Basic Formal Ontology for Bioinformatics

Barry Smith[1,2], Anand Kumar[1], Thomas Bittner[1]

[1]Institute for Formal Ontology and Medical Information Science, Saarland University, Saarbrücken, Germany
[2]Department of Philosophy, Buffalo, New York.

1. Introduction

Two senses of 'ontology' can be distinguished in the current literature. First is the sense favored by information scientists, who view ontologies as *software implementations* designed to capture in some formal way the consensus conceptualization shared by those working on information systems or databases in a given domain. [Gruber 1993] Second is the sense favored by philosophers, who regard ontologies as *theories* of different types of entities (objects, processes, relations, functions) [Smith 2003]. Where information systems ontologists seek to maximize reasoning efficiency even at the price of simplifications on the side of representation, philosophical ontologists argue that representational adequacy can bring benefits for the stability and resistance to error of an ontological framework and also for its extendibility in the future.

In bioinformatics, however, a third sense of 'ontology' has established itself, above all as a result of the successes of the Gene Ontology (hereafter: GO), which is a tool for the representation and processing of information about gene products and functions [Gene Ontology Consortium 2000]. GO is, as the GO Consortium puts it, a 'controlled vocabulary', and its authors have focused neither on software implementations nor on the logical expression of theories. Their efforts have been directed, rather, toward providing a practically useful framework for keeping track of the biological annotations that are applied to gene products in a variety of contexts [Gene Ontology Consortium 2001]. This means that when faced with the trade-off between (1) formal and ontological coherence, and (2) the speedy population of the vocabulary with biological concepts, preference was given by the GO consortium overwhelmingly to the latter. In what follows we provide a survey of GO, and of some general lessons about the role of ontology in bioinformatics which we can learn from both its successes and its failures.

3. The Structure of GO

GO provides a controlled vocabulary for the description of cellular components, molecular functions, and biological processes. It is being developed in tandem with work on a variety of biological databases within the framework of the umbrella project OBO, of "Open Biological Ontologies"[1] constructed on the basis of the GO methodology. GO is currently being used within a variety of biological databases including Uniprot, the largest available protein database, TIGR, InterPro, the Enzyme Commission's Enzyme

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

database,[2] and in a variety of other contexts. It is also being used as the main vocabulary for many of the microarray related tools developed for the analysis of data.[3]

The May 10, 2004 edition of GO contains 1409 component, 7430 function and 8465 process terms, making 17304 terms in all, of which some 11020 are provided with informal definitions. The terms are organized in hierarchies structured by means of two kinds of links, the one (*is_a*) indicating the subclass or subtype relation (or in other words that one entity *is more general than* another), the other (*part_of*) indicating that the entity denoted by one term *includes as part* the entity denoted by another. This same two-link structure has been employed also in other important bioinformatics ontologies – for example in the Foundational Model of Anatomy (FMA)[4] [Rosse and Mejino 2003] – and it is of course present also outside bioinformatics, for example in the lexical database WordNet [Fellbaum 1998]. In addition to *is_a* and *part_of* GO, like other similar systems, also uses a variety of other ontological or quasi-ontological expressions in composing its terms and definitions – for example *function*, *component*, *constituent*, *substance*, *action*, *activity*, *process*, *domain*, *complex*, *unit*. Unfortunately, however, these terms, like GO's ontological relations are nowhere rigorously defined in the GO documentation – and this in spite of the claim made in [Gene Ontology Consortium 2001] to the effect that GO 'comprises a set of well-defined terms with well-defined relationships'.

GO's success in serving the biological community has led some researchers to attempt to expand its utility by using GO in tandem with software applications designed to replace manual comparison of the properties of gene products with automatic reasoning. The Gene Ontology Next Generation (GONG) project[5] is attempting to improve GO's suitability for use by computers by rendering GO in a Description Logic format [Wroe *et al.* 2003]. Another effort applies the Protégé 2000[6] frame-based ontology editor and associated tools to browse and edit GO and to verify certain kinds of consistency [Yeh *et al.* 2003].

