Linköping Electronic Articles in
Computer and Information Science
Vol. 0(2001): nr 0

Ontologies for Molecular Biology

Steffen Schulze-Kremer

RZPD Deutsches Ressourcenzentrum fuer Genomforschung GmbH
steffen@rzpd.de
Berlin, Germany

Linköping University Electronic Press
Linköping, Sweden

http://www.ep.liu.se/ea/cis/2001/000/

Published on August 22, 2001 by
Linköping University Electronic Press
581 83 Linköping, Sweden

Linköping Electronic Articles in
Computer and Information Science
ISSN 1401-9841
Series editor: Erik Sandewall

Typeset by the author using LaTeX
Formatted using étendu *style*

Recommended citation:
<Author>. <Title>. Linköping Electronic Articles in Computer and Information Science, Vol. 0(2001): nr 0. http://www.ep.liu.se/ea/cis/2001/000/. *August 22, 2001.*

This URL will also contain a link to the author's home page.

Abstract

Molecular biology has a communication problem. There are many databases using their own labels and categories for storing data objects and some using identical labels and categories but with a different meaning. A prominent example is the concept "gene" which is used with different semantics by major international genomic databases. Ontologies are one means to provide a semantic repository that can be used to systematically order relevant concepts in molecular biology and to bridge the different notions in various databases by explicitly specifying the meaning of and relation between the fundamental concepts in an application domain. Here, the upper level and a database branch of a prospective ontology for molecular biology is presented and compared to other ontologies with respect to suitability for molecular biology (http://igd.rz-berlin.mpg.de/~www/oe/mbo.html).

1 Introduction

There are a multitude of databases accessible over the Internet that cover genomic [5], cellular [7], structure [4], phenotype [12] and other types of biologically relevant information [2]. Even for one type of information, e.g. DNA sequence data, there exist several databases of different scope and organisation [5, 8, 3].

Unfortunately, naming conventions of data objects, object identifier codes and record labels differ between databases and do not follow a unified scheme. But worse, even the meaning of important high level concepts that are fundamental to many molecular biology databases is ambiguous.

One prominent example is the concept *gene*. For GDB [5], a *gene* is a DNA fragment that can be transcribed and translated into a protein; for Genbank [3] and GSDB [8], however, a *gene* is a "DNA region of biological interest with a name and that carries a genetic trait or phenotype" which includes non-structural coding DNA regions like intron, promoter and enhancer. There is a clear semantic difference between those two notions of *gene* but both continue to be used interchangeably causing misunderstanding and making the integration of databases non-trivial.

To eliminate semantic confusion in molecular biology, it will be therefore necessary to have a list of most important and frequently used concepts coherently defined so that database managers could use such set of definitions either to create new database schemata or to provide an exact, semantic specification of the concepts used in an existing schema.

To become generally acceptable in the molecular biological community such semantic compendium (also often called controlled vocabulary) must be accessible electronically and without licensing charges, preferably using a world wide web browser; be intuitively comprehensible without special computer programming background; be able to cope with natural language features as e.g. homonyms; be capable to perform logical inference over the set of concepts to provide for generalisation and explanation facilities; exhaustively cover the application domain; and be coordinated but open for input from the community. Also, software to manage a semantic repository must be made available.

One way to consistently and transparently create such set of definitions for molecular biology is by using an ontology. By adhering to a commonly agreeable ontology, uncertainties and misunderstandings about the semantic relations between database entries from different databases can be eliminated. When all relevant concepts of an application domain will have been specified in an ontology, a computer program can search for *concepts* instead of words in a set of heterogeneous, autonomous databases [14]; carry out semantic consistency checks; and detect ill-formed statements and interpret well-formed ones [15].

In this report, two well-known ontologies, Cyc and μKosmos are examined with respect to their applicability in molecular biology. Since both are found not to fully satisfy that purpose the foundations of a new, prospective ontology for molecular biology are laid out.

