o.*

[5] This is only partially true. The existential claim $\exists o \in O$ in the semantics of Φ_{RSS} is conditioned on there being a subclass of $t^{\mathcal{I}}$ for which the condition c holds. A sentence of the form (recursive-some-some t g) is trivially true when t is such that $\exists O \in U_O(O \subseteq I(t) \wedge c(O))$ is false. If desirable, this can be repaired by adding $\exists O \in U_O(O \subseteq I(t) \wedge c(O))$ as a condition in the definition of Φ_{RSS}.

The only-some pattern is approximately the inverse of the all-some pattern. However, it is not exactly the case: a sentence of the form $(\mathsf{only\text{-}some}\ \mathrm{t}\ \mathrm{g})$ demands only that *if* something is an instance of the class $\mathrm{g}^{\mathcal{I}}$ *and* it is a feature of an organism, *then* the organism is of the class $\mathrm{t}^{\mathcal{I}}$. On the other hand, a sentence of the form $(\mathsf{all\text{-}some}\ \mathrm{g}\ \mathrm{t})$, if allowed in $\mathcal{L}_{\mathrm{OF}}$ (with $\triangleright$ replaced by $\triangleleft$ in the semantics, the symbol '$\triangleleft$' denoting the inverse of the relation denoted by '$\triangleright$'), would claim that *every* instance of the class $\mathrm{g}^{\mathcal{I}}$ is actually possessed by an instance of $\mathrm{t}^{\mathcal{I}}$. This non-equivalence is also clearly seen in the translation of $(\mathsf{only\text{-}some}\ \mathrm{t}\ \mathrm{g})$ into the corresponding DL formula:

$$\mathrm{g} \sqsubseteq \forall \mathsf{hasFeature}^{-}.\mathrm{t}\ ,$$

which is not equivalent to $\mathrm{t} \sqsubseteq \exists \mathsf{hasFeature}.\mathrm{g}$ (where $\mathsf{hasFeature}^{-}$ is the inverse of $\mathsf{hasFeature}$). Again, this is related to the problem of existential claims, mentioned earlier in this section, and discussed further in Sec. A.5.1.

A.3 Monotonic Inference in $\mathcal{L}_{\mathrm{OF}}$

The purpose of the framework for asserting relations between classes of organisms and classes of their features, as introduced in Ch. 5, is to allow one to create taxonomically specified partitions of the GO on demand. This section describes a number of rules that can be used to perform such inferences in $\mathcal{L}_{\mathrm{OF}}$. It can be proven that the rules are sound — the relation of logical consequence between the premises and the conclusion of each rule is that of entailment, i.e., the inference rules are monotonic. The proofs are trivial, and are not given in this text.

A.3.1 Inference from Φ_{AS}-Sentences

The following rules specify inference patterns involving sentences of the form Φ_{AS} as premises.

Theorem A.3.1 (Down-propagation of Φ_{AS} within the TS) *Let* $\mathrm{g} \in \mathrm{V_G}$ *be a term in the Gene Ontology and* $\mathrm{t} \in \mathrm{V_t}$ *be a term in the Taxonomy of Species,*

and $\mathcal{I}$ *be a* $\mathcal{L}_{\text{OF}}$*-interpretation such that* $(\textsf{all-some t g})^{\mathcal{I}} = \text{T}$. *Then for every* $\text{t}' \in \text{V}_\text{T}$ *such that* $(\textsf{subsumed-by t' t})^{\mathcal{I}} = \text{T}$, $(\textsf{all-some t' g})$ *is also true under* $\mathcal{I}$*:*

$$\frac{(\textsf{all-some t g}),\ (\textsf{subsumed-by t' t})}{(\textsf{all-some t' g})}\text{R}_{\text{AS|T}} \tag{A.23}$$

The rule $\text{R}_{\text{AS|T}}$ says that the relation $\triangleright$ ('has as a feature') propagates with the all-some semantics *downwards* along the hierarchy of the Taxonomy of Species (i.e., it is inherited by more specific taxa from more general taxa).

Theorem A.3.2 (Up-propagation of Φ_{AS} within the the GO) *Let* t, g, *and* $\mathcal{I}$ *be as in Theorem A.3.1. Then for every* $\text{g}' \in \text{V}_\text{G}$, $(\textsf{subsumed-by g g'})$ *implies* $(\textsf{all-some t g'})$ *under* $\mathcal{I}$*:*

$$\frac{(\textsf{all-some t g}),\ (\textsf{subsumed-by g g'})}{(\textsf{all-some t g'})}\text{R}_{\text{AS|G}} \tag{A.24}$$

The rule $\text{R}_{\text{AS|G}}$ says that the relation $\triangleright$ propagates with the all-some semantics *upwards* along the hierarchy of the Gene Ontology (i.e., it is inherited by more general classes from more specific classes).

A.3.2 Inference from Φ_{SS}-Sentences

The following rules specify inference patterns involving sentences of the form Φ_{SS} as premises.

Theorem A.3.3 (Up-propagation of Φ_{SS} within the TS) *Let* $\text{g} \in \text{V}_\text{G}$, $\text{t} \in \text{V}_\text{T}$, *and* $\mathcal{I}$ *be such that* $(\textsf{some-some t g})^{\mathcal{I}} = \text{T}$. *Then for every* $\text{t}' \in \text{V}_\text{T}$, $(\textsf{subsumed-by t t'})$ *implies* $(\textsf{some-some t' g})$ *under* $\mathcal{I}$*:*

$$\frac{(\textsf{some-some t g}),\ (\textsf{subsumed-by t t'})}{(\textsf{some-some t' g})}\text{R}_{\text{SS|T}} \tag{A.25}$$

The rule $\text{R}_{\text{SS|T}}$ says that the relation $\triangleright$ propagates with the some-some semantics *upwards* along the hierarchy of the Taxonomy of Species.

Theorem A.3.4 (Up-propagation of Φ_{SS} within the GO) *Let* g, t, *and* $\mathfrak{I}$ *be as in Theorem A.3.3. Then for every* $g' \in V_G$, $(\text{subsumed-by } g\ g')$ *implies* $(\text{some-some } t\ g')$ *under* $\mathfrak{I}$*:*

$$\frac{(\text{some-some } t\ g),\ (\text{subsumed-by } g\ g')}{(\text{some-some } t\ g')} R_{SS|G} \tag{A.26}$$

The rule $R_{SS|G}$ says that the relation $\triangleright$ propagates with the some-some semantics *upwards* along the hierarchy of the Gene Ontology.

A.3.3 Inference from Φ_{RSS}-Sentences

The following rules specify inference patterns involving sentences of the form Φ_{RSS} as premises.

Theorem A.3.5 (Down-propagation of Φ_{RSS} within the TS) *Let* $g \in V_G$, $t \in V_T$, *and* $\mathfrak{I}$ *be such that* $(\text{recursive-some-some } t\ g)^{\mathfrak{I}} = T$. *Then for any* $t' \in V_T$ *such that* $c(t^{\mathfrak{I}}) = T$, $(\text{subsumed-by } t'\ t)$ *implies* $(\text{recursive-some-some } t'\ g)$ *under* $\mathfrak{I}$*:*

$$\frac{(\text{recursive-some-some } t\ g),\ (\text{subsumed-by } t'\ t), c(t'^{\mathfrak{I}})}{(\text{recursive-some-some } t'\ g)} R_{RSS|T} \tag{A.27}$$

The rule $R_{RSS|T}$ says that the relation $\triangleright$ propagates with the recursive some-some semantics *downwards* along the hierarchy of the Taxonomy of Species — provided that the recursion condition c is fulfilled. Note that the rule is not purely syntact, as it explicitly refers to the semantics: it requires that c holds for $t^{\mathfrak{I}}$. This inconvenience can be avoided if the requirement $c(t'^{\mathfrak{I}})$ is converted into a syntactic statement, which can be easily achieved if the universe of $\mathcal{L}_{OF}$ is extended with taxonomic ranks (kingdom, class, order, etc.), and the vocabulary of $\mathcal{L}_{OF}$ is extended with the corresponding rank-

terms (kingdom, class, order, etc.:

$$V = V_G \cup V_T \cup V_r \tag{A.28}$$
$$V_r = \{\text{kingdom}, \text{class}, \text{order}, \ldots\} \tag{A.29}$$
$$U = U_O \cup U_T \cup U_o \cup U_t \cup U_r \tag{A.30}$$
$$U_r = \{\text{kingdom}, \text{class}, \text{order}, \ldots\} \tag{A.31}$$

A new syntactic form is needed for asserting the ranks of taxa; for example, $(\text{rank-of } r\ t)$, where $r \in V_r$ and t is as before, with the following semantics:

$$(\text{rank-of } r\ t)^{\mathfrak{I}} = \begin{cases} T & \text{if } t^{\mathfrak{I}} \text{ is of the rank } r^{\mathfrak{I}}, \\ F & \text{otherwise.} \end{cases} \tag{A.32}$$

Using this notation, (A.27) can be rewritten by replacing $c(t'^{\mathfrak{I}})$ with

$$((\text{rank-of species } t') \text{ or } (\text{rank-of genus } t') \text{ or} \ldots)$$

that is, a disjunction of all sentences of the form $(\text{rank-of } r\ t')$, where $r \in V_r$ is such that $c(t'^{\mathfrak{I}})$ is true if $(\text{rank-of } r\ t')$, i.e., and where species, genus, $\ldots \in V_r$ are individual terms representing taxonomic ranks above a certain level (e.g., species), as demanded by the condition c. (See App. B for more details on taxa and ranks.)

Theorem A.3.6 (Up-propagation of Φ_{RSS} within the GO) *Let* t, g, *and* $\mathfrak{I}$ *be as in Theorem A.3.5. Then for any* $g' \in V_G$, $(\text{subsumed-by } g\ g')$ *implies* $(\text{recursive-some-some } t\ g')$ *under* $\mathfrak{I}$*:*

$$\frac{(\text{recursive-some-some } t\ g),\ (\text{subsumed-by } g\ g')}{(\text{recursive-some-some } t\ g')} R_{\text{RSS|G}} \tag{A.33}$$

The rule $R_{\text{RSS|G}}$ says that the relation $\triangleright$ propagates with the recursive some-some semantics *upwards* along the hierarchy of the Gene Ontology. Note that $\Phi_{\text{RSS|G}}$ does not invoke the condition c; this is because c applies to taxa, and in $\Phi_{\text{RSS|G}}$ the condition and the consequence refer to the same taxon.

A.3.4 Inference from Φ_{OS}-Sentences

The following rules specify inference patterns involving sentences of the form Φ_{OS} as premises.

Theorem A.3.7 (Up-propagation of Φ_{OS} within the TS) *Let* $g \in V_G$, $t \in V_T$, *and* $\mathcal{I}$ *be such that* $(\text{only-some } t\ g)$ *is true under* $\mathcal{I}$. *Then for every* $t' \in V_T$, $(\text{subsumed-by } t\ t')$ *implies* $(\text{only-some } t'\ g)$ *under* $\mathcal{I}$:

$$\frac{(\text{only-some } t\ g),\ (\text{subsumed-by } t\ t')}{(\text{only-some } t'\ g)} R_{OS|T} \tag{A.34}$$

The rule $R_{OS|T}$ says that the relation $\triangleright$ propagates with the only-some semantics *upwards* along the hierarchy of the Taxonomy of Species.

Theorem A.3.8 (Down-propagation of Φ_{OS} within the GO) *Let* g, t, *and* $\mathcal{I}$ *be as in Theorem A.3.7. Then for every* $g' \in V_G$, $(\text{subsumed-by } g'\ g)$ *implies* $(\text{only-some } t\ g')$ *under* $\mathcal{I}$:

$$\frac{(\text{only-some } t\ g),\ (\text{subsumed-by } g'\ g)}{(\text{only-some } t\ g')} R_{OS|G} \tag{A.35}$$

The rule $R_{OS|G}$ says that the relation $\triangleright$ propagates with the only-some semantics *downwards* along the hierarchy of the Gene Ontology.

A.3.5 Φ_{SS}-Sentences versus Φ_{RSS}, Φ_{AS}, and Φ_{OS}-Sentences

In addition to the rules specified earlier in this section, it is also desirable to specify rules that would allow to infer assertions of one form from assertions of a different form.

Theorem A.3.9 (Φ_{RSS} implies Φ_{SS}) *Let* $g \in V_G$ *and* $t \in V_T$, *and* $\mathcal{I}$ *be such that* $(\text{recursive-some-some } t\ g)^{\mathcal{I}} = \top$. *Then* $(\text{some-some } t\ g)$ *is true under* $\mathcal{I}$:

$$\frac{(\text{recursive-some-some } t\ g)}{(\text{some-some } t\ g)} R_{RSS \to SS} \tag{A.36}$$

The rule $R_{RSS\rightarrow SS}$ says that the recursive some-some pattern implies the some-some pattern. Similarly, the rules $R_{AS\rightarrow SS}$ and $R_{OS\rightarrow SS}$ say that the some-some pattern can be inferred from the all-some and the only-some patterns, respectively:

$$\frac{(\text{all-some t g})}{(\text{some-some t g})}R_{AS\rightarrow SS} \tag{A.37}$$

$$\frac{(\text{only-some t g})}{(\text{some-some t g})}R_{OS\rightarrow SS} \tag{A.38}$$

Note, however, the subtle issue that sentences of the form Φ_{SS} involve existential claims, while sentences of the other forms do not. For the rules $\Phi_{AS\rightarrow SS}$ and $\Phi_{OS\rightarrow SS}$ to be valid, the definitions A.2.1 and A.2.4 would have to be modified accordingly, to involve such existential claims as well. This issue is further discussed in Sec. A.5.1.

A.4 Non-Monotonic Inference in $\mathcal{L}_{OF}$

The rules specified above allow for monotonic (deductive) inference about the relations between classes of organisms and classes of features. However, it might be useful to include non-monotonic reasoning as well.[6] For example, one might want to interpret sentences of the form Φ_{SS} as expressing some sort of *prototypicality* or *defaultness*. One could thus perform *downward* propagation of the some-some pattern along the hierarchy of the Taxonomy of Species — in addition to the monotonic *upward* propagation.

Furthermore, with an appropriate extension to $\mathcal{L}_{OF}$ allowing one to speak about instances (individual organisms and features) it would be possible to make both monotonic and non-monotonic inferences about individuals. An assertion of the form (some-some t g) could thus be used to defeasibly infer that, given a particular organism from the taxon denoted by t, the organism does have a feature of the type denoted by g, while the analogous inference from an assertion of the form Φ_{AS} would be non-defeasible — that

[6] See, e.g., Antonelli [16], Koons [211], and Jaszczolt [182] for a brief but informative introduction to non-monotonic logic, defeasible inference, and default reasoning.

is, if o is an organism from the taxon denoted by t, then from the assertion (all-some t g) it follows *deductively* that o has a feature of the type denoted by g. Note that the definition of the some-some pattern given in Sec. A.2.3 does not refer to prototypicality, and that defeasible inference is not directly supported by the formalism. Other extensions, e.g., involving numerical quantitation of prototypicality allowing for reasoning with various degrees of support or certainty, are of course possible.

A.5 Discussion

This chapter provides a logical specification of the framework informally introduced earlier in Ch. 5. It defines patterns that can be used to assert relations between classes of organisms and classes of their features, by meaningfully connecting terms in the Gene Ontology with terms in the Taxonomy of Species. It also specifies a number of deductive inference patterns, and explains that patterns for non-deductive inference can be defined in addition to what is provided by the framework.

A.5.1 Existential Claims

In Sec. A.2.3, the Definition A.2.2 introduces semantics for sentences that make existential claims: a sentence of the form (some-some t g) says that there *exists* such an organism of the type $t^{\mathcal{I}}$ that has as a feature of the type $g^{\mathcal{I}}$. The some-some and the recursive some-some patterns make existential claims, while the all-some and the only-some patterns do not. One explanation for this could be that sentences of the latter form are *definitional*, while those of the former are *descriptional*. That is, definitional sentences say that for an organism to be classified as an instance of a particular taxon, it must have some specific features. Conversely, the descriptive sentences do not say that if an organism is classified as an instance of a particular taxon, then it must have some specific features — what they say is simply that some organisms of the taxon happen to have such features. Furthermore, it is

reasonable to read 'exists' as *has existed*, or perhaps as *has been observed* or *there is evidence for that there existed*, which allows the framework to be used to make assertions about extinct species. In any case, the exact meaning of 'exists' has to be defined carefully and precisely.

To avoid the incoherence between sentences involving existential claims and those that do not involve such claims, two general solutions can be considered.