All of these efforts, however, accept GO as it is, and seek to supplement it with formal reasoning tools. They thus ignore the degree to which the existing architecture of GO and similar systems harbor problems which stand in the way of such formalizing efforts effectively by making much of GO's content inaccessible to automatic reasoning tools. In a series of papers [Smith *et al.* 2003, Smith *et al.* 2004, Smith and Rosse 2003] we have attempted to demonstrate, through the analysis of a wide range of examples, that by taking account of certain organizing principles drawn from philosophical ontology, GO's consistency and coherence, and thus its future applicability in the automated processing of biological data, can be enhanced. Here we take such analyses further in an attempt to demonstrate how the category system of the Basic Formal Ontology (BFO) [Smith 2003a] can provide a representational framework for ontologies like GO which can facilitate their integration into larger systems that are able to support reasoning across biomedical information drawn from a variety of different sources.

[2] http://molbio.info.nih.gov/molbio/db.html
[3] http://www.geneontology.org/GO.tools.html
[4] http://sig.biostr.washington.edu/projects/fm/index.html
[5] http://gong.man.ac.uk/background.html
[6] http://protege.stanford.edu

4. Basic Formal Ontology (BFO)

At its core, BFO gives a formal account of the distinctions between:

(a) universal and particular
(b) continuant and occurrent
(c) dependent and independent,
(d) formal and material.

4.1 Universals and Particulars

BFO distinguishes first between universals (also called kinds, classes, species, types) and particulars (also called individuals, exemplars, instances, tokens) [Smith 2003b, Bittner *et al* 2004]. Examples of universals are: the species *E. coli*, the function: *to boost insulin production*. Examples of particulars are: *the E. coli bacterium now existing in this Petri dish*, *the function of this gene to boost insulin production in the beta cells in your pancreas*. The terms in ontologies like GO correspond, in philosophical terminology, to universals, that is to entities which are multiply instantiable. Thus the universal corresponding to the term *cell* is *instantiated* by every actual cell. This instantiation relation will be represented below by means of the relational expression '**inst**'.

The distinction between universals and particulars allows us to provide a more coherent account of the *is_a* and *part_of* relations than that which is presented in the current GO and OBO documentation. In particular it supports a distinction between:

(i) the *is-a* relation as a relation between universals, for example, every human *is-an* animal (so that whenever **inst**(*x*, *human*) we also have **inst**(*x*, *animal*)).
(ii) the part relation as a relation between particulars (referred to in what follows by means of the relational expression **part**), for example: your arm is part of your body;
(iii) various part relations asserted to hold between universals (where as we shall see, they in fact hold more properly speaking only *via* the particulars by which these universals are instantiated), for example, as in GO: *nucleus* part_of *cell*.

4.2 Continuants vs. Occurrents

Orthogonal to the distinction between universals and particulars is that between continuants and occurrents. Continuants, as the name implies, are entities which endure, or continue to exist through time. Organisms, cells, chromosomes, molecules are all continuants: they preserve their identity from one moment to the next, even while undergoing a variety of different sorts of changes. The parts and boundaries of continuants – for example your arms and legs, the outer surface of your skin – are also continuants.

Where the principal mark of a continuant is that it is exists in full at any time at which it exists at all. Occurrents (also called events, processes, activities) are marked by the fact that they never exist in full in any single instant of time; rather, they are such as to unfold themselves in their successive phases – in the way in which, for example, the process of embryological development unfolds itself through the successive phases distinguished by developmental biologists. Note that **part** relations never cross the continuant/occurrent divide. Where your arm is **part** of you, your youth is **part** of that process which is your

life. [Bittner and Smith 2003, Bittner and Donnelly 2004, Bittner at al 2004, Bittner 2004].

4.3 Dependent vs. Independent Entities

A third distinction is that between dependent and independent entities. This reflects the fact that while some entities (planets, people, molecules, atoms) have an inherent ability to exist without support from other entities, others require such support in order to exist: a viral infection is dependent upon certain instances of a given virus and also upon the organism which is infected; the function of an organ is dependent upon the existence of the organ which it is the function of [Grenon and Smith 2004].