2 What is an Ontology (and what isn't)?

Ontology was originally perceived by ancient philosophers as the study of *being*. They asked "What does the statement 'X *is*' mean?" and "Which things *are*?"[1]. In modern times, computer science uses ontology in a narrower sense as a "specification of a conceptualization" [6] or, in other

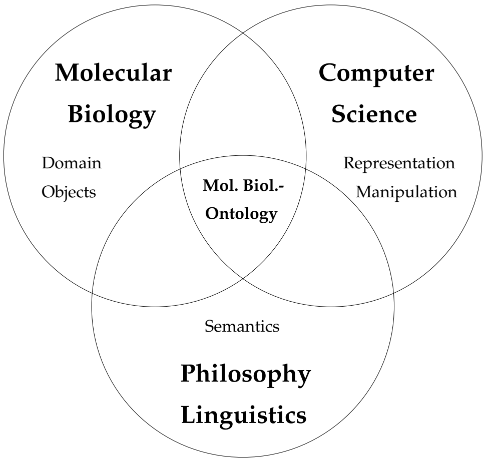

Figure 1: Molecular biologists discover facts that need to be organised and stored in databases. Computer scientists provide techniques for data representation and manipulation. Philosophers and linguists help organise the *meaning* behind database labels.

words, as a concise and unambiguous description of what principal entities are relevant in an application domain and how they can relate to each other. The entities can be objects, processes, functions, predicates, or of other type depending on the selected representation formalism (Figure 1). The formal definition of the components of an ontology and heuristics for ontology construction are given elsewhere [16].

An ontology is *not* a collection of facts that arise from an actual, specific situation but it defines and provides all semantic entities and their potential interactions that would be necessary to completely describe that situation. *Neither* is an ontology a model for an application domain (which would be a theory), but a compendium that holds all necessary "building blocks" with rules of how and which entities can relate to each other and which ones are semantically incompatible. For example, "transcription of a gene" is an ontologically valid expression whereas a "transcription of a cell" is not. *Nor* is an ontology a database schema which defines categories and their data types in a database but which need not represent ontological relations between entities in the real world.

The graphical representation of an ontology in general is not a tree but a semantic net or conceptual graph [18] because there are two or more types of links ("is a member of" and "is a subset of" plus additional domain specific relations) which can give rise to circular loops in the graph. However, if only the "is a subset of" or "is a member of" relation is displayed the ontological graph becomes a tree.

3 Ontologies for Molecular Biology

Two ontologies are discussed with respect to their applicability in molecular biology. Several other approaches exist but which are either sparsely populated, specialised to other domains or in general do not easily connect to molecular biology. After that, parts of a new, prospective ontology for molecular biology are presented. An interactive, graphical rep-

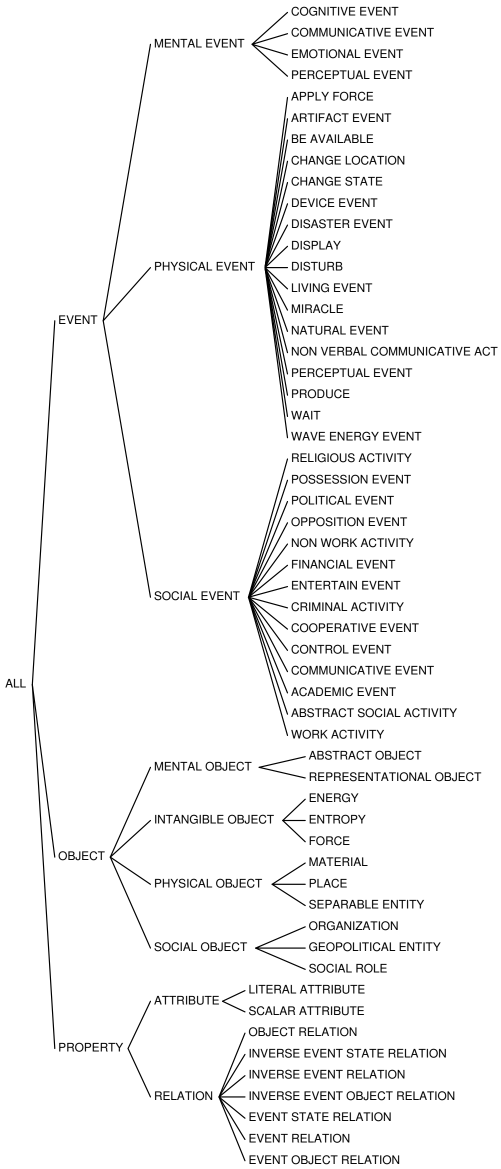

Figure 2: Upper Level of μKosmos Ontology. Links represent the "is a subclass of" relation. Hyphenation was removed from original concept names for better reading.