1. Avoid the existential claim by conditioning the semantics of existential sentences (sentences of the form Φ_{SS}) on the existence of the purported instances. For example, consider this modified definition of the semantics of Φ_{SS}-sentences:

$$\left(\textsf{some-some t g}\right)^{\mathcal{I}} = \begin{cases} \mathrm{T} & \text{if } \exists o \in I(t) \Rightarrow \exists f \in I(g) : o \triangleright f, \\ \mathrm{F} & \text{otherwise,} \end{cases} \tag{A.39}$$

   Analogously, the (conditioned) existential claim in the semantics of Φ_{RSS}-sentences can be further conditioned as follows:

$$\left(\textsf{recursive-some-some t g}\right)^{\mathcal{I}} = \begin{cases} \mathrm{T} & \text{if } \forall O \in \mathcal{U}_O : \\ & \quad (O \subseteq I(t) \ \wedge \ c(O)) \Rightarrow \\ & \quad (\exists o \in O \ \Rightarrow \exists f \in I(g) : o \triangleright f), \\ \mathrm{F} & \text{otherwise,} \end{cases} \tag{A.40}$$

2. Alternatively, extend the existential claim to hold for non-existential sentences as well. For example, consider this modified definition of the semantics of Φ_{AS}-sentences:

$$\left(\textsf{all-some t g}\right)^{\mathcal{I}} = \begin{cases} \mathrm{T} & \text{if } \exists o \in I(t) \wedge \\ & \quad \forall o \in I(t) \ \exists f \in I(g) : o \triangleright f \\ \mathrm{F} & \text{otherwise,} \end{cases} \tag{A.41}$$

Existential claims made at the level of concepts (class descriptions) are rather uncommon in description logics-based KR formalisms, and such claims are usually made by asserting the existence of particular instances. That is, existential claims are made not in TBoxes, but rather in ABoxes (Baader

and Nutt [23] Baader et al. [22]). However, in the context of the development of biomedical ontologies which are intended to speak of the real world and the organisms which exist, or have existed in it, existential claims are essential, and they are actually included in the ontologies, albeit not necessarily explicitly. While the model-theoretic semantics of statements such as (part-of apoptosome cytosol) is that all apoptosomes are parts of cytosols, though there need not be any apoptosomes at all to make the statement true, the implicit assumption is that there actually are (or at least have been) apoptosomes out there. That is, the inclusion of a term in a biomedical ontology automatically implies the existence (whether present or past) of instances of the corresponding type.

A.5.2 The Choice of *species*

In Sec. A.2.3, the recursion condition c on taxa needed for defining the recursive some-some pattern Φ_{RSS} was based on the rank *species*. Why species? What is so special about species for it to be chosen as the basis for the condition c? There are two reasons for which this choice has been made:

1. Species are not just like any other taxa, in that a species is typically defined as a population of organisms that are capable of interbreeding with each other and have fertile offspring; taxa of higher ranks cover organisms that are not necessarily capable to interbreed, and taxa of subspecies ranks cover organisms that are capable of interbreeding with organisms outside of those taxa. Of course, this is just one possible way of understanding the notion of species; there are over a dozen different so-called 'biological species concepts' (which amounts to the so-called 'species-pluralism'), and there are also different ontological views on what species are (individuals, sets, types). See, e.g., Ghiselin [126], Mayden [247], or Ereshefsky [102] for further details. Nevertheless, however defined, species have long been in the center of interest of biologists and philosophers alike.

2. Species are explicitly referred to by the Gene Ontology: its documentation speaks fo 'species-specificity', 'species-specific terms', etc.

The sentence 'all birds lay eggs', for example, should (usually) not be taken literally as meaning that every individual bird lays eggs, but rather as meaning that in every species of birds there are individuals who lay eggs. But to explain the meaning as in the latter case, one has to be able to refer to species. If desirable, other ranks may be used to specify the recursion condition, as in Sec. A.3.3.

A.5.3 Translation to Other Formalisms

As discussed earlier in this chapter (Sec. A.2.3), it may not be straightforward to translate some of the forms in $\mathcal{L}_{\mathrm{OF}}$ into a description logic-based representation formalism. However, more expressive formalisms may be used, such as IKL, a recently proposed, "extremely expressive" logical formalism designed for interchange and archiving of information in a network of logic-based reasoners.[7] For example, the recursive some-some pattern (recursive-some-some t g) (validity) can be defined in IKL as in Fig. A.1.

```
(iff (valid-for go-term tax-term)
     (forall ((taxon t) (organism o))
             (if (and (subsumed-by t (tnb tax-term))
                      (has-rank t species))
                 (exists ((feature f))
                         (and (instance-of f (tnb go-term))
                              (has-feature o f))))))
```

Figure A.1: Definition of validity expressed in IKL. taxon and organisms denote the class of taxa and the class of organisms, respectively; species denotes the rank *species*. tnb is a dereference function ('thing named by').

[7] http://www.ihmc.us/users/phayes/IKL/GUIDE/GUIDE.html.

Appendix B

The Taxonomy of Species and Taxonomic Databases — A Critical Assessment

Since its conception by Carl Linnaeus some 250 years ago, the Linnaean Taxonomy of Species (TS) has been one of the most important inventions designed with the intention of systematizing knowledge about life and its forms.[1] And although it doesn't currently seem to be a hot topic in general biology-related journals, it has always been used as a source of reference and a standard vocabulary for indexing information about biological systems. Chapters 4, 5, and App. A discuss various aspects of the dependencies between the Gene Ontology and the Taxonomy of Species. This appendix takes a closer look at the data in publicly available taxonomic databases.

[1] One could argue that it is life forms that are systematized, rather than our knowledge about them. However, one could also argue that we classify life forms, and that the Taxonomy of Species is a systematization of such classifications into a single, coherent taxonomy — thus, it is a systematization of knowledge.

B.1 Introduction

In Ch. 4, 5, and A, the discussion has been constrained to the particular implementation of the Taxonomy provided by the National Center for Biotechnology Information (NCBI; Wheeler at el. [385]) — the NCBI Taxonomy.[2] However, there are a number of problematic issues related to the classification (or rather, classifications) of life forms. Specifically:

1. There isn't just one definition of 'species'. There is no unique, commonly agreed view on what species are. While from the point of view of a knowledge engineer developing an ontology a species is just a class like any other, its representation as a class is not so obvious to a biologist or a philosopher.

2. There isn't just one taxonomy of species. Besides the traditional Linnaean one, originally based on morphological similarity, there is, for example, the phylogenetic (evolutionary, cladistic) classification based on clades.[3]

3. There isn't just one taxonomical database. Even when only (some version of) the Linnaean Taxonomy is considered, there is no unique implementation of it — there are a number of taxonomic databases, and they are not consistent in details.

B.1.1 The Need for Taxonomic Annotation

The issues mentioned above are important not only in the context of the framework proposed earlier in this thesis. The use of a species classification system[4] is essential for successful navigation among public resources for

[2] http://www.ncbi.nlm.nih.gov/Taxonomy/

[3] PhyloCode (see further text) defines a *clade* as "an ancestor (an organism, population, or species) and all of its descendants."

[4] More generally, an organism classification system; the distinction between various infraspecies groupings, such as strains, is crucial for annotation of experimental data from studies conducted in model organisms.

supporting research in biology. Among the databases and services listed in the annual database supplement to *Nucleic Acids Research* (Galperin [122]), many either are dedicated to a particular species or to a more inclusive taxon, or, if they claim to be generic, include content that is usually associated with taxonomic information that specifies, e.g., the taxonomic status of the organisms used in original studies from which the data come.

For example, there are databases that specialize in gene regulation in eukaryotes, plant gene expression, or plant comparative genomics, and those that are specific to plant and fungal virus genes and genomes, or to structural virology. There is an *Archaeal* genome browser, and databases for bacterial comparative genomics and bacterial insertion sequences, sequences in prokaryotic genomes, microbial genomes, oomycetes and microbial genomes. There are phylogenetic databases for primate species and animal gene families; databases of mammalian microRNAs and gene promoters, invertebrate genes, and cereal genomes; human meiotic recombination hot spots, and mouse protein subcellular localization, *Drosophila* RNAi screening, comparative genomics of *Shigella*, comparative genomics of *Listeria* species, etc.

On the other hand, there are databases of orthologous promoters,[5] multiple genomes, and multi-species orthologs. There is Homologene, a system for annotated genes of 18 completely sequenced eukaryotic genomes; RefSeq, including protein sequences, representing almost 3000 organisms; EntrezGenome, which provides access to over 250 complete microbial genomic sequences, more than 2100 viral genomic sequences, and over 800 reference sequences for eukaryotic organelles. Those databases are comprehensive (in the sense of their taxonomic scope) and may, in some not so distant future, evolve to provide information on virtually *any* class of organisms studied well enough. It is thus essential that the data be indexed with references to those classes of organisms the records apply to.

The situation is no different in the case of biomedical ontologies. The OBO Foundry website lists ontologies covering gross anatomy of *C. elegans* (WBbt), cereal plant development (GRO), anatomy of *Dictyostelium dis-*

[5] Orthologous features are features that are homologous (similar due to shared ancestry) and are the result of a speciation event (emergence of new species from an old one).

coideum (DDANAT), human diseases (DOID), fungal gross anatomy (FAO), adult gross anatomy of mice (MAO), *Plasmodium* life cycle (PLO), etc. The OBO Foundry is an effort focused on providing controlled, structured vocabularies for the annotation of biological data — but, notably, the ontologies it lists are not explicitly annotated with the taxonomic groups they correspond to. (The taxonomic scope of OBO ontologies is addressed only in the ontologies' names, e.g., 'C. elegans gross anatomy', 'Plant structure', or 'Plasmodium life cycle'. Otherwise, there is no way to automatically select those ontologies that are relevant or specific to a particular taxon. The problem is more general: there is no explicit, structured classification whatsoever of the ontologies according to various properties of the domains they cover — structure, function, metadata (e.g., the Evidence Codes ontology, ECO). Indeed, as the number of ontologies collected under the umbrella of OBO increases, an ontology of OBO ontologies would be much in place.)

In what follows, first two issues are reviewed briefly — the 'concepts' of species and the distinction between the cladistic (phylogenetic, evolutional) and the traditional (morphology-based) classifications. It shall then be interesting to explore the issue of discrepancies between taxonomic databases.

B.2 The Problem of Species

In his *Species Concepts: the Basis for Controversy and Reconciliation*, Ghiselin [126] provides a brief discussion of (and argues for) the notion of species as individuals. Ghiselin distinguishes two perspectives on the 'concept' of species: species according to what he calls the 'biological consensus' (i.e., species as populations), and species according to what he calls the 'philosophical consensus' (i.e., species as individuals). Following this view, species are not immutable, extensionally defined entities. Rather, species are individuals, individual populations of organisms; each particular species is an individual of which the organisms are components or parts. A species is neither a class nor a set of which the organisms would be instances or members, respectively. (Ghiselin uses the term 'extensional class' as a synonym to 'set', and in opposition to 'class' used to talk about abstract entities which

may have different instances at different times. He also distinguishes classes from natural kinds in that all of the latter are classes, but not all classes are natural kinds — natural kinds are those classes "for which there exist laws of nature".)

Thus, the species *Homo sapiens* is an individual; all individual species are instances of the category *species*.[6] The class of species, in turn, is subsumed by the class of populations: all species are populations, though not all populations are species. Furthermore, any other taxonomic group is also an individual. However, as Mayden [247] explains, the systematic community argues that supraspecific taxa[7] do not exist in nature; they are manifestations of the historic past through ancestor-descendant relationships and are given proper names that systematists superimpose on a phylogenetic tree to identify monophyletic groups. One consequence of this is that taxa, like all individuals,[8] "can change indefinitely without ceasing to be the same individual".

Opposed to this view are, of course, a number of theories that would like to see species (and other taxa) as classes of organisms, understood either extensionally (as sets) or intensionally (as kinds, universals). Mayden [246, 247] provides an overview of quite a few[9] such accounts of the nature of species; besides the Biological Species Concept (BSC) mentioned above, there are the Evolutionary Species Concept (ESC), Phylogenetic Species Concept (PSC), Genetic Species Concept (GSC), Ecological Species Concept (EcSC), and others, most of which assume that species are classes. It seems that such views are more intuitive for the inexperienced. As Turner expressed it,

> "Some people even like to think of them [of species] as 'individuals', although, whether the use of the technical philosophical definition of this word is helpful to many biologists, is perhaps debatable." (Turner [373])

[6] Ghiselin also seems to synonymize the terms 'class' and 'category'.

[7] Taxa of a rank above species.

[8] More precisely, like all continuants; according to BFO, occurrents — which are individuals — are changes, but themselves do not change.

[9] In this case, 'quite a few' means more than 20.

To avoid such discussions, the earlier chapters adopted the naive view and talked about species as of classes. If desirable, the definitions given there may be modified so that organisms are not *instances*, but rather *parts* of taxa; for example,

$$(\textsf{all-some t g})^{\mathfrak{I}} = \begin{cases} \mathrm{T} & \text{if } \forall o \preceq \mathrm{I}(t)\ \exists f \in \mathrm{I}(g) : o \rhd f, \\ \mathrm{F} & \text{otherwise,} \end{cases} \tag{B.1}$$

where '$o \preceq \mathrm{I}(t)$' replaces '$o \in \mathrm{I}(t)$' and is taken to mean that the organism o is a part of the taxon $\mathrm{I}(t)$. (But I have no intention to engage here in a discussion on whether an organism can be a part of a non-species taxon.)

B.2.1 OntoClean: Species are Metaclasses

Interestingly, the case of species is an excellent example of how philosophical considerations may influence the requirements for expressivity of a representational formalism. In their *Evaluating Ontological Decisions with OntoClean*, Guarino and Welty argue:

> "Perhaps the most confusing of all distinctions is between the two relations subsumption and instantiation. We have often found the subsumption relationship misused when instantiation was actually intended. The canonical example of this is species/animal. While most introductory courses teach the difference between classes, such as *Mammal* or *Human*, and instances, such as *Chris*, they stop short of explaining how second-order classes, such as *Species*, would fit into the picture. *Human* is a subclass of *Mammal*, and *Chris* is a *Human* and therefore a Mammal. Is *Human* also a subclass of *Species*?
>
> "When we perform the analysis described on all these classes, we find that the identity criteria of *Species* are quite different from that of *Human*. Intuitively, species seem to be identified by their position in a biological taxonomy (for example, genus and differentia). On the other hand, we can assume that instances of *Human* are identified, in the simplest case, through the location in space/time of their bodies; two humans are different if they are at different places at the same time.

> "If *Human* was a subclass of *Species*, it would inherit its identity criteria. This can't be the case, since genus and differentia do not help in distinguishing one human from another. Therefore, *Species* cannot subsume *Human*. In fact, *Human* turns out to be an instance of *Species*, and subsumption is not instantiation." (Guarino and Welty [150, original emphasis])

Clearly, Guarino and Welty argue that *Chris* is an instance of *Human*; furthermore, *Human* is an instance of *Species*. Thus, *Human* must be a class,[10] and *Species* must be a class of classes, a *metaclass*.[11] Therefore, to represent the (meta)class *Species* those authors seem to require (though they do not say that in their article) a higher-order language, in which they could not only apply the predicate Human to the constant Chris, but also the predicate Species to the predicate Human, e.g.:

$$\mathsf{Human}(\mathsf{Chris}),\ \mathsf{Species}(\mathsf{Human}) \tag{B.2}$$

However, the issue here seems to be the result of confusing a class with an individual, both of which are called with the same name 'Human'. Under the assumption that species are individuals, the problem can be easily solved. *Species* is a class, but its instances are not classes themselves, they are individuals. Humans are individuals as well, but they are not instances of *Human*-the-species, but rather its parts. Of course, one may argue that humans do instantiate a class which we'd like to call 'Human'; but then *Human*-the-class and *Human*-the-species are obviously not the same: they are two distinct and quite different entities, which just happen to have been given the same name (which obviously breaks the principle of univocity; see, e.g., Smith [321]). This can be formalized as follows:

$$\begin{aligned}&\mathsf{Human}_{\mathrm{class}}(\mathsf{Chris}),\ \mathsf{Species}(\mathsf{Human}_{\mathrm{individual}}),\\ &\mathsf{PartOf}(\mathsf{Chris}, \mathsf{Human}_{\mathrm{individual}}),\end{aligned} \tag{B.3}$$

[10] It appears that Guarino and Welty do not make sufficiently clear what they mean by 'class' in that article: they talk about classes *representing* wholes, *polysemous* classes, and also about *instances of* classes.

[11] And if *Species* is a metaclass of species classes, *Species* itself is an instance of the metametaclass *Taxon* of taxonomic metaclasses such as *Genus*, *Order*, etc.

where subscripts are used to distinguished the two identically named entities, and $\mathsf{Human}_{\text{class}} \neq \mathsf{Human}_{\text{individual}}$.