Both the continuant/occurrent and the dependent/independent distinction apply at the level of both universals and particulars. Thus the *functioning* of my heart here and now (a particular occurrent) is dependent on my heart itself and on its function (both particular continuants) in a way which reflects exactly parallel dependence relations among the corresponding universals.

4.4 Formal vs. Material

Biology deals primarily with entities referred to by material terms such as *cell*, *nucleus*, *organism*, *death*. Ontologies deal also with the various formal relations by which these material entities are connected together. [Smith and Grenon 2004] Material terms are characterized by the fact that they apply to entities in one domain of reality only; formal relations by the fact that they can hold between entities which span domains. Examples of formal relations are: identity, dependence, instantiation, parthood. Such relations will receive a special treatment in the formal framework to be outlined below.

5. Using Formal Ontology to Improve the Definitional Resources of Bioinformatics

GO and related terminology systems in biomedical informatics such as the FMA or the Semantic Network of the Unified Medical Language System[7] (UMLS), provide not only terms arranged in hierarchies but also *definitions*. A definition is a group of words or symbols designed to explain the meaning of some other word or symbol. Good definitions are those which advance communication and understanding; bad definitions are those which hinder or at least fail to advance communication or understanding. All definitions are made up of two parts, the definiendum and the definiens. The definiendum is whatever word, symbol, or group of words is being defined; the definiens is whatever words are being used to do the defining. Thus, in the statement 'a definition is a group of words or symbols designed to explain the meaning of some other word or symbol,' the definiendum is 'a definition' and the definiens is everything to the right of 'is'. In the regimented format we shall employ in what follows we use the special symbol '$=_{def}$' to substitute for 'is'. In a good definition, all the terms which constitute the definition should be defined, unless they are common English terms, in which case (if they are not stop words, such as 'the' or 'is') they should be linked to some external lexical resource such as WordNet. In a good definition, further the definiens should be of the same part of speech as the definiendum, so that the latter can be substituted for the former in different

[7] http://www.nlm.nih.gov/research/umls/

sentential contexts in such a way as to preserve not only grammatically but also truth value.

5.1 Problems with OBO Definitions

We shall now examine some examples of definitions, deriving not only from GO but also from the collection of Open Biological Ontologies (OBO) to which it belongs, attempting to formalize these examples using BFO predicate logic. Our goal is to illustrate and resolve some of the problems by which definitions in standard bioinformatics resources are affected, both in order to illustrate the present state of the discipline and also in order to suggest future paths towards improvement.

Our first example we take from the file "Rel.Definitions" of the OBO Relationship Ontology (also headed "Gene Ontology Definitions").[8] This file is to a degree misnamed, since the relations with which it deals would in some cases more adequately be treated as primitives, and thus as entities for which no definitions can or should be supplied. We will examine the proposed 'definitions' nonetheless, since they serve to illustrate some of the problems which relate to OBO definitions more generally, including those to be found in GO. In OBO's relationship ontology we find for example the following definition of *is_a*:

> Represents subsumption relationships. The subject (child) is the more specific term; the object (parent) is the more general term. Corresponds exactly to the daml and rdfs property "subClassOf", which means the following always holds:
> Entity instance_of TermX
> TermX is_a TermY
> implies:
> Entity instance_of TermY
> For example, if Entity is a specific gene, then if that gene is assigned to class X then it is implicitly a member of class Y (and parents of Y, because is_a is itself a transitive_relationship).

If we attempt to rephrase the initial part of this definition in our preferred regimented format, then this yields:

Is_a $=_{\text{def}}$ represents subsumption relationships. The error here is three-fold:

(i) the definiens is contains terms which are no easier to understand than the terms used in the definiendum; the definition thus provides little aid to the understanding;

(ii) definiens and definiendum are not intersubstitutable, so that the former cannot be said to have captures the meaning of the latter;

(iii) what is provided is not in fact a definition of '*is_a*' at all; rather, what is defined is '"*is_a*"', which is to say a certain syntactic expression, which 'represents subsumption relations'.