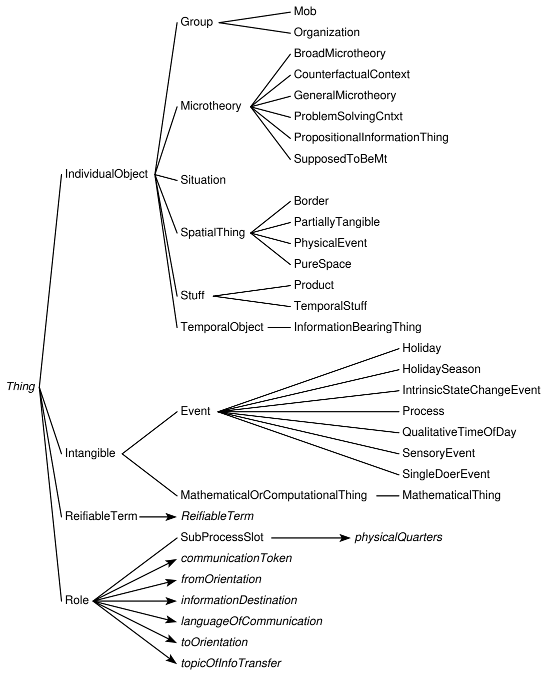

Figure 3: Upper Level of Cyc Ontology. Straight lines indicate "is a subclass of" relation, arrows and italics denote "is a member of" relation (instances). The leading special characters "#$" in concept names have been removed for clarity.

resentation of all public classes and instances of μKosmos, Cyc and the prospective ontology for molecular biology was prepared with a new ontology editor [17] and is accessible for browsing on the world wide web at http://igd.rz-berlin.mpg.de/~www/oe/mbo.html.

3.1 μKosmos

μKosmos [11] is an ontology developed for machine translation of business and financial texts between English, Spanish and other languages. The ontology contains 4790 annotated concepts which are publicly accessible as HTML hypertext and Lisp code. Ontological entries in μKosmos are linked to dictionaries in several languages. Thus, the ontology can serve as an inter-lingual framework to map the meaning of words from different languages. The links to the language dictionaries are not publicly available.

The upper level of the μKosmos ontology is graphically depicted in Figure 2.

There are several features of μKosmos that limit its suitability for use in molecular biology. μKosmos does not include a lot of concepts from molecular biology because it was developed mainly for translation of financial and corporate texts. μKosmos cannot store homonyms, i.e. one word with different (and possibly disjunct) meanings at different places in the ontological tree. There is no ontology editor available with μKosmos and no interactive graphical browsing facilities as at http://igd.rz-berlin.mpg.de/~www/oe/mbo.html.

3.2 Cyc

Cyc [9] is an ontology originally developed to cover everyday common-sense knowledge. Of the reported "tens of thousands" ontological entries a subset of about 2200 are publicly available as HTML hypertext with ample documentation. Cyc was not built to support a specific application but with the intention to cover even subtle semantic distinctions that a person has to consider when communicating in daily life. There is no ontology editor available with Cyc and no interactive graphical browsing facilities are provided except static HTML hypertext. Cyc has an inference engine and a natural language interface with the ontological entries linked to an English dictionary but all of which are not available to the public. The complete version of Cyc is commercially available.

Although Cyc contains a large and detailed collection of well documented concepts it is of limited use for molecular biology for several reasons.

Cyc does not include a significant portion of concepts relevant to molecular biology since it was designed to be a universal ontology. Only very basic knowledge about chemistry and biology has been added.

Although the authors of Cyc state that they "generally only list a non-redundant series of supersets" or "the incommensurably most specific (i.e., smallest) supersets of each collection" this rule is violated on several occasions. For example, #$Collection has listed the supersets #$Intangible, #$Thing and #$Set of which #$Thing is a superset of #$Intangible which in turn is a superset of #$Set. There are also several cases where two concepts are listed to be the superset of each other, e.g. #$Stuff and #$IndividualObject.