Note that, unlike to Chris, $\mathsf{Human}_{\text{individual}}$ does not denote an individual human, but rather a species (i.e., an individual, an instance of the class *Species*). In a formalism where names of classes are distinguished from names of individuals syntactically, e.g., by using lower case for the former and upper case for the latter, B.3 can be rewritten as

$$\mathsf{human}(\mathsf{Chris}),\ \mathsf{species}(\mathsf{Human}),\ \mathsf{partOf}(\mathsf{Chris},\ \mathsf{Human}). \tag{B.4}$$

There is an interesting consequence of the distinction exemplified in the $\mathsf{Human}_{\text{class}}$-$\mathsf{Human}_{\text{individual}}$, or human-Human, case. In general, we would talk about a species, say s, referred to with a name, say n_s, and a class, say c, referred to with a name, say n_c. (The confusion discussed above arises if $n_s = n_c$.) Suppose that c is such that, at some time t, all and only organisms of the species s are of the class c. That is, every organism o that exists at t is an instance of c if and only if it is a part of s. However, species evolve: they are individuals that gain and lose parts, and the same species s may have an organism o as part at some time t, but not at another time t', $t' \neq t$.[12] If we take classes to be extensional entities (entities defined by their extensions, i.e., by the totalities of those entities taken to be their instances), then classes cannot change, but rather cease to exist when any of their instances cease to exist.[13] If, on the other hand, we take classes to be intensional entities (entities defined by their intension, i.e., by how the entities taken to be their instances are), then a class may have different extensions at different times, depending on the existence of different entities that match the class's intension, i.e., entities that instantiate the class at different times. (Thus, in a sense, classes understood intensionally can evolve, though it is their extensions and not intensions that change in time.)

[12] We may assume that an organism is part of one and only one species, and also that it is part of the same species throughout its life, though this is not essential here, and may even be wrong. But if we do, then '$o \preceq_t s$' and '$o \not\preceq_{t'} s$' may only mean that o does not exist at t'.

[13] Alternatively, a class, understood as an extensional entity, might be said to be partly present or non-present when some or all, respectively, of their instances cease to exist. See, e.g., Bittner et al. [45] for an account in which extensional entities (called 'collections' there) can be attributed partial presence or non-presence.

In the first case (classes as extensional entities), if s gains a part (a new organism), it no longer corresponds to the same class c, because the totality of all organisms of s is not the one that defines c. Since c is the class of all and only those organisms that are parts of the species s at t, then c cannot be the class of all and only those organisms that are parts of s at t′, after s has gained (or lost) a part. In the other case (classes as intensional entities), s may still correspond to the same class c after it has gained or lost parts; but species evolve not only in that they gain or lose parts, but also in that their new parts may be (and usually are) different in some respect from all other parts the species have ever had before. It may thus be that, at some time t′, s has as part an organism o which does not match the intension of the class c, and thus, again, c is not, at t′, the class of all organisms of the species s. Let us now return to the example taken from Guarino and Welty. We may provide an intensional definition of what it means to be a human in a way such that all and only those entities that are instances of the class *human* are also parts of the species *human*. But as the species evolves, new part-of-*human*-the-species organisms may appear that no longer are instances of *human*-the-class thus defined.

B.3 Taxonomic Classifications and Nomenclatures

The Linnaean Taxonomy of Species is based on morphological criteria: organisms are classified into more or less inclusive taxa according to the degree to which they resemble each other structurally. This traditional approach is accompanied by fairly complex naming schemes, e.g., the International Code of Botanical Nomenclature (ICBN; St. Louis Code [1]), the International Code of Zoological Nomenclature (ICZN),[14] and others. On the other hand, there is a phylogeny-based approach to the definition of biological taxa, which is the basis of the recently proposed alternative nomenclature for naming groups of organisms, the PhyloCode;[15] this approach — the phylogenetic systematics — is based on evolutionary relations between

[14] http://www.iczn.org/ .
[15] http://www.ohiou.edu/phylocode/ .

groups of organisms, past and present. These two schemes are largely incompatible, and are the topic of fierce debates. In *Stems, nodes, crown clades, and rank-free lists: is Linnaeus dead?*, Benton notes:

> "The proposals of phylogenetic nomenclature are to translate cladistic phylogenies directly into classifications, and to define taxon names in terms of clades. The method has a number of radical consequences for biologists: taxon names must depend rigidly on the particular cladogram favoured at the moment, familiar names may be reassigned to unfamiliar groupings, Linnaean category terms (e.g. phylum, order, family) are abandoned, and the Linnaean binomen (e.g. Homo sapiens) is abandoned. ... The consequences of this semantic maelstrom have not been worked out. In practice, phylogenetic nomenclature will be disastrous, promoting confusion and instability, and it should be abandoned. It is based on a fundamental misunderstanding of the difference between a phylogeny (which is real) and a classification (which is utilitarian). Under the new view, classifications are identical to phylogenies, and so the proponents of phylogenetic nomenclature will end up abandoning classifications altogether." (Benton [38])

The Linnaean Taxonomy is by far the most widely used one. The Gene Ontology, as well as other OBO ontologies, refers to it by means of the identifiers of the NCBI Taxonomy database. Therefore, in the rest of this chapter it is mostly the traditional taxonomy and the corresponding taxonomic databases that are in focus, though, occasionally, PhyloCode is also mentioned where relevant.

B.4 Problems with the Taxonomy

B.4.1 Taxonomic Databases

The NCBI Taxonomy database neither is, nor purports to be, an authoritative source of ultimate taxonomic information. Where desirable, several

different sources are used further in this appendix, such as:

- the Integrated Taxonomic Information System (ITIS),[16] one that claims to be authoritative, but yet still incomplete and possibly inaccurate;
- the Tree of Life Web Project (ToL),[17] a collaborative effort which provides information about the diversity of organisms on Earth, their evolutionary history (phylogeny), and characteristics;
- the Universal Biological Indexer and Organizer (uBio),[18] a set of tools that includes a taxonomic name server, a name bank, and a classification bank;
- the World Biodiversity Database (WBD),[19] a continuously growing taxonomic database and information system that allows you to search and browse a number of online species banks covering a wide variety of organisms.

B.4.2 Taxonomic Ranks

While building a classificatory hierarchy, it is usually desirable to have a clearly specified, but possibly extendable, set of criteria, which serve to distinguish the more or less inclusive classes into which individuals would be classified (see, e.g., Frank [117]). In the case of the Linnaean Taxonomy, taxa are assigned to various *ranks* which can reasonably be seen as playing the role of predefined levels of generality or specificity. With ranks, it is possible, e.g., to compare the generality of two taxa that are not related directly, that is, taxa such that none of them is an ancestor of the other. For example, both *Mammalia* (mammals) and *Amphibia* (amphibians) are taxa of the rank *class*; they can be said to be of the same generality, even though

[16] http://www.itis.gov/ .
[17] http://tolweb.org/tree/ .
[18] http://www.ubio.org/ .
[19] http://nlbif.eti.uva.nl/bis/ .

they are not siblings in the taxonomy (as reported by the NCBI Taxonomy database).

Whether ranks are or are not an essential part of the Taxonomy, has long been the matter of much debate. Although ranks are used by both ICBN and ICZN to specify nomenclatural conventions, they were abandoned in the original version of PhyloCode. Yet, recently, the advocates of PhyloCode seem to have modified their views and admit ranks as a component of the classification scheme (Pickett [282]). Ranks of taxa beyond the species level are often denied to have a precise and practical meaning, and seem to be treated as a historical legacy. In PhyloCode, the nomenclatural conventions are not dependent on ranks — a taxon can have a name which corresponds syntactically to a rank other than the actual rank of the taxon. However, despite their lack of clear, direct correspondence to, e.g., temporal and genetic distances between species and their common ancestors, ranks have proved useful for the purpose of building and maintaining the Taxonomy. Irrespectively of what ranks do or might mean, they are interesting from the point of view of supporting automated inference about the relative significance of taxa.

Both ICBN and ICZN use ranks; unfortunately, the codes differ in details, and it is not clear what approach to ranks will PhyloCode adapt. It is thus difficult to speak of just one hierarchy of ranks.[20] Table B.1 shows a hierarchy of ranks compiled from ICBN, ICZN, and the NCBI Taxonomy. (Ranks mentioned in the codes but for which no taxa were found are omitted.) For most ranks sanctioned by ICZN, there are ranks sanctioned by ICZN that have identical names. However, it is not clear whether such homonymous ranks should be seen as reflecting the same significance of taxa.[21] The ranks *Division* (ICZN) and *Phylum* (ICBN) are supposed to be equivalent. ICZN sanctions only a fixed set of ranks in the *Genus* and *Species* groups, namely, *Genus* and *Species*, and their subranks *Subgenus* and *Subspecies*, respectively.[22] ICBN is more liberal: it allows more non-primary ranks in these

[20] Ranks form a linear order. The hierarchy of ranks can thus be seen as a tree in which each non-leaf node has exactly one successor.

[21] Or, perhaps, the same names are used in both codes to refer to *the same*, rather than *homonymous*, ranks.

[22] http://www.iczn.org/, Art. 42 and 45.

groups, e.g., the secondary ranks *Section* and *Series*, and their further-order subranks *Subsection* and *Subseries* — in the *Genus* group — and the secondary ranks *Variety* and *Form*, and their further-order subranks *Subvariety* and *Subform* — in the *Species* group, and, in principle, this list is extendable.[23] (Both ICBN and ICZN use the term 'rank group', but with slightly different meanings. In ICBN, a rank group encompasses a primary rank and all non-primary ranks that lie *below this rank*, but *above the next primary rank* below. In ICZN, a rank group includes any non-primary rank whose name is derived from the name of the primary rank in the group.)

While ICZN sanctions only to a fixed set of named ranks, ICBN allows new ranks to be created if needed. Although primary and secondary ranks are fixed (see Sec. B.4.3), and thus only further-order ranks can be added to the existing system, it is not entirely clear what are the rules for creating and naming such new ranks. The hierarchies of ranks used by ICBN and ICZN already differ, and an arbitrary extension made to one of them may only increase incoherence. There are in use (in the NCBI Taxonomy database) ranks that are not mentioned in any of the codes. The position of such ranks in the hierarchy is not explicitly given, and can only be inferred from the position of the taxa that are of these ranks. For example, in the zoological part of the Taxonomy,[24] there are in use ranks such as *Superclass*, *Superorder*, *Infraclass* and *Infraorder*, which are sanctioned by ICBN but not by ICZN; there are also in use primary and other ranks which are not sanctioned by ICZN: the ranks *Kingdom* and *Domain* are unofficial in this sense. Further examples are the ranks *Species group* and *Species subgroup*.

B.4.3 Rank Orders

The hierarchy of ranks is linear (totally ordered) — every rank is either above or below any other rank. Yet the hierarchy is not quite a flat list, in that each rank is assigned to one of several levels or orders: there are *principal* (or *primary*) ranks, *secondary* ranks, and *further* ranks. These levels

[23] http://www.bgbm.org/iapt/nomenclature/code/SaintLouis/0000St.Luistitle.htm, Art. 4.
[24] The taxon *Metazoa* (a *Kingdom* taxon) and all taxa below it.

presumably reflect some form of significance or importance of ranks (which should not be confused with the significance of taxa). There are also ranks that do not have any explicit level of importance; in what follows, such ranks are called 'no-order' ranks, and the term 'further' is reserved to those ranks that are mentioned as such in ICBN. ICZN does not explicitly assign orders to its ranks. As shown in Tab. B.1, the counts of taxa of further- and no-order (unofficial) ranks are not marginal.

Is there any real meaning behind the rank orders? Intuitively, secondary and further ranks, understood as supplementary to the primary ranks, provide finer levels of detail for the classification of organisms. In this sense, one might use rank orders to view the taxonomy at different levels of detail, simply by hiding or revealing taxa that are of ranks beyond or below a particular order. For example, one might want to see only the primary taxa included in the classification of a particular species-level taxon. Figure B.1 shows the taxonomic classification of the subspecies *Homo sapiens sapiens*; if only primary taxa were selected, the view would be compressed to that shown in Fig. B.2. However, for this intuition to make sense, any classification should include a taxon of a lower-order rank only if it also includes a taxon of that higher-order rank which the lower-order rank is a refinement of; i.e., a classification should not include a subclass taxon if it does not include a class taxon, etc. (see Tab. B.1 again). Otherwise, fine-grained details would be provided when no coarse-grained information is present. Unfortunately, this intuition fails, as there are many cases that contradict this should-be rule. For example, the classification of the species *Citrus limon* includes a subclass taxon, but no class taxon; the classification of *Lama glama* includes a suborder taxon, but no order taxon; *Corydalis sempervirens* is classified under a subfamily, but not under a family.

Rank name	In ICBN	In ICZN	Taxon count
· · · · Superkingdom (Domain)			3
Kingdom	•		3
· · · · Superdivision (Superphylum)			4
Division (Phylum)	•		79
· · · Subdivision (Subphylum)	•		14
· · · · Superclass			5
Class	•		212
· · · Subclass	•		101
· · · · Infraclass			11
· · · · Superorder			53
Order	•		1010
· · · Suborder	•		339
· · · · Infraorder			72
· · · · Superfamily		•	649
Family	•	•	5735
· · · Subfamily	•	•	1629
· · Tribe	•	•	1053
· · · Subtribe	•	•	202
Genus	•	•	37336
· · · Subgenus	•	•	576
· · · · Species group			182
· · · · Species subgroup			80
Species	•	•	222438
· · · Subspecies	•	•	6240
· · Variety	•		1683
· · Form	•		142
no rank			19840

Table B.1: Taxonomic ranks according to ICBN and ICZN, ordered by the purported significance of taxa assigned to them. Primary ranks in boldface. Secondary, further, and unordered ranks marked with · · , · · · , and · · · · , respectively. Taxon counts as found in NCBI Taxonomy, June 2006.

domain: *Eukaryota*
—: *Fungi/Metazoa group*
kingdom: *Metazoa*
— (subkingdom): *Eumetazoa*
— (superphylum): *Bilateria*
—: *Coelomata*
—: *Deuterostomia*
phylum: *Chordata*
subphylum: *Craniata*
— (subphylum): *Vertebrata*
superclass: *Gnathostomata*
—: *Teleostomi*
—: *Euteleostomi*
—: *Sarcopterygii*
—: *Tetrapoda*
—: *Amniota*
class: *Mammalia*
—: (subclass): *Theria*
—: (infraclass): *Eutheria*
superorder: *Euarchontoglires*
order: *Primates*
suborder (infraorder): *Simiiformes*
—: *Catarrhinii*
superfamily: *Hominoidea*
family: *Hominidae*
(**subfamily**: *Homininae*)
—: *Homo/Pan/Gorilla group*
genus: *Homo*
species: *Homo sapiens*
subspecies: *Homo sapiens sapiens*

Figure B.1: Complete classification of *Homo sapiens sapiens*. Ranks and taxa found in databases other than the NCBI Taxonomy are in parentheses. Taxa with no ranks in the NCBI Taxonomy are preceded by an em-dash (‘ — ’). Ranks explicitly mentioned by ICZN are in boldface.

kingdom: *Metazoa*
phylum: *Chordata*
class: *Mammalia*
order: *Primates*
family: *Hominidae*
genus: *Homo*
species: *Homo sapiens*

Figure B.2: Modified classification of *Homo sapiens sapiens* — only taxa of primary ranks are shown. Only primary ranks are visible. Ranks explicitly mentioned in ICZN are in boldface.

To estimate the scale of the problem, we have conducted a small experiment: Among 5,000 randomly picked species taxa, in nearly 500 cases — approximately 10% — the rule was disobeyed. (The study was based on data from the NCBI Taxonomy database. NCBI Taxonomy and other taxonomic databases, notably ITIS, often disagree on taxon names and their ranks, and the results may reflect an idiosyncrasy of the source used.) Rank orders seem to be a quite *ad hoc* invention, and persist presumably only for historical reasons. Primary ranks are primary because they were conceived first: the original taxonomy devised by Linnaeus contained only the ranks *Kingdom*, *Order*, *Genus* and *Species*, while secondary and further ranks, as well as no-order ranks, are those added by later systematists. But apart from providing this historical insight, rank orders seem to bring no real benefit to the somewhat chaotic system.

B.4.4 Rank Names

The taxonomic ranks are named (see Table B.1). Primary and secondary ranks have arbitrary names; further-order ranks have (according to ICBN) names formed by extending the names of primary and secondary ranks with the prefix 'sub-'. This naming policy is too constrained — what if new ranks are to be added? The sets of primary and secondary ranks are fixed (none of the codes allows one to create a new primary rank); ranks may be added

only with names created according to the prefixing scheme. But this means that there may only be a relatively small number of ranks, and that it may be impossible to assign a rank to some taxa.

For example, in the classification of *H. sapiens sapiens* (see Fig. B.1), what would be the names and orders appropriate for ranks to which the five taxa placed between the superclass *Gnathostomata* and the class *Mammalia* might be assigned? Clearly, due to the rank naming and ordering policy, a separation of the ranks *Superclass* and *Class* by another rank is inconceivable. Should a rank such as *Subsubsuperclass* be introduced?

B.4.5 Taxa and Their Ranks

One possible use of the rank system could be to provide a means for characterizing the significance of ranks. The category selected for any particular taxon reflects the significance of the group in some way, and it is determined by the placement of that taxon in a classification list (Benton [38]). In principle, it would be useful to consider two taxa of the same rank as equally general, irrespective of their actual taxon-wise distance from the root of the Taxonomy, or, alternatively, from their leaf-level subtaxa.[25] However, this may be difficult in practice: a large proportion of taxa do not have ranks, and ranks are not used in a coherent manner.