If we analyze the remainder of the definition then we can infer that GO's *is_a* is either identified with the standard inclusion (subset) relation of set theory (which does not correspond to the actual practice of the OBO ontologies in using this relation) or it provides only sufficient but not necessary conditions for *is_a* to hold, and thus is not a true definition. In addition it involves potentially problematic terms such as 'assigned to'

[8] http://cvs.sourceforge.net/viewcvs.py/*checkout*/obo/obo/ontology/OBO_REL/rel.definitions?rev=1.2

and 'implicitly' which are themselves nowhere defined, and other terms, such as 'more general than', which serve to make the proposed definition circular.

The 'definition' of *part_of* proposed by OBO reads:

> Used for representing partonomies. The subject (child node) of the relationship is the subpart; the object (parent node) is the superpart. Part_of can be used in various contexts – spatial, compositional, temporal. The context can usually be inferred from the terms it relates (for instance, in a process ontology, it means sub_process_of).

This definition shies away from formal rigor entirely, and amounts to little more than a series of remarks about (some aspects of) the *part_of* relation. The proposed definition, too, is circular (the term 'partonomy', which is itself defined in terms of 'part', is used within the definition). Moreover, the definition does not address the serious difficulties involved in giving a coherent account of *part_of* when once the difference between universals and particulars is taken into account [Smith and Rosse 2004]. GO's Documentation[9] has recently (March 2004) been revised in an attempt to take account of this latter problem.

5.2 Problems with GO Definitions

In order to dig more deeply into what is needed in the way of formally rigorous definitions by the biomedical informatics ontologies of the future, we will carry out a series of semi-formal restatements of GO's definitions using the language of the first-order predicate logic (which we take to be a language of sufficient expressive power to capture the main sorts of distinctions which are of ontological importance). In each case we begin by providing the definiendum on the left-hand side and the definiens provided by GO on the right.

5.2.1 Definitions Involving Chemical Formulae

GO asserts in its documentation that it will provide definitions in ordinary English. Occasionally, however, it imports chemical expressions, as for example in:

(+)-borneol dehydrogenase activity	*Catalysis of* the reaction: (+)-borneol + NAD+ = (+)-camphor + NADH + H+.

To unpack the formal content of this definition would require linkage to an external ontology of chemical reactions (and this in turn would require a degree of formal rigor which both GO and the available chemical ontologies are still far from realizing). Thus we can formalize here only part of what GO is attempting. We use '**inst**(*x*, (+)-*borneol dehydrogenase activity*)' to signify that *x* is an instance of the universal (+)-*borneol dehydrogenase activity* (where '(+)' stands for the isomeric state of the enzyme borneol dehydrogenase). Note that expressions like '(+)-*borneol dehydrogenase activity*' are not predicates, in our framework. Rather, they are names of universals. We restrict predicates to a small group of formal relations, which are expressed in bold face, as in '**inst**(*x*, *A*)' or '**part**(*x*, *y*)'. Occasionally such predicates are given special symbols, for example '**=**(*x*, *y*)', for '*x* and *y* are identical'.

[9] http://www.geneontology.org/GO.usage.html#partof

One problem with GO's definition of (+)-*borneol dehydrogenase activity* is that it uses the same expression '+' in three different ways – to indicate isomerization and ionization and also as the usual combination operator employed in representing chemical reactions. (Such ambiguous use of operators is in fact characteristic of GO. [Smith *et al*. 2004]

In order to give a first approximation to a formal rendering of GO's definition as follows, we assume that we can import from some chemistry ontology an expression '*R*' designating the class of reactions of the given type. We then have:

> **inst**(*x*, (+)-*borneol dehydrogenase activity*) $=_{def}$ **inst**(*x*, *catalysis*) and ∃*y*(**inst**(*y*, *R*) and **acts_on**(*x*, *y*))

(Here and elsewhere '∀' and '∃' are the usual universal and existential quantifiers of the predicate calculus and lower case roman italic letters are variables representing individuals.) The above definition thus asserts that *x* is an instance of (+)-*borneol dehydrogenase activity* if and only if *x* is an action of catalysis and there is some instance *y* of the given reaction-type *R* and *x* acts on *y*. **Acts_on** is a relation, in this case a kind of *regulation* between two processes.