#$Thing, the "universal set of everything", has as its immediate subclasses #$IndividualObject, #$Intangible, and #$Role of which all three are overlapping because there exist intangible #$IndividualObject(s) and a #$Role is something both individual and intangible (Figure 3). The definition of #$Thing as the set of everything also faces Russel's set dilemma.

Though most definitions in Cyc seem philosophically well established, what is visible to the public is counterintuitive in some places. For example, #$Situation is defined to be "a state of affairs" with superclass #$IndividualObject which is a "discrete, not abstract entity that can have parts but not elements or subsets", suggesting that not only objects involved in a #$Situation but also #$Situation itself is a tangible entity since no link to #$Intangible exists.

The concept #$Stuff, defined as a discrete object that "when divided into pieces remains of the same type" (e.g. water) includes "physical entities like wood", "temporal entities like the event of a person running" and abstract things like "a piece of English text". One problem with the definition of #$Stuff is its granularity: on a molecular scale wood can well be divided into components that no longer are wood. Similarly, English text can be divided into letters which are neither distinctively English nor text anymore.

The criterion used to subclassify a concept in Cyc is not always stated explicitly. In many cases, subclasses in one class overlap semantically or are created using different subclassifying criteria. No homonyms are found in the public parts of Cyc. Naming of concepts is sometimes confusing, e.g. #$Thing vs. #$SomethingExisting; #$PartiallyTangible vs. #$PartiallyIntangible; #$IntangibleObject vs. #$IntangibleStuff.

Cyc contains a hierarchy of classes containing only classes that in some cases mirrors a similar hierarchy of classes containing instances but which

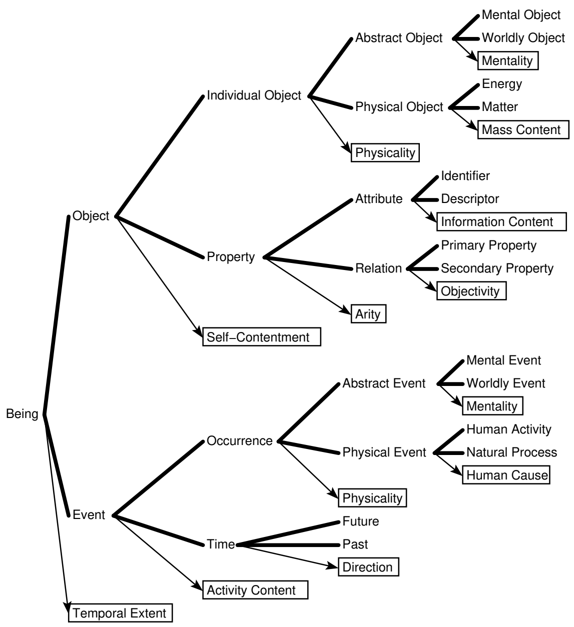

Figure 4: Upper Level of a prospective Molecular Biology Ontology. Links represent the "is a subclass of" relation. No instances are present; discriminating criteria have arrows and boxes; thick lines denote disjunct subclasses.

does not convey any new information. This adds to the confusion when searching for a concept. All these properties of the Cyc ontology make it difficult to locate the appropriate position for an existing concept or for a new one to be added.

3.3 Molecular Biology Ontology

An ontology for molecular biology should become a repository for all relevant concepts that are required to describe biological objects, experimental procedures and computational aspects of molecular biology. Although this looks like an impossible task at first sight it does not mean to compile all knowledge about molecular biology nor does it imply being able to explain every biological phenomena. It just means collecting all types of entities that a molecular biologist includes in her professional thinking and placing those concepts appropriately in a "is a subset of" and "is a member of" hierarchy plus annotating them with additional properties.

By doing so in a consistent manner, where the discriminating criteria of subclassifying each concept are made explicit, the definition of a concept becomes the path from its own node to the root node of the ontology. As an example, it can be read as "*Name* is an $\Rightarrow$ *Identifier* is an $\Rightarrow$ *Attribute* is a $\Rightarrow$ *Property* is a $\Rightarrow$ ...", for the case of only "is a subclass of" relationships ($\Rightarrow$). Relations of type "is a member of" ($\rightarrow$) inherit only from their direct parent node(s) because being an instance of a concept is different from being a subclass of a class. For example, "Aristotle is a $\rightarrow$ *Name* is an $\Rightarrow$ *Identifier* is an $\Rightarrow$ *Attribute* is a ..." means that Aristotle is one *Name* but not an *Attribute*, whereas *Name* is an *Attribute*.