Few species have classifications including taxa of all available ranks. Worse still, in many cases such classifications include taxa of secondary and further-order ranks, but not of some of the primary ranks (see Sec. B.4.3). Unranked taxa may be divided into two groups: those for which it is possible to assign an existing rank without contradicting the respective code's rules, and those for which there is no rank available, unless a new one is created. In the former case it is, in principle, possible to infer the rank from the ranks of the closest ancestor and successors, but there may be more than one rank available. In the latter case, new ranks should be created, but this may be hindered by the rank-naming policy of taxonomic codes (see Sec. B.4.4).

[25] In the Taxonomy, leaf-level taxa are of different ranks: not only species, but also subspecies and other taxa may be leaves.

Many of the taxa found in the classification of *H. sapiens sapiens* (Fig. B.1) are unranked — according to the NCBI Taxonomy database, only 50% of taxa in this classification have ranks. The taxa *Theria* and *Eutheria* can only be assigned to the ranks *Subclass* and *Infraclass*, respectively, since they are placed immediately between a *Class* and a *Superorder*. (Indeed, this is how they are ranked according to ITIS.) There is no guess, however, to be done about the ranks of *Teleostomii*, *Euteleostomii*, *Sarcopterygii*, *Tetrapoda*, and *Amniota*, as there are no ranks between *Superclass* and *Class*. Five new ranks between *Superclass* and *Class* would have to be invented; what would be the names and orders of these five ranks?

If all taxa were ranked, comparing the generality of two taxa would be straightforward and could be done in constant time (e.g., by using a two-dimensional matrix of precomputed rank-rank comparisons). For unranked taxa, the process is more complicated: the closest ranked ancestors and successors of the respective taxa must be found and compared, and the result may still be uncertain. In the absence of explicit ranks, the count of intermediate taxa on the path to the root of the taxonomy may be used, but this purely structural measure of depth does not necessarily correspond well to how systematists understand the generality of taxa. A species taxon may be placed very shallowly (e.g., *Pleurochrysis sp.* is only 4 (NCBI) or 7 (ITIS) taxa from the root) or very deeply (e.g., *Jordanella pulchra* is 39 (NCBI) or 15 (ITIS) taxa from the root). Likewise, for every rank in the *Species* group, there is at least one leaf taxon of that rank in the Taxonomy.

For example, the taxon *Didelphidae* could thus be seen as more general than the taxon *Homo*, because the former is of the rank *Family* and the latter is of the rank *Genus* (Table B.1) — even though none of them is an ancestor or successor of the other. Clearly, the taxon *Teleostomii* is more general than the taxon *Catarrhinii*, because the latter is subsumed by the former (Fig. B.1) — even though none of them has an explicit rank. *Teleostomii* is also more general than *Pterygota* (Fig. B.3), because the former is placed between a superclass and a class, while the latter is placed between a class and a subclass.

However, it is impossible to decide whether the taxon *Bacillariophycidae* is

more, equally, or less general than any of the taxa *Dicondylia*, *Pterygota* and *Palaeoptera* (Fig. B.3). If no rank information were available and structural depth were used as an estimate of taxon generality, any of the latter three taxa would have to be considered much more specific than *Bacillariophyceae*, simply because they are placed three times deeper in the hierarchy.[26]

...†		...‡
class: *Insecta*	=	class: *Bacillariophyceae*
(no rank): *Dicondylia*		
(no rank): *Pterygota*	≈	(no rank): *Bacillariophycidae*
subclass: *Palaeoptera*		
order: *Odonata*	=	order: *Naviculales*
...		...

Figure B.3: Rank-based mapping of taxa between two partial lineages of *Odonata* and *Naviculales*; data from NCBI Taxonomy. Omitted ancestors: †13 taxa above, ‡4 taxa above. Symbols: = unambiguous taxon-taxon mapping (corresponding explicit ranks); ≈ uncertain mapping.

B.4.6 Taxa and Their Names

One of the principal goals of ICBN and ICZN is to normalize the nomenclature of taxa. Some of the central rules employed by these systems are: each taxon has exactly one official name; no two taxa have the same official name; the name of a taxon reflects its rank. Unfortunately, taxonomic databases do not necessarily obey these rules in a consistent and coherent manner.

In the NCBI Taxonomy, all taxa have exactly one official name. However, some taxa seem to have been given different names in different databases. Since each database employs its own system of unique taxon identifiers,

[26]Comparing the generality of taxa such as *Insecta* and *Bacillariophyceae* seems to be a highly speculative business, but the very term 'rank' is directly suggestive of there being some sort of generality or significance that is common to both those class-ranked taxa mentioned above.

the only way to match such differently named taxa across databases is to use additional information, e.g., their ranks and position in the respective hierarchies. For example, the kingdom that includes all plants is named 'Viridiplantae' in NCBI Taxonomy, but 'Planta' in ITIS. Since both databases include exactly one kingdom taxon that covers plants,[27] it is reasonable to believe that 'Viridiplantae' and 'Planta' are two distinct official names for the same taxon. There is no taxon named 'Planta' in NCBI Taxonomy, and no taxon named 'Viridiplantae' in ITIS — so that these two names are not treated as a synonyms by either database alone.

Table B.2 shows further two examples: the taxonomic classification of two genera, *Bacillus* and *Ficus*. In the case of *Ficus*, one may be relatively confident that the alignment is correct, and that 'Sorbeoconcha' is synonymous with 'Neotaeniaglossa', though it would not be easy to align these two names on a purely lexical basis. Alignment in the case of *Bacillus* is less obvious.

Rank	*Bacillus*		*Ficus*	
	NCBI	ITIS	NCBI	ITIS
kingdom	—	***Monera***	***Metazoa***	***Animalia***
phylum	***Firmicutes***	***Bacteria***	*Mollusca*	*Mollusca*
class	***Bacilli***	***Schizomycetes***	*Gastropoda*	*Gastropoda*
order	***Bacillales***	***Eubacteriales***	***Sorbeoconcha***	***Neotaenioglossa***
family	*Bacillaceae*	*Bacillaceae*	*Ficidae*	*Ficidae*
genus	*Bacillus*	*Bacillus*	*Ficus*	*Ficus*

Table B.2: Partial classifications of two genera, *Bacillus* and *Ficus*, as in NCBI Taxonomy and ITIS. Names that differ in NCBI and ITIS appear in boldface.

ICZN employs what it calls the 'principle of homonymy':[28] "When two or more taxa are distinguished from each other they must not be denoted by the same name." On the other hand, ICBN does allow homonyms.[29] Although rules for the acceptance and rejection of new taxon names are rel-

[27] It would have been rather surprising had it been otherwise.
[28] ICZN, Art. 52.
[29] ICBN, Art. 53.

atively simple, a considerable number of exceptional situations make the whole system unreasonably complicated and difficult to maintain. There are many examples of multiple taxa with the same name, both within a single taxonomic database and between databases, and both in the zoological and the botanical part of the taxonomy. NCBI's attempt to disambiguate such cases is based on what they call 'unique names': homonymous taxon names extended with a disambiguation string that may, very irregularly, reflect a taxon's rank, the name of one of its ancestors, or a general, non-scientific description of the organism.[30]

While most cases of homonymy to be found in both NCBI and ITIS involve taxa placed in distinct kingdoms,[31] there are also cases of homonymous taxon names within one kingdom. For example, the name 'Morganella' denotes three distinct genera: one proteobacterial, one fungal (in NCBI), and one animal (in ITIS). For disambiguation, the two former are given, in the NCBI Taxonomy, the unique names 'Morganella ⟨proteobacterium⟩' and 'Morganella ⟨fungus⟩', respectively. There are also two taxa named 'Branchiura': one of them is a genus under the phylum *Annelida*, the other is a subclass under the phylum *Arthropoda*, both in the zoological part of the taxonomy. NCBI disambiguates them as *Branchiura* ⟨*Annelida*⟩ and *Branchiura* ⟨*Crustacea*⟩. Similarly, the name 'Chlamydiae' is used to refer to both the phylum *Chlamydiae* as well as to its subordinate class taxon *Chlamydiae*. To disambiguate, NCBI uses the unique name 'Chlamydiae (class)' for the latter. Analogous solution has been adopted for the case of 'Spirochaetes', 'Actinobacteria', 'Fusobacteria', etc. Another interesting case from the NCBI Taxonomy is that of the genus *Drosophila* and its homonymous subgenus. While officially the latter is called 'Drosophila subg. Drosophila', NCBI reports 'Drosophila' as the official name, and hence needs to disambiguate the name, using the unique names 'Drosophila ⟨fruit fly, genus⟩' and 'Drosophila ⟨fruit fly, subgenus⟩', respectively.

While there is no formal requirement, taxon names should best reflect some

[30] And, indeed, this scheme resembles the 'sensu' clauses of species-specific GO terms.

[31] Which, by the way, is legal according to the codes, as neither of them seems to care about what the other does: "The name of an animal taxon identical with the name of a taxon which has never been treated as animal is not a homonym for the purposes of zoological nomenclature." ICZN, Art. 52.7.

essential properties of the organisms covered; unfortunately, names of lower-rank taxa (e.g., genera and species) often do not reflect any such properties.[32] Examples such as *Agathidium bushi*, *Bufonaria borisbeckeri*, *Abra cadabra*, *Ytu brutus*, and others, abound.[33] Yet another kind of nomenclatural problems is related to the principles of forming names of higher-order taxa. The are a few naming rules, which do not, however, guarantee a clean naming practice, and may be somewhat cumbersome to use. Consider the example of the family *Asteraceae* (see Fig. B.4):

> "The subfamily of the family *Asteraceae* Martinov (nom. alt., *Compositae* Adans.) including *Aster* L., the type of the family name, is irrespective of priority to be called *Asteroideae* Asch., and the tribe and subtribe including *Aster* are to be called *Astereae* Cass. and *Asterinae* Less., respectively. However, the correct name of the tribe including both *Cichorium* L., the type of the subfamily name *Cichorioideae* W. D. J. Koch (1837), and *Lactuca* L. is *Lactuceae* Cass. (1815), not *Cichorieae* D. Don (1829), while that of the subtribe including both *Cichorium* and *Hyoseris* L. is *Hyoseridinae* Less. (1832), not *Cichoriinae* Sch. Bip. (1841) (unless the *Cichoriaceae* Juss. are accepted as a family distinct from *Compositae*)." [ICBN, Art. 19.4, Ex. 4]

Interestingly, taxonomic databases (e.g., NCBI Taxonomy) use the names 'Cichorieae' and 'Cichoriinae' for the tribe and subtribe mentioned above, which appears to be a more intuitive solution, despite ICBN's instructions.

As a final example of inconsistently realized nomenclatural conventions, consider the rule, adopted by both ICBN and ICZN, demanding that the name of a taxon should reflect the rank of the taxon by means of a rank-specific suffix added to the respective type name. Unfortunately, the two codes use different suffixes to express corresponding ranks, as well as use the same suffixes for different ranks. For example, the suffix '-oidea' is reserved by ICZN for superfamilies, but is also legal in names of taxa of other ranks, as in, e.g., *Ranoidea* (a genus), *Asteroidea* (a class). The suffix '-idae'

[32] The name of a genus "may be taken from any source whatever, and may even be composed in an absolutely arbitrary manner". ICBN, Art 20.1.

[33] http://home.earthlink.net/m̃isaak/taxonomy.html provides an impressively extensive archive of such (un)usual taxon names.

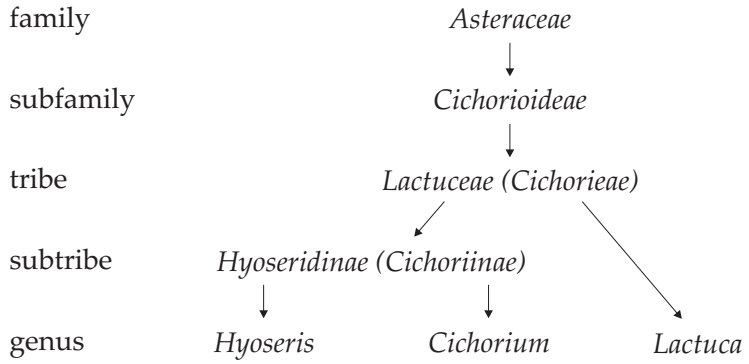

Figure B.4: Partial classification of the family *Asteriaceae*, according to ICBN, Art. 19.4. Taxon names discouraged by ICBN, but used by the NCBI Taxonomy database as official names, are in parentheses.

is reserved by ICZN for families, but rather for subclasses in ICBN, as in, e.g., *Hominidae* (a family), *Bryidae* (a subclass), etc. Other codes may decide on yet other nomenclatural principles. In consequence, it may be difficult to predict the rank of a taxon from its name alone. In PhyloCode, taxon names are completely independent of ranks. Although clades are hierarchically related, assignment of a categorical rank (e.g., genus, family, etc.) is not part of the formal naming process and has no bearing on the spelling or application of taxon names (PhyloCode, Art. 3).

B.5 Discussion

This appendix provides a qualitative and quantitative insight into the current state of taxonomic information available in public databases, such as the NCBI Taxonomy database, the Integrated Taxonomic Information System, etc. The information distributed among those different sources does

not seem to be coherently represented, and the Taxonomy of species itself is not internally consistent. Chapter 5 and App. A show how the Gene Ontology and the Taxonomy of Species can be meaningfully connected to provide for automated partitioning of the GO. However, App. B should be convincing that it is important not to rest on the assumption that the NCBI Taxonomy is the only available source of taxonomic data, or that it is a reliable source, or even that such a source exists at all. Therefore, all reasoning over the Gene Ontology based on links with the Taxonomy of Species should be done with care.

Bibliography

[1] *International Code of Botanical Nomenclature (St Louis Code)*, volume 138 of *Regnum Vegetabile*. Koeltz Scientific Books, Königstein, 1999. 167

[2] A decade of genome-wide biology. *Nature Genetics*, 37, 2005. Editorial. 2

[3] Editorial. *Nucleic Acids Research*, 34(supplement 2):W1, 2006. 45

[4] A. Aamodt. *A Knowledge Intensive, Integrated Approach to Problem Solving and Sustained Learning*. PhD thesis, Norwegian University of Science and Technology, 1991. 10

[5] A. Aamodt. A knowledge-representation system for integrating general and case-specific knowledge. In *Proceedings of IEEE TAI-94, International Conference on Tools with Artificial Intelligence*, 1994. 10

[6] A. Aamodt. Modeling the knowledge contents of cbr systems. In *Proccedings of the Workshop Program at the Fourth International Conference on Case-Based Reasoning*, pages 32–37, 2001. 10

[7] A. Aamodt and M. Nygard. Different roles and mutual dependencies of data, information, and knowledge - an AI perspective on their integration. *Data Knowledge Engineering*, 16(3):191–222, 1995. 29, 30

[8] A. Aamodt and E. Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *Artificial Intelligence Communications*, 7:39–52, 1994. 10

[9] H. Abelson and G. J. Sussman. *Structure and Interpretation of Computer Programs*. MIT Press, second edition, 1996. 80

[10] M. Adams, S. Celniker, R. Holt, C. Evans, J. Gocayne, P. Amanatides, S. Scherer, P. Li, R. Hoskins, R. Galle, et al. The genome sequence of *Drosophila melanogaster*. *Science*, 287:2185–2195, 2000. 104

[11] U. Alon. *An Introduction to Systems Biology. Design Principles of Biological Circuits*. Chapman & Hall/CRC, 2007. 8, 29

[12] G. Alterovitz, M. Xiang, M. Mohan, and M. F. Ramoni. Go pad: the gene ontology partition database. *Nucleic Acids Res*, 35(Database issue):D322–D327, Jan 2007. 58, 106

[13] R. Altman. Annual progress in bioinformatics 2006. *Briefings in Bioinformatics*, 7:209–210, 2006. 28

[14] R. Altman, M. Bada, X. Chai, M. W. Carillo, R. Chen, and N. Abernethy. RiboWeb: An ontology-based system for collaborative molecular biology. *IEEE Intelligent Systems*, 14:68–76, 1999. 57

[15] P. Angenendt, H. Lehrach, J. Kreutzberger, and J. Glïċ$\frac{1}{2}$kler. Subnanoliter enzymatic assays on microarrays. *Proteomics*, 5(2):420–425, Feb 2005. 33

[16] G. A. Antonelli. Non-monotonic logic. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. 2006. 153

[17] G. Antoniou and F. van Harmelen. *Handbook on Ontologies*, chapter Web Ontology Language: OWL, pages 67–92. Springer-Verlag Berlin Heidelberg, 2004. 60, 68

[18] M. E. Aranguren, S. Bechhofer, P. Lord, U. Sattler, and R. Stevens. Understanding and using the meaning of statements in a bio-ontology: recasting the Gene Ontology in OWL. *BMC Bioinformatics*, 8:57, 2007. 12

[19] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–29, May 2000. 3, 56, 99

[20] E. M. Awad and H. M. Ghaziri. *Knowledge Management*. Pearson Education, Inc., 2004. 29, 134

[21] M. Aydede. The language of thought hypothesis. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2004. 80

[22] F. Baader, I. Horrocks, and U. Sattler. *Handbook on Ontologies*, chapter Description Logics, pages 3–28. Springer-Verlag Berlin Heidelberg, 2004. 60, 68, 156