Note that '*R*' has the character of a black box. Because GO is not linked to a chemistry ontology, it stops short from the point of view of supporting inferences.

5.2.2 Definitions Involving Loops

hemolysis	The processes that cause hemolysis, the lytic destruction of red blood cells with the release of intracellular hemoglobin, in another organism.

This second example of a GO definition is not merely circular; that is the definiens does not merely contain the definiendum as part. In addition it transforms the definiendum in a way which implies that it is in fact some different term that is being defined. The result amounts to an assertion to the effect that hemolysis is that which causes hemolysis, or in other words (using (*) to flag the presence of logico-ontological incoherence):

> (*) **inst**(*x*, *hemolysis*) $=_{def}$ ∃y (**inst**(*y*, *hemolysis*) and **causes**(*x*, *y*))

This is a particularly poignant expression of a general problem in GO, where definitions contain definitions of *other* terms as part. Another general problem turns on the fact that many GO definitions go beyond what is properly required from a well-constructed definiens by providing *additional information*, which (however helpful to human users of GO) should not for formal reasons, be part of the definition itself [Smith *et al.* 2003]. To rectify this problem here the added clause pertaining to release of intracellular hemoglobin must also be deleted. A better definition, clearing up these problems, would state simply:

Hemolysis	the lytic destruction of red blood cells

5.2.3 Definitions containing terms missing from GO

hermaphrodite genital morphogenesis	Formation and development of organized structures in hermaphrodite genitals.

Here one specific type of morphogenesis is defined as a process (an occurrent), which is dependent on a certain independent continuant: the genitals. Moreover it is asserted that the given process is such as to lead to the formation of certain organized structures within this continuant. Unfortunately 'formation' and 'development' are not themselves defined within GO; for this purpose one would need a formal machinery for talking about temporal sequence. The best we can do in the way of providing a formal rendering faithful to GO's definition scheme would be:

(*) **inst**(*x*, *hermaphrodite genital morphogenesis*) $=_{\text{def}}$ $\exists y$(**inst**(*x*, *organized structures in hermaphrodite genitals*) and **inst**(*y*, *formation-and-development*) and **part**(*y*, *x*))

A better definition could be achieved already by examining the compositional structure of the GO term itself, and we understand that GO's developers are currently initiating a reformulation of GO's terms and definitions along compositional lines: the definitions for compound terms being built up in stages out of the definitions of simple terms. This would mean obtaining definitions for terms like *morphogenesis*, *hermaphrodite* and *genital*, from external sources such as WordNet (or a version of WordNet that has been validated for purposes of biomedical research [Smith and Fellbaum 2004]). Defining *hermaphrodite genital morphogenesis* on the basis of such definitions (and thus following the definition principles adopted by the FMA) [Rosse *et al*. 1998] will bring clarity and systematicity to GO's definitional structure, since the definition itself would bear on its face its manner of having been logically derived.

5.2.4 Definitions with a high degree of unintelligibility

drinking behavior	The specific actions or reactions of an organism relating to the intake of liquids, especially water.

While all the terms used in this definition will be perfectly familiar to the average speaker of English, this definition is still unintelligible in the technical sense that, the definiendum is easier to understand than the definiens. This means that the definition is at best redundant. Definitions of terms in a controlled vocabulary should however *help* the user of the vocabulary, and if they do so then they will in addition have the right sort of compositional structure that they can be useful further to software tools for information referral.

A further problem with this definition is that, here again, constituent expressions such as 'specific', 'action', 'reaction' are not themselves defined in GO; moreover 'related to' is an expression too broad to convey any coherent meaning.

pronucleus	The nucleus of either the ovum or the spermatozoon following fertilization.