One difficulty when compiling an ontology for molecular biology is identifying the subtle connotations that are hidden in everyday language. For example, the chromosome of an E. coli bacterium is a DNA molecule which is a physical object. The sequence of that DNA, however, is an abstract object that is not contained in E. coli at all but that can be subjected to mathematical analysis. Therefore the concepts DNA and DNA sequence will reside on quite remote branches in an ontological graph.

The upper level of a prospective Ontology for Molecular Biology is shown in Figure 4. Starting from the root node *Being* which includes anything that *is*, the classes *Object* and *Event* are disjoint and discriminated based on their temporal extent. An *Object* remains an *Object* even in a single moment whereas an *Event* when dissected into single moments looses its identity. This holds also for all subclasses of *Object* and *Event*. The class *Object* is further subclassified into *Individual Object* and *Property*. Both can be thought of as instantaneous, i.e. they keep their identity even if looked at only for one moment. The two are discriminated based on self-contentment. An *Individual Object* can stand alone whereas a *Property* always needs another *Object* or *Event* to refer to. A *Property* is further subclassified based on arity into *Attribute*, a property with only one argument, and *Relation*, a property relating two or more *Being*s. Hereby, the logical grammar of words, not their surface structure must be considered. For example, in the statement "Paris is beautiful", beautiful is not a logical attribute to Paris because this statement necessarily involves a second entity, the speaker, and thus becomes one binary and one unary relation: "She thinks, Paris is beautiful" or in Prolog syntax: think (she, beautiful (Paris)).

Attribute can be subclassified into *Identifier* and *Descriptor* based on whether it just labels an entity or whether it carries additional information about it. *Relation* can be subclassified analogous to Locke [10] into *Secondary Property* relations that involve personal judgement and *Primary Property* factuals describing intersubjective measurable relations.

Individual Object is subclassified based on physicality into *Abstract Object*, which has no physical equivalent *per se* (except capable of being represented neurologically or in writing, etc.) and *Physical Object*, which must have a defined spatial extension and/or energy content and is similar to Popper's "World 1" [13]. *Abstract Object* is further subclassified based on mentality, i.e. whether it refers to an object within the mind or to an object in the outside world, into *Mental Object* (similar to Popper's "World 2") and *Worldly Object* (similar to Popper's "World 3"). Although energy and matter are equivalent in nuclear physics a given object can be only of one type at a time. Hence, *Physical Object* has been subclassified based on mass content into *Energy* and *Matter*.

On the other branch of the ontology *Event* is subclassified based on activity into *Occurrence*, where at least one object participates and (pure) *Time*, where nothing happens. This is the notion of absolute time which is no longer valid in relativistic physics and astronomy. The reason for nevertheless holding on to the belief of absolute time here is justified by the intended scope of the ontology for molecular biology: physical processes in living organisms have so far never been known to reach the realm of relativistic physics. *Time* is further subclassified according direction into *Past* and *Future*. Because presence strictly lasts one moment only, it does not appear in this branch.

Analogous to abstract and physical objects, *Occurrence* is subclassified based on physicality into *Abstract Event* and *Physical Event*, and further *Abstract Event* based on mentality into *Mental Event* (similar to Popper's

Concept	Discriminator	Value
Being	Temporal Extent	instanteneous
Object	Self-Contentment	yes
Individual Object	Physicality	no
Abstract Object	Mentality	no
Worldly Object	Domain Specific Usage	yes
Domain Specific Wordly Object	Subject Domain	mathematics
Mathematical Object	Complexity	high
Composed Mathematical Object	Application Specificity	yes
Applied Mathematical Object	Application Domain	computer science
Computer Science Object	Subject Matter	theory of data
Theory Of Data Object	Subject Matter	data structure
Data Structure Object	Representation Formalism	object-oriented
Object-Oriented Data Structure	Implementation	OPM
OPM Database Object	Database	GDB
GDB Database Object	Subject Matter	genomic
GDB Genome Object	Object Class	DBObject
DBObject	Subclass	MappingObject
MappingObject	Subclass	Map
Map	Subclass	*LinkageMap*

Figure 5: Semantic Hierarchy for the GDB database category *LinkageMap*. A *LinkageMap* is a *Being* with instantaneous temporal extent, an *Object* which is self-contenting, etc. until the concept *LinkageMap* is reached right at the bottom.