[23] F. Baader and W. Nutt. *The Description Logics Handbook. Theory, Implementation, Applications*, chapter Basic Description Logics, pages 47–100. Cambridge University Press, 2003. 145, 156

[24] J. Bacon. Tropes. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2002. 88

[25] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, 1999. 9

[26] P. G. Baker, C. A. Goble, S. Bechhofer, N. W. Paton, R. Stevens, and A. Brass. An ontology for bioinformatics applications. *Bioinformatics*, 15(6):510–520, Jun 1999. 58

[27] P. Baldi and S. Brunak. *Bioinformatics: The Machine Learning Approach*. MIT Press, 2001. 28

[28] J. Bard, S. Y. Rhee, and M. Ashburner. An ontology for cell types. *Genome Biology*, 6:R21, 2005. 96

[29] M. T. Barrett. Stacking the chips for biological discovery. *Nature Genetics*, 37 Suppl:S1, Jun 2005. 36

[30] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, and R. Edgar. NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Research*, 35(Database issue):D760–D765, Jan 2007. 38, 47

[31] G. Bealer. Universals. *The Journal of Philosophy*, 90(1):5–32, January 1993. 89

[32] S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D. McGuinness, P. F. Patel-Schneider, and L. A. Stein. *OWL Web Ontology Language reference. W3C Recommendation 10 February 2004*, 2004. 71

[33] V. Beisvåg, L. Jølsum, W. Kuśnierczyk, M. Langaas, B. K. Alsberg, H. Bergum, J. Komorowski, A. K. Sandvik, and A. Lægreid. eGOn: A new tool for mapping microarray data onto the Gene Ontology structure. In *Proceedings of the 1st Workshop on Standards and Ontologies for Functional Genomics*, 2002. 7, 58

[34] V. Beisvåg, F. K. R. Jünge, H. Bergum, L. Jølsum, S. Lydersen, C.-C. Günther, H. Ramampiaro, M. Langaas, A. K. Sandvik, and A. Lægreid. GeneTools — application for functional annotation and statistical hypothesis testing. *BMC Bioinformatics*, 7:470, 2006. 7, 27, 43, 58

[35] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, and J. Morissette. Bio2RDF: Towards a mashup to build bioinformatics knowledge system. In *Proceedings of the 16th International World Wide Web Conference (WWW2007)*, 2007. 53

[36] R. Bellman. *Dynamic Programming*. Dover Publications, 2003. 134

[37] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. GenBank. *Nucleic Acids Research*, 35(Database issue):D21–D25, Jan 2007. 38

[38] M. J. Benton. Stems, nodes, crown clades, and rank-free lists: is Linnaeus dead? *Biological Reviews*, 75:633–648, November 2000. 168, 176

[39] H. Berman, K. Henrick, H. Nakamura, and J. L. Markley. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Research*, 35(Database issue):D301–D303, Jan 2007. 40

[40] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res*, 28(1):235–242, Jan 2000. 40

[41] J. Bermúdez. Nonconceptual mental content. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2003. 80

[42] P. A. Bernstein. Middleware: A model for distributed system services. *Communications of the ACM*, 39(2):86–98, 1996. 44

[43] M. Bhatt, C. Wouters, A. Flahive, W. Rahayu, and D. Taniar. Semantic completeness in sub-ontology extraction using distributed methods. In *Proceedings of Computational Science and Its Applications (ICCSA 2004)*, 2004. 99

[44] T. Bittner. Axioms for parthood and containment relations in bio-ontologies. In *KR-MED 2004: Workshop on Formal Biomedical Knowledge Representation*, pages 4–11, 2004. 117

[45] T. Bittner, M. Donnelly, and B. Smith. Individuals, universals, collections: On the foundational relations of ontology. *AI Communications, IOS Press*, 13:247–258, 2004. 89, 91, 166

[46] S. Blackburn. *Oxford Dictionary of Philosophy*. Oxford Paperback Reference. Oxford University Press, second edition, 2005. 67, 90, 92

[47] J. Blake, D. P. Hill, and B. Smith. Gene Ontology annotations: What they mean and where they come from. In *Proceedings of the 10th*. 15, 58

[48] J. A. Blake and C. J. Bult. Beyond the data deluge: Data integration and bio-ontologies. *Journal of Biomedical Informatics*, 39(3):314–320, Jun 2006. 58

[49] O. Bodenreider. Using UMLS semantics for classification purposes. In *Proceedings of AMIA Symposium*, pages 86–90, 2000. 58

[50] O. Bodenreider, J. A. Mitchell, and A. T. McCray. Biomedical ontologies. In *Pacific Symposium on Biocomputing*, pages 76–78, 2005. 59

[51] O. Bodenreider, B. Smith, and A. Burgun. The ontology-epistemology divide: A case study in medical terminology. In A. Varzi and L. Vieu, editors, *Proceedings of the International Conference on Formal Ontology and Information Systems, FOIS2004*, 2004. 12, 70, 126

[52] O. Bodenreider, B. Smith, A. Kumar, and A. Burgun. Investigating subsumption in DL-based terminologies: A case study in SNOMED-CT. In *Proceedings of the 1st International Workshop on Formal Biomedical Knowledge Representation (KR-MED 2004)*, 2004. 88

[53] O. Bodenreider and R. Stevens. Bio-ontologies: current trends and future directions. *Briefings in Bioinformatics*, 7(3):256–274, Sep 2006. 59

[54] D. Booth. URIs and the myth of resource identity. In *Proceedings of the WWW2006 Workshop on Architecture and Philosophy of the Web. Identity, Reference, and the Web (IRW2006)*, 2006. 77

[55] R. Boyd. Scientific realism. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Summer 2002. 78

[56] A. Brazma. On the importance of standardisation in life sciences. *Bioinformatics*, 17(2):113–114, Feb 2001. 27, 40, 49

[57] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genetics*, 29(4):365–371, Dec 2001. 47

[58] A. Brazma, M. Krestyaninova, and U. Sarkans. Standards for systems biology. *Nature Reviews Genetics*, 7:593–605, 2006. 46, 48

[59] J. F. Brinkley, D. Suciu, L. T. Detwiler, J. H. Gennari, and C. Rosse. A framework for using reference ontologies as a foundation for the semantic web. In *AMIA Symposium Proceedings*, 2006. 96

[60] P. O. Brown and D. Botstein. Exploring the new world of the genome with DNA microarrays. *Nature Genetics*, 21, 1999. 2

[61] J. J. Burbaum and N. H. Sigal. New technologies for high-throughput screening. *Current Opinion in Chemical Biology*, 1(1):72–78, Jun 1997. 32

[62] A. Burgun. Desiderata for domain reference ontologies in biomedicine. *Journal of Biomedical Informatics*, 39(3):307–313, Jun 2006. 12

[63] A. Burgun and O. Bodenreider. Mapping the UMLS Semantic Network into general ontologies. pages 81–85, 2001. 69

[64] K. Burrage, L. Hood, and M. A. Ragan. Advanced computing for systems biology. *Briefings in Bioinformatics*, 7(4):390–398, Dec 2006. 29

[65] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, and R. Apweiler. The Gene Ontology Annotation (GOA) Database: Sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Research*, 32:D262–D266, 2004. 58, 127

[66] E. B. Camon, D. G. Barrell, E. C. Dimmer, V. Lee, M. Magrane, J. Maslen, D. Binns, and R. Apweiler. An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. *BMC Bioinformatics*, 6, 2005. 126

[67] A. M. Campbell and L. J. Heyer. *Discovering Genomics, Proteomics, and Bioinformatics*. Pearson Education, Inc., second edition, 2007. 2

[68] R. Caspi, H. Foerster, C. A. Fulcher, R. Hopkinson, J. Ingraham, P. Kaipa, M. Krummenacker, S. Paley, J. Pick, S. Y. Rhee, C. Tissier, P. Zhang, and P. D. Karp. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Research*, 34:D511–D516, 2006. 10

[69] M. Cassman. Barriers to progress in systems biology. *Nature*, 438(7071):1079, Dec 2005. 46

[70] W. Ceusters and B. Smith. A realism-based approach to the versioning and evolution of biomedical ontologies. In *Proceedings of the AMIA Symposium*, 2006. 12, 127

[71] W. Ceusters, B. Smith, A. Kumar, and C. Dhaen. Mistakes in medical ontologies: Where do they come from and how can they be detected? In *Ontologies in Medicine: Proceedings of the Workshop on Medical Ontologies*, 2003. 12, 88

[72] W. Ceusters, B. Smith, A. Kumar, and C. Dhaen. Ontology-based error detection in SNOMED-CT. In *Proceedings of Medinfo*, 2004. 88

[73] D. D. Chamberlin and R. F. Boyce. SEQUEL: A structured English query language. In *International Conference on Management of Data, Proceedings of the 1974 ACM SIGFIDET (now SIGMOD) workshop on Data description, access and control*, pages 249–264, 1974. 43

[74] B. Chandrasekaran, J. R. Josephson, and V. R. Benjamins. What are ontologies, and why do we need them? *IEEE Intelligent Systems*, 14:20–26, 1999. 68

[75] V. K. Chaudhri, A. Farquhar, R. Fikes, P. D. Karp, and J. P. Rice. *Open Knowledge Base Connectivity 2.0.3*, 1998. 71, 72, 73, 81

[76] K.-H. Cheung, K. Y. Yip, A. Smith, R. Deknikker, A. Masiar, and M. Gerstein. YeastHub: a semantic web use case for integrating data in the life sciences domain. *Bioinformatics*, 21 Suppl 1:i85–i96, Jun 2005. 47

[77] J. J. Cimino. In defense of the Desiderata. *Journal of Biomedical Informatics*, 39(3):299–306, Jun 2006. 12, 82

[78] J. J. Cimino and X. Zhu. The practical impact of ontologies on biomedical informatics. *Methods in Information in Medicine*, 46:124–135, 2006. 59

[79] J. I. Clark, C. Brooksbank, and J. Lomax. It's all GO for plant scientists. *Plant Physiology*, 138:1268–12797, 2005. 102

[80] P. Clark and B. Porter. *KM—The Knowledge Machine 2.0 User's Manual*. 72

[81] T. Clark, S. Martin, and T. Liefeld. Globally distributed object identification for biological knowledgebases. *Briefings in Bioinformatics*, 5(1):59–70, Mar 2004. 49, 50

[82] W. M. Claudino, A. Quattrone, L. Biganzoli, M. Pestrin, I. Bertini, and A. D. Leo. Metabolomics: Available results, current research projects in breast cancer, and future applications. *Journal of Clinical Oncology*, May 2007. 34

[83] D. Connolly. A pragmatic theory of reference for the web. In *Proceedings of the 15th International World Wide Web Conference (WWW2006)*, 2006. 51

[84] F. M. Couto, M. J. Silva, V. Lee, E. Dimmer, E. Camon, R. Apweiler, H. Kirsch, and D. Rebholz-Schuhmann. GOAnnotator: Linking protein GO annotations to evidence text. *Journal of Biomedical Discovery and Collaboration*, 1, 2006. 126

[85] D. R. Cox and N. Reid. *The Theory of the Design of Experiments*. Chapman & Hall/CRC, 2000. 132

[86] J. Davies, R. Studer, and P. Warren, editors. *Semantic Web Technologies: Trends and Research in Ontology-based Systems*. Wiley, 2006. 98

[87] R. Davis and W. Hamscher. Model-based reasoning: troubleshooting. pages 297–346, 1988. 10

[88] J. Day-Richter, M. Harris, M. Haendel, The Gene Ontology OBO-Edit Working Group, and S. Lewis. OBO-Edit—an ontology editor for biologists. *Bioinformatics*, 2007. 21, 103, 125

[89] M. J. L. de Hoon, B. Chapman, and I. Friedberg. Bioinformatics and computational biology with Biopython. *Genome Informatics*, 14:298–299, 2003. 44

[90] J. Demeter, C. Beauheim, J. Gollub, T. Hernandez-Boussard, H. Jin, D. Maier, J. C. Matese, M. Nitzberg, F. Wymore, Z. K. Zachariah, P. O. Brown, G. Sherlock, and C. A. Ball. The Stanford Microarray Database: implementation of new analysis tools and open source release of software. *Nucleic Acids Research*, 35(Database issue):D766–D770, Jan 2007. 37

[91] V. Devedžić. Understanding ontological engineering. *Communications of the ACM*, 45(4ve):136–144, 2002. 55, 69, 98

[92] M. Diehn, G. Sherlock, G. Binkley, H. Jin, J. C. Matese, T. Hernandez-Boussard, C. A. Rees, J. M. Cherry, D. Botstein, P. O. Brown, and A. A. Alizadeh. SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Research*, 31(1):219–223, Jan 2003. 43

[93] F. M. Donini. Complexity of reasoning. In *The Description Logic Handbook*. Cambridge University Press, 2003. 60

[94] S. Drăghici, S. Sellamuthu, and P. Khatri. Babel's tower revisited: a universal resource for cross-referencing across annotation databases. *Bioinformatics*, 22(23):2934–2939, Dec 2006. 59

[95] W. Dubitzky. Understanding the computational methodologies of systems biology. *Briefings in Bioinformatics*, 7(4):315–317, Dec 2006. 29

[96] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, 2000. 2

[97] G. M. Duyk. Sharper tools and simpler methods. *Nature Genetics*, 32 Suppl:465–468, Dec 2002. 33

[98] M. Ehrig, S. Handshuh, A. Hotho, A. Maedche, B. Motik, D. Oberle, C. Schmitz, S. Staab, L. Stojanovic, N. Stojanovic, R. Studer,

G. Stumme, Y. Sure, J. Tane, R. Volz, and V. Zacharias. The Karlsruhe view on ontologies. Technical report, Institute of Applied Informatics and Foral Description Methods (AIFB) University of Karlsruhe, D-76128 Karlsruhe, Germany, 2003. 54

[99] J. A. Eisen. Environmental shotgun sequencing: Its potential and challenges for studying the hidden world of microbes. *PLoS Biology*, 5(3):e82, 2007. 33

[100] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceeding of the National Academy of Science USA*, 95(25):14863–14868, Dec 1998. 3

[101] M. Ereshefsky. *The Poverty of the Linnaean Hierarchy: A Philosophical Study of Biological Taxonomy*. Cambridge Studies in Philosophy and Biology. Cambridge University Press, 2000. 113

[102] M. Ereshefsky. Species. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2006. 113, 156

[103] M. Ereshefsky and M. Matthen. Taxonomy, polymorphism and history: An introduction to population structure theory. *Philosophy of Science*, 72:1–21, 2005. 100

[104] T. Etzold and P. Argos. SRS—an indexing and retrieval tool for flat file data libraries. *Computational Applied Bioscience*, 9(1):49–57, Feb 1993. 41

[105] T. Etzold and P. Argos. Transforming a set of biological flat file libraries to a fast access network. *Computational Applied Bioscience*, 9(1):59–64, Feb 1993. 41

[106] T. Etzold and G. Verde. Using views for retrieving data from extremely heterogeneous databanks. *Proceedings of the Pacific Symposium on Biocomputing*, pages 134–141, 1997. 41

[107] J. Euzenat and P. Shvaiko. *Ontology matching*. Springer-Verlag, Heidelberg (DE), 2007. 59, 60

[108] W. J. Ewens and G. Grant. *Statistical Methods in Bioinformatics: An Introduction*. Springer Verlag, 2005. 28

[109] T. A. Eyre, F. Ducluzeau, T. P. Sneddon, S. Povey, E. A. Bruford, and M. J. Lush. The HUGO Gene Nomenclature Database, 2006 updates. *Nucleic Acids Research*, 34(Database issue):D319–D321, Jan 2006. 49

[110] E. Falkenberg, W. Hesse, P. Lindgreen, B. Nilsson, J. Oei, C. Rolland, R. Stamper, F. V. Assche, A. Verrijn-Stuart, and K. Voss. FRISCO : A Framework of Information System Concepts. Technical report, The IFIP WG 8.1 Task Group FRISCO, 1996. 79

[111] J.-B. Fan, M. S. Chee, and K. L. Gunderson. Highly parallel genomic assays. *Nature Reviews Genetics*, 7(8):632–644, Aug 2006. 32

[112] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27:861–874, 2006. 16

[113] N. Fedoroff, S. Racunas, and J. Shrager. Making biological computing smarter. *The Scientist*, 19(11):20–21. 27

[114] M. Fitting. Intensional logic. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2007. 88

[115] L. Floridi. Semantic conceptions of information. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2007. 30

[116] J. A. Fox, S. McMillan, and B. F. F. Ouellette. A compilation of molecular biology web servers: 2006 update on the Bioinformatics Links Directory. *Nucleic Acids Research*, 34(supplement 2):W3–5, 2006. 45

[117] A. U. Frank. Distinctions produce a taxonomic lattice: Are these units of mentalese? In *Proceedings of the 4th International Conference on Formal Ontology in Information Systems (FOIS2006)*, 2006. 169

[118] A. Gagnemi and V. Presutti. A grounded ontology for identity and reference of web resources. In *Proceedings of the 16th International World Wide Web Conference (WWW2007)*, 2007. 51