	Thus, in the fertilized ovum, there are two pronuclei, one originating from the ovum, the other from the spermatozoon that brought about fertilization; they approach each other, but do not fuse until just before the first cleavage, when each pronucleus loses its membrane to release its contents.

This definition tells us more than what a pronucleus, a continuant is. It also tells about the process of fertilization, an occurrent, in which the pronucleus is involved, and this extra information should strictly speaking be eliminated from this definition and provided rather in the context of a definition of 'fertilization'.

The protonucleus is an example of a continuant which transforms itself within a given period of time while preserving its identity. A given ovum or spermatozoon nucleus results from the transformation of a pronucleus as an adult results from the transformation of a child.

Because GO does not have temporal operators it is not able to do justice to relations of this sort. To achieve this end we need to add the machinery to reason about what holds at specific times. We thus introduce variables t, t_1, … to range instants of time, together with a temporal relation *earlier than*, symbolized **earl**, holding between them. We also need to temporalize the instantiation relation, yielding a framework in which the above definition might be formalized for example along the following lines:

> **inst**(x, *pronucleus*, t) $=_{\text{def}}$ $\exists t_1 \exists y$((**inst**(y, *ovum nucleus*, t_1) or **inst**(y, *spermatozoon nucleus*, t_1)) and **earl**(t_1, t) and **derives_from**(y, x)))

One problem with this definition is that it does not contain the information that a pronucleus is the nucleus of either ovum or sperm only before they fuse to form the zygote. To get these matters clear one would need to represent two processes: fertilization and zygote formation, which follow one another in close succession and assert that the pronucleus is that continuant which exists between these two processes. None of the GO definitions reflects this fact explicitly.

6. Granularity in Biomedical Ontologies

The holy gail of contemporary biomedical informatics is to find ways of bridging the granularity gap between molecular biological phenomenon at the one extreme and clinically relevant phenomena at the other. The granularities with which we have to deal in biomedicine include:

> whole organism: e.g. *human body, bacteria*
> organ system (higher organisms only): e.g. *digestive system*, *respiratory system*
> organ (higher organisms only): e.g. *pharynx*, *esophagus*
> tissue and tissue samples (multicellular and higher organisms): e.g. *adipose tissue*
> cell: e.g. *epithelial cell*, *ovum*
> subcellular: e.g. *cell membrane*, *nucleus*
> molecular: e.g. *ligand*

We can illustrate the role played by such granular levels in current bioinformatics by examining examples of definitions drawn from GO:

regulation of organ size	Any process that modulates the frequency, rate or extent of growth of an organ of an organism.
cell	The basic structural and functional unit of all organisms. Includes the plasma membrane and any external encapsulating structures such as the cell wall and cell envelope.
organellar ribosome	A ribosome contained within a subcellular organelle.

While there are references to entities at various levels of granularities present within these and other definitions, GO recognizes granularity only informally and haphazardly. Thus its biological process ontology comprises primarily entities at the whole organism and cellular levels; its cellular constituent ontology primarily entities at the cellular and subcellular levels; its molecular function ontology primarily functions/processes/activities at the molecular level. What GO and similar systems do not provide is any means for representing the fact that given entities belong to given levels of granularity. Moreover there are many types of entities – including organisms – which do not fall within the scope of GO at all. Thus when we consider the definition of "regulation of organ size" above contains two terms "organ" and "organism" referring to entities which fall outside the scope of GO. In addition, this definition contains the composite expression "organ of organism", which immediately raises the question whether there are, from GO's perspective, also organs which exist outside of organisms.