"World 2") and *Worldly Event* (similar to Popper's "World 3"). *Physical Event* is similar to Popper's "World 1" and subclassified based on whether it is done or initiated by human intention into *Human Activity* and *Natural Process.*

Similar reasoning was applied to ontologically capture the meaning of the GDB database category *LinkageMap*, which is defined as a "database object class used to store maps based upon frequency of recombination between genomic segments, resulting in the ordering of markers along a chromosome backbone, usually measured in centiMorgans". Note that this is ontologically not the same as a linkage map itself which is an abstract concept with certain mathematical properties, nor is it an actual linkage map which is a concrete instance of the class of linkage maps for a particular organism and chromosome.

The complete path from root node *Being* to *LinkageMap* is summarised in Figure 5. The meaning of the database category *LinkageMap* is captured by a series of ontological specifications. This example shows how a semantic definition of a molecular biological concept can be extracted from its mere position in an ontological graph. Similarly, semantic differences and the least general common concept of a pair of concepts can be found by following the graph upwards along "is a subclass of" and "is a member of" links until both paths meet in one concept.

It follows a more biologically oriented example, an ontological fragment on cDNA microarrays as depicted in Figure 6. We start out with the most general class of cDNA microarrays *cDNA Array.* This is subclassified by the phase of the target, whether it is on the chip or in solution. Chips with targets fixed are further subclassified based on the type of *Measurement* and then on the main type of *Application.* Chips that catch their targets from solution are also subclassified by *Measurement* into *One Channel* (e.g. radioactive labelling) and *Two Channel* (e.g. fluorescent labelling) devices. Further subclassification is done by *Support Material,* which could be nylon membranes or glass slides. The final subclassification in this example is by *Probe Type,* here defined as whether only one species of nucleic acid is brought over the chip or a mixture of nucleic acids. Thick red lines indicate

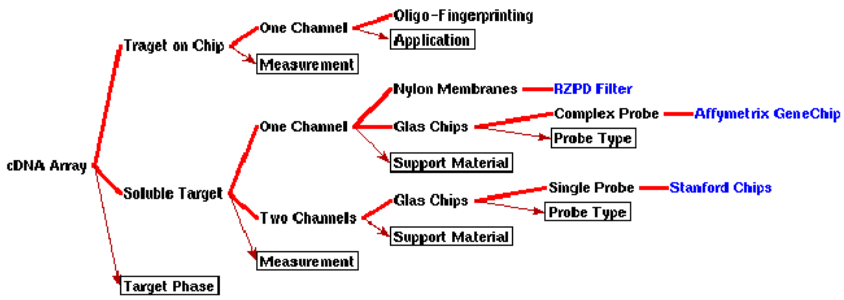

Figure 6: Various types of DNA microarrays are subclassified based on their properties.

disjunct subclasses, concepts with boxes and arrows are discriminating criteria for the respective superclass and blue concepts denote instances of a particular class.

4 Discussion

In this work, the communication problem in molecular biology has been looked at from the viewpoint of semantic integration. To improve the current situation of non-unified and ambiguous vocabulary the only solution is to develop a core of commonly agreeable definitions and using those to implement user interfaces to and between databases. Those definitions must be connected to each other so that no ambiguities over relations between concepts remain and that a computer may infer from specialised to generalised concepts and vice versa.

An ontology as an explicit and hierarchical specification of the relevant concepts in an application domain is one means to develop such semantic repository. Here, two ontologies from the literature, μKosmos and Cyc, have been reviewed with respect to suitability in molecular biology. The μKosmos ontology is found not to be transparent and precise enough to collect and sort scientific concepts concerning molecular biology. Cyc contains a lot of knowledge about semantic distinctions in daily life but seems to be too complicated and overloaded with concepts not relevant for molecular biology (and probably too expensive) to be of use here.