[119] M. Y. Galperin. The molecular biology database collection: 2004 update. *Nucleic Acids Research*, 32, January 2004. 37

[120] M. Y. Galperin. The molecular biology database collection: 2005 update. *Nucleic Acids Research*, 33(Database issue):D5–24, Jan 2005. 37

[121] M. Y. Galperin. The molecular biology database collection: 2006 update. *Nucleic Acids Research*, 34:D3–D5, 2006. 37, 124

[122] M. Y. Galperin. The molecular biology database collection: 2007 update. *Nucleic Acids Research*, 35(Database issue):D3–D4, Jan 2007. 37, 43, 161

[123] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider. Sweetening ontologies with dolce, 2002. 61

[124] M. R. Genesereth and N. J. Nilsson. *Logical Foundations of Artificial Intelligence*. Morgan Kaufman Publishers, 1987. 68

[125] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10):R80, 2004. 3

[126] M. T. Ghiselin. Species concepts: the basis for controversy and reconciliation. *Fish and Fisheries*, 3:151–160, 2002. 156, 162

[127] A. Ginsberg. The big schema of things. two philosophical visions of the relationship between language and reality and their implications for the semantic web. In *Proceedings of the 15th International World Wide Web Conference (WWW2006)*, 2006. 51, 77

[128] C. Goble. The state of the nation in data integration. In *Proceedings of the 16th International World Wide Web Conference (WWW2007)*, 2007. 47

[129] C. A. Goble, S. Pettifer, NewAuthor2, and C. Greenhalgh. Knowledge integration: In silico experiments in bioinformatics. In I. Foster and C. Kesselman, editors, *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufman, 2003. 51

[130] C. A. Goble, R. Stevens, G. Ng, S. Bechhofer, N. W. Paton, P. G. Baker, M. Peim, and A. Brass. Transparent access to multiple bioinformatics information sources. *IBM Systems Journal*, 40:532–551, 2001. 58

[131] S. Golder and B. A. Huberman. The structure of collaborative tagging systems. *Journal of the American Society for Information Science and Technology*, 2006. 59

[132] A. Gómez-Pérez. *Handbook on Ontologies*, chapter Ontology Evaluation, pages 251–273. Springer Verlag, 2004. 59

[133] A. Gómez-Pérez, M. Fernández-López, and O. Corcho. *Ontological Engineering*. Advanced Information and Knowledge Processing. Springer-Verlag London Limited, first edition, 2004. 12, 55, 66, 71, 72, 73, 98

[134] B. M. Good, E. M. Tranfield, P. C. Tan, M. Shehata, G. K. Singhera, J. Gosselink, E. B. Okon, and M. D. Wilkinson. Fast, cheap and out of control: a zero curation model for ontology development. *Pacific Symposium on Biocomputing*, pages 128–139, 2006. 59

[135] B. M. Good and M. D. Wilkinson. The Life Sciences Semantic Web is full of creeps! *Briefings in Bioinformatics*, 7(3):275–286, Sep 2006. 51

[136] P. Grenon. BFO in a nutshell: A bi-categorial axiomatization of BFO and comparison with DOLCE. Technical report, Institute for Formal Ontology and Medical Information Science, 2003. 61, 78, 86

[137] P. Grenon. Knowledge management from the ontological standpoint. In *Proceedings of the WM2003 Workshop on Knowledge Management and Philosophy*, 2003. 78, 86

[138] P. Grenon. Nuts in BFO's nutshell: Revisions to the bi-categorial axiomatization of BFO. Technical report, Institute for Formal Ontology and Medical Information Science, 2003. 78

[139] P. Grenon and B. Smith. Persistence and ontological pluralism. *Metaphysica*, 2007. 93

[140] P. Grenon, B. Smith, and L. Goldberg. *Ontologies in Medicine*, chapter Biodynamic Ontology: Applying BFO in the Biomedical Domain, pages 20–38. IOS Press, 2004. 61, 78, 86

[141] M. J. F. Grimnes. *ImageCreek: A knowledge level approach to case-based image representation*. PhD thesis, Norwegian University of Science and Technology, 1998. 10

[142] S. Grossman, S. Bauer, P. N. Robinson, and M. Vingron. Improved detection of overrepresentation of Gene-Ontology annotations with parent-child analysis. *Bioinformatics*, 23:3024–3031, 2007. 134

[143] T. R. Gruber. Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In N. Guarino and R. Poli, editors, *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Deventer, The Netherlands, 1993. Kluwer Academic Publishers. 55, 66

[144] T. R. Gruber. A translation approach to portable ontology specification. *Knowledge Acquisition*, 5(2):199–220, 1993. 67, 81

[145] M. Gu. *Knowledge-Intensive Conversational Case-Based Reasoning*. PhD thesis, Norwegian University of Science and Technology, 2006. 10

[146] N. Guarino. Understanding, building and using ontologies. In *Proceedings of the 10th Knowledge Acquisition for Knowledge-Based Systems Workshop (KAW1997)*, volume 46, pages 293–310, Duluth, MN, USA, 1997. Academic Press, Inc. 54, 63, 76

[147] N. Guarino. Formal ontology and information systems. In N. Guarino, editor, *Proceedings of the 1st International Conference on Formal Ontologies in Information Systems (FOIS98) Trento, Italy*, pages 3– 15. IOS Press, 1998. 68

[148] N. Guarino and P. Giaretta. Ontologies and knowledge bases: Towards a terminological clarification. In N. Mars, editor, *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing (KBKS'95)*, pages 25–32. University of Twente, Enschede, The Netherlands, IOS Press, Amsterdam, The Netherlands, 1995. 54, 63, 67, 77

[149] N. Guarino and M. A. Musen. Applied Ontology: Focusing on content. *Applied Ontology*, 1:1–5, 2005. 64

[150] N. Guarino and C. Welty. Evaluating ontological decisions with OntoClean. *Comunications of the ACM*, 45:61–65, 2002. 73, 165

[151] R. V. Guha and P. Hayes. Lbase: Semantics for languages of the semantic web, 2003. 144

[152] C.-C. Gunther, M. Langaas, S. Lydersen, V. Beisvåg, F. R. K. Junge, H. Bergum, and A. Lægreid. Statistical hypothesis testing of association between two reporter lists within the GO-hierarchy. In *European Meeting of Statisticians*, 2005. 134

[153] C.-C. Gunther, M. Langaas, S. Lydersen, V. Beisvåg, F. R. K. Junge, H. Bergum, and A. Lægreid. Statistical hypothesis testing of association between two lists of genes. In *Proceedings of the Workshoop on Statistics for Gene and Protein Expression*, 2006. 134

[154] D. Gusfield. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge University Press, 1997. 28

[155] M. A. Haendel, F. Neuhaus, D. S. Osumi-Sutherland, P. M. Mabee, J. José L. V. Mejino, C. J. Mungall, and barry Smith. *Anatomy Ontologies for Bioinformatics: Principles and Practice*, chapter CARO — The Common Anatomy Reference Ontology. 2007. 87

[156] H. Halpin. Identity, reference, and meaning on the web. In *Proceedings of the 15th International World Wide Web Conference (WWW2006)*, 2006. 51

[157] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(Database issue):D514–D517, Jan 2005. 53

[158] J. Han and M. Kaber. *Data Mining. Concepts and Techniques*. Morgan Kaufman, 2005. 2

[159] D. Hanisch, F. Sohler, and R. Zimmer. ToPNet—an application for interactive analysis of expression data and biological networks. *Bioinformatics*, 20(9):1470–1471, Jun 2004. 8

[160] E. R. Harold and W. S. Means. *XML in a Nutshell*. O'Reilly Media, Inc., 2004. 46

[161] P. Hayes. In defense of ambiguity. In *Proceedings of the WWW2006 Workshop on Architecture and Philosophy of the Web. Identity, Reference, and the Web (IRW2006)*, 2006. 77

[162] P. Hayes and C. Menzel. A semantics for the knowledge interchange format. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI2001)*, 2001. 73, 144

[163] M. L. Hekkelman and G. Vriend. MRS: a fast and compact retrieval system for biological data. *Nucleic Acids Research*, 33(Web Server issue):W766–W769, Jul 2005. 33, 41

[164] H. Hermjakob. The HUPO Proteomics Standards Initiative—overcoming the fragmentation of proteomics data. *Proteomics*, 6 Suppl 2:34–38, Sep 2006. 14

[165] P. Hieter and M. Boguski. Functional genomics: it's all how you read it. *Science*, 278(5338):601–602, Oct 1997. 2

[166] D. P. Hill, J. A. Blake, M. S. McAndrews-Hill, and B. Smith. Annotating genes to the Gene Ontology: Connections between experiments, genes and knowledge representation. 2007. (under review). 15, 21

[167] K. Hinkelmann and O. Kempthorne. *Design and Analysis of Experiments*. John Wiley and Sons, 2005. 132

[168] R. Hoffmann and A. Valencia. A gene network for navigating the literature. *Nature Genetics*, 36(7):664, Jul 2004. 8

[169] E. Hofsli, L. Thommesen, F. Yadetie, M. Langaas, W. Kuśnierczyk, U. Falkmer, A. K. Sandvik, and A. Lægreid. Identification of novel growth factor-responsive genes in neuroendocrine gastrointestinal tumour cells. *British Journal of Cancer*, 92:1506–1516, 2005. 5, 37

[170] T. Hofweber. Logic and ontology. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, Stanford, CA, winter 2003 edition, 2003. 67

[171] J. D. Hoheisel. Microarray technology: beyond transcript profiling and genotype analysis. *Nature Reviews Genetics*, 7(3):200–210, Mar 2006. 34, 37

[172] R. W. Hooft, C. Sander, M. Scharf, and G. Vriend. The PDBFINDER database: a summary of PDB, DSSP and HSSP information with added value. *Computational Applied Bioscience*, 12(6):525–529, Dec 1996. 36, 41

[173] S. Horst. The computational theory of mind. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2005. 80

[174] T. J. P. Hubbard, B. L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. C. Dyer, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, J. Herrero, R. Holland, K. Howe, K. Howe, N. Johnson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, C. Melsopp, K. Megy, P. Meidl, B. Ouverdin, A. Parker, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, J. Severin, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, M. Wood, T. Cox, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, P. Flicek, A. Kasprzyk, G. Proctor, S. Searle, J. Smith, A. Ureta-Vidal, and E. Birney. Ensembl 2007. *Nucleic Acids Research*, 35(Database issue):D610–D617, Jan 2007. 37, 44

[175] D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. R. Pocock, P. Li, and T. Oinn. Taverna: a tool for building and running workflows of services. *Nucleic Acids Research*, 34(Web Server issue):W729–W732, Jul 2006. 51

[176] P. J. Hunter and T. K. Borg. Integration from proteins to organs: the Physiome project. *Nature Reviews Molecular Cell Biology*, 4(3):237–243, Mar 2003. 27

[177] T. R. Hvidsten. *Predicting Function of Genes and Proteins from Sequence, Structure and Expression Data*. PhD thesis, Department of Information Technology, Uppsala University, 2004. 4

[178] T. R. Hvidsten, J. Komorowski, A. K. Sandvik, and A. Lægreid. Predicting gene function from gene expressions and ontologies. In *Pacific Symposium on Biocomputing*, pages 299–310, 2001. 4

[179] T. R. Hvidsten, A. Lægreid, and J. Komorowski. Learning rule-based models of biological process from gene expression time profiles using Gene Ontology. *Bioinformatics*, 19(9):1116–23, 2003. 4

[180] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, Oct 2004. 35

[181] A. D. Irvine. Russellś paradox. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Summer 2004. 73

[182] K. M. Jaszczolt. Defaults in semantics and pragmatics. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2006. 153

[183] T.-K. Jenssen, A. Lægreid, J. Komorowski, and E. Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28(1):21–28, May 2001. 8

[184] I. Johansson. Bioinformatics and biological reality. *Journal of Biomedical Informatics*, 39(3):274–287, Jun 2006. 78

[185] I. Johansson, B. Smith, K. Munn, N. Tsikolia, K. Elsner, D. Ernst, and D. Siebert. Functional anatomy: a taxonomic proposal. *Acta Biotheoretica*, 53:153–166, 2005. 117

[186] N. C. Jones and P. A. Pevzner. *An Introduction to Bioinformatics Algorithms*. MIT Press, 2004. 28

[187] J. R. Josephson and S. G. Josephson. *Abductive Inference. Computation, Philosophy, Technology*. Cambridge University Press, first edition, 1996. 31

[188] C. A. Joslyn, S. M. Mniszewski, A. Fulmer, and G. Heaton. The Gene Ontology Categorizer. *Bioinformatics*, 20 Suppl 1:i169–i177, Aug 2004. 135

[189] D. Jurafsky and J. H. Martin. *Speech and Language Processing*. Prentice-Hall, Inc., 2000. 52

[190] J. Kaiser. Public-private group maps out initiatives. *Science*, 296(5569):827, 2002. 78

[191] A. Kanapin, S. Batalov, M. Davis, J. Gough, S. Grimmond, H. Kawaji, M. Mgrane, H. Matsuda, C. Schönbach, and R. Teasdale. Mouse proteome analysis. *Genome Research*, 13:1335–1344, 2003. 103

[192] M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research*, 34(Database issue):D354–D357, Jan 2006. 10

[193] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. The KEGG resource for deciphering the genome. *Nucleic Acids Research*, 32(Database issue):D277–D280, Jan 2004. 10

[194] E. Kawas, M. Senger, and M. D. Wilkinson. BioMoby extensions to the Taverna workflow management and enactment software. *BMC Bioinformatics*, 7:523, 2006. 51

[195] J. Kelso, J. Visagie, G. Theiler, A. Christoffels, S. Bardien, D. Smedley, D. Otgaar, G. Greyling, C. V. Jongeneel, M. I. McCarthy, T. Hide, and W. Hide. eVOC: a controlled vocabulary for unifying gene expression data. *Genome Research*, 13(6A):1222–1230, Jun 2003. 96

[196] R. E. Kenyon, Jr. On the use of quotation marks. *A Review of General Semantics*, 51, 1994. 69

[197] S. Kerrien, Y. Alam-Faruque, B. Aranda, I. Bancarz, A. Bridge, C. Derow, E. Dimmer, M. Feuermann, A. Friedrichsen, R. Huntley, C. Kohler, J. Khadake, C. Leroy, A. Liban, C. Lieftink, L. Montecchi-Palazzi, S. Orchard, J. Risse, K. Robbe, B. Roechert, D. Thorneycroft, Y. Zhang, R. Apweiler, and H. Hermjakob. IntAct—open source resource for molecular interaction data. *Nucleic Acids Research*, 35(Database issue):D561–D565, Jan 2007. 38

[198] P. Khatri and S. Draghici. Ontological analysis of gene expression data: Current tools, limitations, and open problems. *Bioinformatics*, 21(18):25–29, 2005. 58

[199] P. Khatri, C. Voichita, K. Kattan, N. Ansari, A. Khatri, C. Georgescu, A. L. Tarca, and S. Draghici. Onto-Tools: new additions and improvements in 2006. *Nucleic Acids Research*, 35(Web Server issue):W206–W211, Jul 2007. 59

[200] D. Khlentzos. Semantic challenges to realism. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2004. 78

[201] M. E. I. Kipp and D. G. Campbell. Patterns and inconsistencies in collaborative tagging systems: An examination of tagging practices, 2006. 59

[202] H. Kitano. Computational systems biology. *Nature*, 420(6912):206–210, Nov 2002. 8, 29, 32

[203] H. Kitano. Systems biology: a brief overview. *Science*, 295(5560):1662–1664, Mar 2002. 29, 32

[204] G. Klima. The medieval problem of universals. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2004. 92

[205] K. Kline, D. Kline, and B. Hunt. *SQL In A Nutshell*. O'Reilly Media, Inc., 2004. 43

[206] S. Knudsen. *Analysis of DNA Microarray Data*. John Wiley and Sons, Inc., 2002. 36

[207] I. S. Kohane, A. T. Kho, and A. J. Butte. *Microarrays for an Integrative Genomics*. MIT Press, 2003. 36

[208] J. Köhler, S. Philippi, and M. Lange. SEMEDA: ontology based semantic integration of biological databases. *Bioinformatics*, 19(18):2420–2427, Dec 2003. 58

[209] J. Köhler, A. Rüegg, A. Skusa, and B. Smith. Quality control for terms and definitions in ontologies and taxonomies. *BMC Bioinformatics*, 7:212–220, 2006. 133

[210] J. Kolodner. *Case-Based Reasoning*. Morgan Kaufmann Publishers, 1993. 10

[211] R. Koons. Defeasible reasoning. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2005. 153

[212] K. Korta and J. Perry. Pragmatics. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2006. 77

[213] D. E. Krane and M. L. Raymer. *Fundamental Concepts of Bioinformatics*. Pearson Education, Inc., 2003. 28

[214] T. Kulikova, R. Akhtar, P. Aldebert, N. Althorpe, M. Andersson, A. Baldwin, K. Bates, S. Bhattacharyya, L. Bower, P. Browne, M. Castro, G. Cochrane, K. Duggan, R. Eberhardt, N. Faruque, G. Hoad, C. Kanz, C. Lee, R. Leinonen, Q. Lin, V. Lombard, R. Lopez, D. Lorenc, H. McWilliam, G. Mukherjee, F. Nardone, M. P. G. Pastor, S. Plaister, S. Sobhany, P. Stoehr, R. Vaughan, D. Wu, W. Zhu, and R. Apweiler. EMBL Nucleotide Sequence Database in 2006. *Nucleic Acids Research*, 35(Database issue):D16–D20, Jan 2007. 38

[215] A. Kumar and B. Smith. The unified medical language system and the gene ontology: Some critical reflections. In A. Günter, R. Kruse, and B. Neumann, editors, *KI2003: Advances in AI*, pages 135–148. 2003. 88, 133, 135

[216] A. Kumar and B. Smith. Enhancing GO for the sake of clinical bioinformatics. In *Proceedings of the Bio-Ontologies Workshop (ISMB2004), Glasgow*, 2004. 133

[217] P. Küngas. *Distributed Agent-Based Web Service Selection, Composition and Analysis Through Partial Deduction*. PhD thesis, Norwegian University of Science and Technology, 2006. 52

[218] W. Kuśnierczyk. Nontological engineering. In B. Bennett and C. Fellbaum, editors, *Proceedings of the International Conference on Formal Ontology in Information Systems*, volume 150 of *Frontiers in Artificial Intelligence and Applications*, pages 39–50. IOS Press, 2006. 13, 54, 70, 75, 77, 95, 129, 144

[219] W. Kuśnierczyk. The logic of relations between the Gene Ontology and the Taxonomy of Species. In *Proceedings of the 10th Bio-Ontologies SIG Workshop*, 2007. 13

[220] W. Kuśnierczyk. Taxonomic partitioning of the Gene Ontology. In *Proceedings of the Dagstuhl Seminar Towards Interoperability of Biomedical Ontologies*, 2007. 13

[221] W. Kuśnierczyk. Taxonomy-based partitioning of the Gene Ontology. In *Journal of Biomedical Informatics*, 2007. To appear. 13

[222] W. Kuśnierczyk. What does a GO annotation mean? In *Proceedings of the 10th Bio-Ontologies SIG Workshop*, 2007. 15

[223] W. Kuśnierczyk, A. Aamodt, and A. Lægreid. Knowledge-intensive case-based support for automated explanation of biological phenomena. In *Proceedings of the 6th International Conference on Case-Based Reasoning (ICCBR)*, 2005. 11

[224] W. Kuśnierczyk, A. Lægreid, and A. Aamodt. Towards automated explanation of gene-gene relationships. In *Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, 2004. 11

[225] W. Kuśnierczyk and E. M. Sonnervik. Learning yest gene function from expression programs and gene ontology. In *Proceedings of the Computer Science Graduate Student Conference (CSGSC2002)*, 2002. 5, 17

[226] A. Lægreid, T. R. Hvidsten, H. Midelfart, J. Komorowski, and A. K. Sandvik. Predicting Gene Ontology biological process from temporal gene expression patterns. *Genome Research*, 13(5):965–79, 2003. 4

[227] P. Lambrix. *Artificial Intelligence Methods and Tools for Systems Biology*, chapter Ontologies in Bioinformatics and Systems Biology. Springer, 2004. 59

[228] P. Lambrix, M. Habbouche, and M. Pérez. Evaluation of ontology development tools for bioinformatics. *Bioinformatics*, 19(12):1564–1571, 2003. 59

[229] P. Lambrix and H. Tan. SAMBO—a system for aligning and merging biomedical ontologies. *Journal of Web Semantics*, 4:196–206, 2006. 60

[230] M. Langaas, B. H. Lindqvist, and E. Ferkingstad. Estimating the proportion of true null hypotheses, with application to DNA microarray data. *Journal of The Royal Statistical Society Series B-statistical Methodology*, 67, 2005. 132

[231] P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armañanzas, G. Santafé, A. Pérez, and V. Robles. Machine learning in bioinformatics. *Briefings in Bioinformatics*, 7(1):86–112, Mar 2006. 29

[232] O. Lassila and D. McGuinness. The role of frame-based representation on the Semantic Web. KSL-01-02. Technical report, Knowledge Systems Laboratory, Stanford University, Stanford, Califor-

nia, 2001. http://www.ida.liu.se/ext/epa/ej/etai/2001/018/01018-etaibody.pdf. 57, 68, 69, 70

[233] V. Lee, E. Camon, E. Dimmer, D. Barrell, and R. Apweiler. Who tangos with GOA? — use of Gene Ontology Annotation (GOA) for biological interpretation of '-omics' data and for validation of automatic annotation tools. *In Silico Biology*, 5(1):5–8, 2005. 58

[234] G. Leech and M. Weisser. *The Oxford Handbook of Computational Linguistics*, chapter Pragmatics and Dialogue, pages 136–156. Oxford University Press, 2004. 77

[235] G. R. Librelotto, M. Martins, and H. Machado. Topic Maps applied to PubMed. In *Proceedings of the Markup Theory and Practice Conference, Montréal, Canada*, 2007. 69

[236] T. G. Lilburn, S. H. Harrison, J. R. Cole, and G. M. Garrity. Computational aspects of systematic biology. *Briefings in Bioinformatics*, 7(2):186–195, Jun 2006. 29

[237] B. Liu, S. Li, and J. Hu. Technological advances in high-throughput screening. *Am J Pharmacogenomics*, 4(4):263–276, 2004. 33

[238] C. M. Lloyd, M. D. B. Halstead, and P. F. Nielsen. CellML: its future, present and past. *Progress in Biophysics and Molecular Biology*, 85(2-3):433–450, 2004. 26

[239] D. J. Lockhart and E. A. Winzeler. Genomics, gene expression and dna arrays. *Nature*, 405(6788):827–836, Jun 2000. 33

[240] H. J. Lowe and G. O. Barnett. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *JAMA*, 271(14):1103–1108, 1994. 58

[241] P. M. Mabee et al. Phenotype ontologies: The bridge between genomics and evolution. *Trends in Ecology and Evolution*, 22:1222–1230, 2007. 96

[242] N. Markosian. Time. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2002. 93

[243] S. Martin, M. M. Hohman, and T. Liefeld. The impact of Life Science Identifier on informatics data. *Drug Discovery Today*, 10(22):1566–1572, Nov 2005. 51

[244] C. Masolo, S. Borgo, A. Gagnemi, N. Guarino, and A. Oltramari. WonderWeb Deliverable D18. Technical report, Laboratory for Applied Ontology — ISTC-CNR, 2003. 12, 61, 65, 78, 90

[245] J. P. Massar, M. Travers, J. Elhai, and J. Shrager. Biolingua: a programmable knowledge environment for biologists. *Bioinformatics*, 21(2):199–207, Jan 2005. 44

[246] R. L. Mayden. A hierarchy of species concepts: the denouement in the saga of the species problem. In M. F. Claridge, H. A. Dawah, and M. R. Wilson, editors, *Species: the Units of Biodiversity*, pages 381–424. Chapman & Hall Ltd., 1997. 163

[247] R. L. Mayden. On biological species, species concepts and individuation in the natural world. *Fish and Fisheries*, 3:171–196, 2002. 156, 163

[248] A. T. McCray. Conceptualizing the world: lessons from history. *Journal of Biomedical Informatics*, 39(3):267–273, Jun 2006. 12

[249] J. W. McGuffee. Programming languages and the biological sciences. *Journal of Computing Sciences in Colleges*, 22(4):178–183, 2007. 44

[250] L. M. McShane, M. D. Radmacher, B. Freidlin, R. Yu, M.-C. Li, and R. Simon. Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics*, 18(11):1462–1469, Nov 2002. 3

[251] C. C. Mello and D. Conte. Revealing the world of RNA interference. *Nature Insight. RNA Interference*, 431(7006):338–342, 2004. 34

[252] C. Menzel. Actualism. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2007. 88

[253] H. Midelfart. *Knowledge Discovery from cDNA Microarrays and* a priori *Knowledge*. PhD thesis, Norwegian University of Science and Technology, 2003. 4, 6, 134

[254] H. Midelfart, W. Kuśnierczyk, T. R. Hvidsten, A. Lægreid, and J. Komorowski. Learning yeast gene function from expression programs and Gene Ontology. In *Conference Proceedings of the Norwegian Biochemical Society Winter Meeting 2002*, 2002. 4, 5, 6

[255] A. Miller. Realism. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2005. 78, 88

[256] B. Miller. Existence. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Summer 2002. 88

[257] T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997. 2

[258] R. Mizoguchi. A step towards ontological engineering. In *Proceedings of the 12th National Conference on AI of JSAI*, pages 24–31, 1998. 55, 98

[259] G. E. Moore. Cramming more components onto integrated circuits. *Electronics*, 38, 1965. 36

[260] C. Mungall. Obol:integrating language and meaning in bio-ontologies. *Comparative and Functional Genomics*, 5:509–520, 2004. 112, 136

[261] S. Myhre, H. Tveit, T. Mollestad, and A. Lægreid. Additional Gene Ontology structure for improved biological reasoning. *Bioinformatics*, 2006. 112

[262] D. Nardi and R. J. Brachman. *The Description Logic Handbook*, chapter An Introduction to Description Logics, pages 1–40. Cambridge University Press, 2003. 71, 146

[263] B. Nebel. *Reasoning and Revision in Hybrid Representation Systems*, volume 422 of *Lecture Notes in Artificial Intelligence*. Springer Verlag, 1990. 146

[264] F. Neuhaus, P. Grenon, and B. Smith. A formal theory of substances, qualities, and universals. In A. Varzi and L. Vieu, editors, *Formal Ontology and Information Systems. Proceedings of the International Conference (FOIS 2004)*. IOS Press, 2004. 89, 90, 91

[265] F. Neuhaus and B. Smith. *Anatomy Ontologies for Bioinformatics: Principles and Practice*, chapter Relations in Anatomical Ontologies. 2007. 130

[266] A. Nikitin, S. Egorov, N. Daraselia, and I. Mazo. Pathway studio–the analysis and navigation of molecular networks. *Bioinformatics*, 19(16):2155–2157, Nov 2003. 8

[267] K. Nørsett, T. Bruland, O. Ween, E. Hofsli, L. Thommesen, K. Misund, T. Strømmen, W. Kuśnierczyk, A. Aamodt, H. Midelfart, T. R. Hvidsten, et al. Systems biology of the normal and diseased gastrointestinal system. In *Proceedings of the 2nd ESF Functional Genomics Conference*, 2005. 5

[268] K. Nørsett, W. Kuśnierczyk, A. Lægreid, M. Langaas, H. Waldum, and A. K. Sandvik. Gene expression profiles in gastric mucosa during therapeutic acid inhibition. In *Proceedings of the 1st ESF Functional Genomics and Disease Conference*, 2003. 5

[269] K. Nørsett, W. Kuśnierczyk, A. Lægreid, M. Langaas, F. Yadetie, S. Kvam, S. Wørlund, H. Waldum, and A. K. Sandvik. Gene expression profiles in hypergastrinemic rats and patients. In *Proceedings of the 2003 Winter Meeting of the Norwegian Biochemical Society*, 2003. 5

[270] K. G. Nørsett, A. Lægreid, W. Kuśnierczyk, M. Langaas, S. Ylving, R. Fossmark, S. Myhre, S. Falkmer, H. L. Waldum, and A. K. Sandvik. Changes in gene expression of gastric mucosa during therapeutic acid inhibition. *European Journal of Gastroenterology and Hepatology*, 2007. (under review). 5

[271] N. F. Noy and D. L. McGuinness. Ontology development 101: A guide to creating your first ontology. 60

[272] C. K. Ogden and I. A. Richards. *The Meaning of Meaning*. 1930. 79, 84, 85

[273] A. Øhrn. *Discernibility and Rough Sets in Medicine: Tools and Applications*. PhD thesis, Norwegian University of Science and Technology, 1999. 4

[274] P. Öztürk. *A Knowledge Level Model of Context Use in Diagnotic Domains*. PhD thesis, Norwegian University of Science and Technology, 2000. 10

[275] C. Pan, J. Kim, L. Chen, Q. Wang, and C. Lee. The HIV positive selection mutation database. *Nucleic Acids Research*, 35(Database issue):D371–D375, Jan 2007. 37

[276] H. Parkinson, M. Kapushesky, M. Shojatalab, N. Abeygunawardena, R. Coulson, A. Farne, E. Holloway, N. Kolesnykov, P. Lilja, M. Lukk, R. Mani, T. Rayner, A. Sharma, E. William, U. Sarkans, and A. Brazma. ArrayExpress–a public database of microarray experiments and gene expression profiles. *Nucleic Acids Research*, 35(Database issue):D747–D750, Jan 2007. 38, 47

[277] B. Parsia and P. F. Patel-Schneider. Meaning and the semantic web. In *Proceedings of the 13th International World Wide Web Conference (WWW2004)*, 2004. 51, 77

[278] Z. Pawlak. Rough sets. *International Journal of Computer and Information Sciences*, 11:341–356, 1982. 4

[279] H. Pearson. What is a gene. *Nature*, 441:399–401, 2006. 57

[280] C. Perez-Iratxeta, M. A. Andrade-Navarro, and J. D.Wren. Evolving research trends in bioinformatics. *Briefings in Bioinformatics*, 8:88–95, 2006. 28

[281] P. A. Pevzner. *Computational Molecular Biology: An Algorithmic Approach*. MIT Press, 2001. 28

[282] K. M. Pickett. The new and improved PhyloCode, now with types, ranks, and even polyphyly: A conference report from the First International Phylogenetic Nomenclature Meeting. *Cladistics*, 21:79–82, 2005. 170

[283] D. Pitt. Mental representation. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2007. 80

[284] E. Plaza and J. L. Arcos. *Overview of Νονσ v. 1.0. Draft*. Institut d'Investigació en Intelligencia Artificial. 72

[285] M. Pocock, T. Down, and T. Hubbard. Biojava: open source components for bioinformatics. *SIGBIO Newsl.*, 20(2):10–12, 2000. 44

[286] R. L. Poidevin. The experience and perception of time. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2004. 93

[287] S. Povey, R. Lovering, E. Bruford, M. Wright, M. Lush, and H. Wain. The HUGO Gene Nomenclature Committee (HGNC). *Human Genetics*, 109(6):678–680, Dec 2001. 49

[288] J. Quackenbush. Data standards for 'omic' science. *Nature Biotechnology*, 22(5):613–614, May 2004. 39

[289] J. Quackenbush, C. Stoeckert, C. Ball, A. Brazma, R. Gentleman, W. Huber, R. Irizarry, M. Salit, G. Sherlock, P. Spellman, and N. Winegarden. Top-down standards will not serve systems biology. *Nature*, 440(7080):24, Mar 2006. 48

[290] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*, 14(8):656–664, 1998. 27

[291] A. 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, Jun 2006. 12

[292] A. L. Rector. Modularisation of domain ontologies implemented in Description Logics and related formalisms including OWL. In *Proceedings of the 2nd International Conference on Knowledge Capture*, 2003. 98

[293] A. L. Rector, W. A. Nowlan, and S. Kay. Foundations for an electronic medical record. *Methods of Information in Medicine*, 30(3):179–186, Aug 1991. 83

[294] M. Reicher. Nonexistent objects. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2006. 88

[295] M. Reimer. Reference. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2006. 79

[296] R. Reiter. On closed world data bases. In H. Gallaire and J. Minker, editors, *Logic and Data Bases*, 1978. 127, 140

[297] O. Ritter, P. Kocab, M. Senger, D. Wolf, and S. Suhai. Prototype implementation of the integrated genomic database. *Computers and Biomedical Research*, 27(2):97–115, Apr. 1994. 43

[298] D. Robb and J. Heil. Mental causation. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2005. 84

[299] A. Robbins. *Unix in a Nutshell*. O'Reilly Media Inc., 2005. 40

[300] D. S. Roos. Bioinformatics — trying to swim in a sea of data. *Science*, 291(5507):1260–1261, February, 16 2001. 38

[301] G. Rosen. Abstract objects. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2006. 88

[302] C. Rosse and J. Mejino. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *Journal of Biomedical Informatics*, 36:478–500, 2003. 77, 96

[303] R. Sætre. *GeneTUC: Natural Language Understanding in Medical Text*. PhD thesis, Norwegian University of Science and Technology, 2006. 18, 133

[304] M. Safran, I. Solomon, O. Shmueli, M. Lapidot, S. Shen-Orr, A. Adato, U. Ben-Dor, N. Esterman, N. Rosen, I. Peter, T. Olender, V. Chalifa-Caspi, and D. Lancet. GeneCards 2002: towards a complete, object-oriented, human gene compendium. *Bioinformatics*, 18(11):1542–1543, Nov 2002. 8, 43

[305] J. Sampson. *A Comprehensive Framework for Ontology Alignment Quality*. PhD thesis, Norwegian University of Science and Technology, 2007. 60