Work has begun on a prospective ontology for molecular biology. The upper level of the ontology is lean and in agreement with traditional arguments from philosophy of science. In contrast to other ontologies, the criterion used for subclassifying a concept is explicitly stated and therefore essential decisions and assumptions behind the ontology are made transparent. Homonyms can be handled by explicit reference to their superclasses. Currently, the ontology has about 1300 concepts with emphasis on molecular biology database entities and should be regarded as work in progress.

References

References

[1] Aristotle. Categories. Translated by E. M. Edghill, 350 BC. http://entisoft.earthlink.net/category.htm.

[2] A. Bairoch. The ENZYME data bank. *Nucleic Acids Research*, 21(13):3155–3156, July 1993. http://www.gdb.org/Dan/proteins/ec-enzyme.html.

[3] Dennis A. Benson, Mark S. Boguski, David J. Lipman, and James Ostell. GenBank. *Nucleic Acids Research*, 25(1):1–6, 1997. http://ncbi.nlm.nih.gov/genbank/query_form.html.

[4] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The Protein Data Bank: a Computer-based Archival File for Macromolecular Structures. *Journal of Molecular Biology*, 112:535–542, 1977.

[5] K. H. Fasman, S. I. Letovsky, R. W. Cottingham, and D. T. Kingsbury. Improvements to the GDB(TM) Human Genome Data Base. *Nucleic Acids Research*, 24(1):57–63, 1996. http://gdbwww.gdb.org.

[6] Tom R. Gruber. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5(2):199–220, 1993. http://www-ksl-svc.stanford.edu:5915.

[7] D. Jacobson and A. Anagnostopoulos. Internet resources for transgenic or targeted mutation research. *Trends in Genetics*, 12(3):117–118, March 1996. http://www.gdb.org/Dan/tbase/tbase.html.

[8] Gifford Keen, Chris Fields, et al. The Genome Sequence DataBase (GSDB): meeting the challenge of genomic sequencing. *Nucleic Acids Research*, 24(1):13–16, 1996. http://wehih.wehi.edu.au/gsdb/gsdb.html.

[9] D. B. Lenat. CYC: A Large-Scale Investment in Knowledge Infrastructure. *Communications of the ACM*, 38(11), November 1995. http://www.cyc.com/public.html.

[10] J. Locke. *An Essay Concerning Human Understanding*, volume 2, 4, chapter 8, 3, pages 23–26, 11–16. Oxford University Press, 1975, 3 edition, 1690. http://www.ilt.columbia.edu/academic/digitexts/locke/understanding/title.html.

[11] K. Mahesh and S. Nirenburg. Meaning Representation for Knowledge Sharing in Practical Machine Translation. In *Proceedings of the FLAIRS-96 Track on Information Interchange, Florida AI Research Symposium*, May 1996. http://crl.nmsu.edu/users/mahesh/onto-intro-page.html.

[12] V. A. McKusick. *Mendelian Inheritance in Man. Catalogs of Human Genes and Genetic Disorders.* Johns Hopkins University Press, Baltimore, MD, 11 edition, 1994. http://www3.ncbi.nlm.nih.gov/omim.

[13] K. R. Popper and J. C. Eccles. *The Self and Its Brain.* Springer, 3 edition, 1985. http://www.springer.de.

[14] B. R. Schatz. Information Retrieval in Digital Libraries: Bringing Search to the Net. *Science*, 275(5298):327–334, January 1997. http://csl.ncsa.uiuc.edu.

[15] S. Schulze-Kremer. *Molecular Bioinformatics - Algorithms and Applications*, pages 13–108. Walter de Gruyter, Berlin, 1996. http://mycroft.rz-berlin.mpg.de/~steffen/expose.html.

[16] S. Schulze-Kremer. Adding Semantics to Genome Databases: Towards an Ontology for Molecular Biology. In *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*, pages 272–275. AAAI Press, CA, 1997. http://igd.rz-berlin.mpg.de/~www/oe/mbo.html.

[17] S. Schulze-Kremer. Building an ontology for molecular biology with a graphical ontology editor. Submitted to *Data & Knowledge Engineering*, 1997.

[18] J. F. Sowa. *Conceptual Structures: Information Processing in Mind and Machine.* Addison-Wesley, Reading, MA, 1984. http://www.cs.adelaide.edu.au/users/peirce.

Contents

List of Figures