[306] S.-A. Sansone, T. Fan, R. Goodacre, J. L. Griffin, N. W. Hardy, R. Kaddurah-Daouk, B. S. Kristal, J. Lindon, P. Mendes, N. Morrison, B. Nikolau, D. Robertson, L. W. Sumner, C. Taylor, M. van der Werf, B. van Ommen, and O. Fiehn. The Metabolomics Standards Initiative. *Nature Biotechnology*, 25(8):846–848, Aug 2007. 14

[307] U. Sauer, M. Heinemann, and N. Zamboni. Getting closer to the whole picture. *Science*, 316(5824):550–551, Apr 2007. 32

[308] M. Schena, editor. *Microarray Biochip Technology*. Eaton Publishing, 2000. 36

[309] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, Oct 1995. 33

[310] A. Schlicker, J. Rahnenfi¿½hrer, M. Albrecht, T. Lengauer, and F. Domingues. Gotax: investigating biological processes and biochemical activities along the taxonomic tree. *Genome Biol*, 8(3):R33, Mar 2007. 124

[311] D. Schober, W. Kuśnierczyk, S. E. Lewis, J. Lomax, members of the MSI and PSI Ontology Working Groups, C. Mungall, P. Rocca-Serra, B. Smith, and S.-A. Sansone. Towards naming conventions for use in controlled vocabulary and ontology engineering. In *Proceedings of the 10th Bio-Ontologies SIG Workshop, Conference on Intelligent Systems for Molecular Biology and European Conference on Computational Biology (ISMB/ECCB2007), Vienna, Austria*, 2007. 13, 70, 78

[312] S. Schulze-Kremer, B. Smith, and A. Kumar. Revising the UMLS Semantic Network. In *Annual Symposium of the American Medical Informatics Association*, 2004. 69, 135

[313] Y. M. Shah, K. Morimura, Q. Yang, T. Tanabe, M. Takagi, and F. J. Gonzalez. Peroxisome proliferator-activated receptor alpha regulates a microRNA-mediated signaling cascade responsible for hepatocellular proliferation. *Molecular Cell Biology*, 27(12):4238–4247, Jun 2007. 131

[314] S. C. Shapiro. *Encyclopedia of Artificial Intelligence*, chapter Artificial Intelligence, pages 54–57. John Wiley and Sons, 1992. 52

[315] J. Shendure, R. D. Mitra, C. Varma, and G. M. Church. Advanced sequencing technologies: methods and goals. *Nature Reviews Genetics*, 5(5):335–344, May 2004. 32

[316] R. Simon, M. Mirlacher, and G. Sauter. Tissue microarrays. *Biotechniques*, 36(1):98–105, Jan 2004. 34

[317] R. Simon, M. Mirlacher, and G. Sauter. Tissue microarrays. *Methods in Molecular Medicine*, 114:257–268, 2005. 34

[318] B. Smith. *Blackwell Guide to the Philosophy of Computing and Information*, chapter Ontology, pages 155–166. Blackwell, 2003. 39, 61, 64, 65, 86, 94, 133

[319] B. Smith. Beyond concepts: Ontology as reality representation. In A. Varzi and L. Vieu, editors, *Proceedings of the International Conference on Formal Ontology and Information Systems, FOIS2004*, 2004. 12, 70, 76, 77

[320] B. Smith. *Experience and Analysis*, chapter Against Fantology, pages 153–170. HPT & ÖBV, Vienna, 2005. 76

[321] B. Smith. Against idiosyncrasy in ontology development. In *Formal Ontology in Information Science (FOIS'2006)*, 2006. 12, 99, 165

[322] B. Smith. From concepts to clinical reality: An essay on the benchmarking of biomedical terminologies. *Journal of Biomedical Informatics*, 39:288–298, 2006. 12, 76

[323] B. Smith. Ontologies for biomedicine — how to make and use them. Proceedings of the 15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) and 6th European Conference on Computational Biology (ECCB), 2007. Tutorial. 27, 57, 59

[324] B. Smith, M. Ashburner, W. Ceusters, C. Mungall, M. Musen, N. Shah, J. Bard, K. Eilbeck, tFW Group, N. Leontis, et al. The OBO Foundry: Remoulding ontology to support data integration. *Nature Biotechnology*, 2007. 95, 98

[325] B. Smith and W. Ceusters. HL7 RIM: An incoherent standard. *Studies in Health Technology and Informatics*, 124:133–138, 2006. 80

[326] 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:R46, 2005. 87, 92, 95, 112, 116, 135

[327] B. Smith, W. Ceusters, and R. Temmerman. Wüsteria. In *Proceedings of Medical Informatics Europe*, 2005. 70, 84, 85

[328] B. Smith and P. Grenon. The cornucopia of formal-ontological relations. *Dialectica*, 58:279–296, 2004. 61, 86, 117

[329] B. Smith, J. Köhler, and A. Kumar. On the application of formal principles to life science data: A case study in the Gene Ontology. In *DILS 2004 - International Workshop on Data Integration in the Life Sciences*, 2004. 12, 133

[330] B. Smith and A. Kumar. On controlled vocabularies in bioinformatics: A case study in the Gene Ontology. *Biosilico: Drug Discovery Today*, 2:146–252, 2004. 133

[331] B. Smith, W. Kuśnierczyk, D. Schober, and W. Ceusters. Towards a coherent terminology for principles-based ontology. In *Proceedings of the 2nd International Workshop on Formal Biomedical Knowledge Representation, KRMED2006*, 2006. 13, 54, 70, 77, 82, 83, 92, 129, 144

[332] B. Smith and C. Rosse. The role of foundational relations in the alignment of biomedical ontologies. In M. Fieschi et al., editors, *Proceedings of MedInfo 2004*, pages 444–448. IOS Press, Amsterdam, 2004. 117

[333] B. Smith and C. Welty. Ontology: Towards a new synthesis. In *Proceedings of the Second International Conference on Formal Ontology in Information Systems (FOIS2001)*, 2001. 11

[334] B. Smith, J. Williams, and S. Schulze-Kremer. The ontology of the Gene Ontology. pages 609–613, 2003. 12

[335] D. W. Smith. Phenomenology. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2005. 82

[336] M. K. Smith, C. Welty, and D. McGuinness. *OWL Web Ontology Language guide. W3C Recommendation 10 February 2004*, 2004. 71

[337] G. K. Smyth, J. Michaud, and H. S. Scott. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, 21(9):2067–2075, May 2005. 3

[338] F. Sohler, D. Hanisch, and R. Zimmer. New methods for joint analysis of biological networks and expression data. *Bioinformatics*, 20(10):1517–1521, Jul 2004. 8

[339] F. Sørmo. *Case-Based Tutoring with Concept Maps*. PhD thesis, Norwegian University of Science and Technology, 2007. 10

[340] J. F. Sowa. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks Cole publishing Co., Pacific Grove, California, 2000. 11, 61, 71, 79

[341] K. A. Spackman and G. Reynoso. Examining SNOMED from the perspective of formal ontological principles. In *Workshop on Formal Biomedical Knowledge Representation (KR-MED2004)*, 2004. 78

[342] A. D. Spear. *Ontology for the Twenty First Century: An Introduction with Recommendations*. http://www.ifomis.uni-saarland.de/bfo/manual/manual.pdf. 60, 61, 69, 76, 78

[343] P. T. Spellman, M. Miller, J. Stewart, C. Troup, U. Sarkans, S. Chervitz, D. Bernhart, G. Sherlock, C. Ball, M. Lepage, M. Swiatek, W. L. Marks, J. Goncalves, S. Markel, D. Iordan, M. Shojatalab, A. Pizarro, J. White, R. Hubley, E. Deutsch, M. Senger, B. J. Aronow, A. Robinson, D. Bassett, C. J. Stoeckert, and A. Brazma. Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biology*, 3(9), Aug 2002. 47

[344] S. Staab and R. Studer, editors. *Handbook on Ontologies*. Springer-Verlag Berlin Heidelberg, first edition, 2004. 66

[345] J. E. Stajich, D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz, C. Dagdigian, G. Fuellen, J. G. R. Gilbert, I. Korf, H. Lapp, H. Lehvï¿½slaiho, C. Matsalla, C. J. Mungall, B. I. Osborne, M. R. Pocock, P. Schattner, M. Senger, L. D. Stein, E. Stupka, M. D. Wilkinson, and E. Birney. The Bioperl toolkit: Perl modules for the life sciences. *Genome Research*, 12(10):1611–1618, Oct 2002. 44, 45

[346] J. E. Stajich and H. Lapp. Open source tools and toolkits for bioinformatics: significance, and where are we? *Briefings in Bioinformatics*, 7(3):287–296, Sep 2006. 44

[347] L. D. Stein. Integrating biological databases. *Nature Reviews Genetics*, 4(5):337–345, May 2003. 44, 57

[348] M. Steup. Epistemology. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2006. 30

[349] M. Steup. The analysis of knowledge. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2006. 30

[350] R. Stevens, P. Baker, S. Bechhofer, G. Ng, A. Jacoby, N. W. Paton, C. A. Goble, and A. Brass. TAMBIS: transparent access to multiple bioinformatics information sources. *Bioinformatics*, 16(2):184–185, Feb 2000. 58

[351] R. Stevens, C. Goble, and S. Bechhofer. Ontology-based knowledge representation for bioinformatics. *Briefings in Bioinformatics*, 1:389–414, 2000. 54, 57

[352] R. Stevens, C. Wroe, P. Lord, and C. Goble. *Ontologies in Bioinformatics*, chapter Ontologies in Bioinformatics, pages 635–657. Springer Verlag, 2004. 54, 57

[353] R. D. Stevens, A. J. Robinson, and C. A. Goble. myGrid: Personalised bioinformatics on the information grid. *Bioinformatics*, 19 Suppl 1:i302–i304, 2003. 51

[354] M. Stevenson and Y. Wilks. *The Oxford Handbook of Computational Linguistics*, chapter Word-Sense Disambiguation, pages 249–265. Oxford University Press, 2004. 77

[355] J. D. Storey and R. Tibshirani. Statistical methods for identifying differentially expressed genes in DNA microarrays. *Methods in Molecular Biology*, 224:149–157, 2003. 132

[356] J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proceeding of the National Academy of Science USA*, 100(16):9440–9445, Aug 2003. 132

[357] L. Strömback, D. Hall, and P. Lambrix. A review of standards for data exchange within systems biology. *Proteomics*, 7:857–867, 2007. 48

[358] R. Studer, V. R. Benjamins, and D. Fensel. Knowledge engineering: Principles and methods. *Data Knowledge Engineering*, 25(1-2):161–197, 1998. 70

[359] B. P. Suomela and M. A. Andrade. Ranking the whole MEDLINE database according to a large training set using text indexing. *BMC Bioinformatics*, 6:75, 2005. 58

[360] B. Swartout, P. Ramesh, K. Knight, and T. Russ. Toward distributed use of large-scale ontologies. In A. Farquhar, M. Gruninger, A. Gómez-Pérez, M. Uschold, and P. van der Vet, editors, *AAAI'97 Symposium on Ontological Engineering, Stanford University, California, USA*, pages 138–148, 1997. 68

[361] C. Swoyer. Properties. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2000. 92

[362] F. Tai and W. Pan. Incorporating prior knowledge of gene functional groups into regularized discriminant analysis of microarray data. *Bioinformatics*, 23:3170–3177, Oct 2007. 133

[363] J. Tamames and A. Valencia. The success (or not) of HUGO nomenclature. *Genome Biology*, 7(5):402, 2006. 49

[364] P. Thagard. Cognitive science. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Summer 2007. 80

[365] The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000. 136

[366] The Gene Ontology Consortium. The Gene Ontology database and informatics resource. *Nucleic Acids Research*, 32:D258–D261, 2004. 56, 98

[367] The Gene Ontology Consortium. The Gene Ontology (GO) project in 2006. *Nucleic Acids Research*, 34:D322–D326, 2006. 56, 98, 100

[368] The UniProt Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Research*, 35(Database issue):D193–D197, Jan 2007. 40

[369] N. J. Thomas. Mental imagery. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2007. 80

[370] S. Tobies. *Complexity Results and Practical Algorithms for Logics in Knowledge Representation*. PhD thesis, Rheinisch-Westfalisch Technisch Hochschule Aachen, 2001. 60

[371] O. Tuason, L. Chen, H. Liu, J. A. Blake, and C. Friedman. Biological nomenclatures: a source of lexical knowledge and ambiguity. *Pacific Symposium on Biocomputing*, pages 238–249, 2004. 77

[372] J. Tuikkala, L. Flo, O. Nevalainen, and T. Aittokallio. Improving missing values estimation in microarray data with gene ontology. *Bioinformatics*, 22(5):556–572, 2006. 103

[373] G. F. Turner. Parallel speciation, despeciation and respeciation: implications for species definition. *Fish and Fisheries*, 3:225–229, 2002. 163

[374] M. Uschold and M. Grüninger. Ontologies: principles, methods, and applications. *Knowledge Engineering Review*, 11(2):93–155, 1996. 69

[375] K. Verspoor, J. Cohn, S. Mniszewski, and C. Joslyn. A categorization approach to automated ontological function annotation. *Protein Science*, 15(6):1544–1549, Jun 2006. 135

[376] A. Visel, S. Minovitsky, I. Dubchak, and L. A. Pennacchio. Vista enhancer browser–a database of tissue-specific human enhancers. *Nucleic Acids Research*, 35(Database issue):D88–D92, Jan 2007. 37

[377] H. M. Wain, E. A. Bruford, R. C. Lovering, M. J. Lush, M. W. Wright, and S. Povey. Guidelines for human gene nomenclature. *Genomics*, 79(4):464–470, April 2002. 49

[378] H. M. Wain, M. J. Lush, F. Ducluzeau, V. K. Khodiyar, and S. Povey. Genew: the Human Gene Nomenclature Database, 2004 updates. *Nucleic Acids Research*, 32(Database issue):D255–D257, Jan 2004. 49

[379] M. S. Waterman. *Introduction to Computational Biology: Maps, Sequences and Genomes*. Chapman Hall, New York, 1995. 28

[380] K. Waters. Molecular genetics. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2007. 32

[381] C. Welty. The ontological nature of subject taxonomies. In N. Guarino, editor, *Proceedings of the 1998 International Conference on Formal Ontology in Information Systems (FOIS'98)*. IOS Press, 1998. 71, 72

[382] E. Werner. All systems go, 2007. Book review. 29

[383] L. Wetzel. Types and tokens. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Summer 2006. 89

[384] D. B. Wheeler, A. E. Carpenter, and D. M. Sabatini. Cell microarrays and RNA interference chip away at gene function. *Nature Genetics*, 37 Suppl:S25–S30, Jun 2005. 34

[385] D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, L. Y. Geer, Y. Kapustin, O. Khovayko, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, J. Ostell, V. Miller, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R. L. Tatusov, T. A. Tatusova, L. Wagner, and E. Yaschenko. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 35(Database issue):D5–12, Jan 2007. 37, 38, 160

[386] R. L. Winslow and M. S. Boguski. Genome informatics: current status and future prospects. *Circulation Research*, 92(9):953–961, May 2003. 2

[387] I. H. Witten and E. Frank. *Data Mining. Practical Machine Learning Tools and Techniques*. Morgan Kaufman, 2005. 2

[388] J. A. Wohlschlegel, E. S. Johnson, S. I. Reed, and J. R. Yates. Global analysis of protein sumoylation in *Saccharomyces cerevisiae*. *Journal of Biological Chemistry*, 279(44):45662–45668, 2004. 103

[389] J. Woleński. Reism. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2004. 88

[390] C. Wouters, T. Dillon, W. Rahayu, E. Chang, A. Hameurlain, R. Cicchetti, and R. Traunmüller. A practical walkthrough of the ontology derivation rules. In *Proceedings of Database and Expert Systems Applications (DEXA2002)*, 2002. 99

[391] C. Wroe, R. Stevens, C. Goble, and M. Ashburner. A methodology to migrate the Gene Ontology to a description logic environment using DAML+OIL. In *Proceedings of the Pacific Symposium on Biocomputing*, 2003. 112

[392] F. Yadetie, I. Bakke, A. Lægreid, W. Kuśnierczyk, J. Komorowski, H. Waldum, and A. K. Sandvik. Analysis of effects of the PPAR α agonist ciprofibrate on rat hepatic gene expressions using cDNA microarrays. In *Proceedings of the 7th IUBMB Conference on Receptor-Ligand Interactions: Molecular, Physiological and Pharmacological Aspects*, 2002. 5

[393] F. Yadetie, A. Lægreid, I. Bakke, W. Kuśnierczyk, J. Komorowski, H. Waldum, and A. K. Sandvik. Liver gene expression in rats in response to the peroxisome proliferator-activated receptor-α agonist ciprofibrate. *Physiological Genomics*, 15:9–19, 2003. 5, 37

[394] T. Yagisawa. Possible objects. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2005. 88

[395] Y. Zhou, J. A. Young, A. Santrosyan, K. Chen, S. F. Yan, and E. A. Winzeler. *In silico* gene function prediction using ontology-based pattern identification. *Bioinformatics*, 21(7):1237–1245, Apr 2005. 135