Part of the problem posed for GO and similar systems by the phenomenon of granularity is that granularity assignments are often species specific: thus what is of cellular granularity for instance of one species may be of whole organism granularity for instances of another. One of GO's fundamental principles however, is that it wants to deal with phenomena which appear in organisms in general and not in organisms of specific types. For human beings, entities are capable of being distinguished at all the mentioned levels of granularity. For unicellular bacteria, in contrast, organism is at the same granularity as a cell and there are no organ or organ system levels in between. GO's failure to do justice to such distinctions goes hand in hand with the absence in its documentation of any definitions of such ontological terms as "basic structural unit" and "basic functional unit" for example as used in its definition of "cell" given above. For "basic" does not mean simply "smallest".

6.1 Formal Representation of Granularity

We have sketched a formal machinery that is able to do justice to these matters explicitly in our [Bittner and Smith 2003, Smith and Brogard 2003] Here we indicate the outlines of one system by which we could represent the phenomenon of granularity in a formal framework of the sort employed above. To this end we first of all define the function '**gr**' which assigns to each entity its level of granularity. We then define:

$$\mathbf{inst}_{\mathrm{gr}=g}(x, A) =_{\mathrm{def}} \mathbf{inst}(x, A) \;\&\; \mathbf{gr}(x) = y$$

Where *g*, *h* … stand in for the different levels of granularity in the list above. On this basis we can also define a transgranular parthood as follows:

$$A_{gr=g}\textbf{\textit{part}}\textbf{-of}_{gr=h}\ B =_{def} (\textbf{inst}_{gr=g}(x, A) \ \& \ \textbf{inst}_{gr=h}(y, B) \ \& \ \textbf{part}(x, y))$$

We can then write for example:

$$\text{inst}_{gr=subcellular}(x, \textit{subcellular organelle}) \rightarrow \exists y(\textbf{inst}_{gr=cell}(y, \text{cell}) \ \& \ x_{gr=subcellular}\text{part}_{gr=cell}, y)$$

$$\textbf{inst}_{gr=cell}(x, \textit{hepatocyte}) \rightarrow \exists y(\textbf{inst}_{gr=organ}(y, \textit{liver}) \text{ and } x_{gr=cell}\textbf{part}_{gr=organ}y)$$

$$\textbf{inst}_{gr=cell}(x, \textit{cell}) \rightarrow \exists y(\textbf{inst}_{gr=organ}(y, \textit{organ}) \text{ and } x_{gr=cell}\textbf{part}_{gr=organ}y)$$

$$\textbf{inst}_{gr=organ}(x, \textit{organ}) \rightarrow \exists y(\textbf{inst}_{gr=organism}(y, \textit{organism}) \text{ and } x_{gr=organ}\textbf{part}_{gr=organism}y)$$

(Here and in what follows initial universal quantifiers have been suppressed.) The transgranular parthood relation reflects the necessity in biomedical informatics reasoning for what we might call ontological zooming: the problem is to develop the formal means whereby we can relate for example our knowledge of molecular biological phenomena with clinical knowledge of physiology and pathology. We can formalize the definitional core of GO's definition of "cell" as follows:

$$\textbf{inst}_{gr=cell}(x, \textit{cell}) =_{def} \textbf{functional\_unit}\ (x, y)\exists y(\textbf{inst}_{gr=organism}(y, \textit{organism}) \ \& \ \textbf{inst}(x, \textit{structural\_unit}) \ \& \ \textbf{inst}(x, \textit{functional\_unit}) \ \& \ \textbf{part}(x, y)$$

Of course, we still need to provide an account of what is meant by 'structural' and 'functional' unit. Thus we might define what it is for one entity to be a functional part of another entity in terms of the fact that the function of the one is itself a part of the function of the other, as the function of the thyroid gland is a part of the function of the endrocrine system.

Unfortunately however an analysis along these lines is not available to GO, since it comprehends functions at the level of molecules only.

7. Structural Classification of Proteins

Problems analogous to those outlines above are not restricted to GO. Thus they can be found for example in the Structural Classification of Proteins[10] (SCOP), which is the largest protein database for protein structures [Lo Conte *et al*. 2004]. The SCOP database provides a detailed and comprehensive description of the relationships between those proteins whose structure is known, including all entries in the Protein Data Bank, combining together classifications made on the basis of structure and evolution. SCOP divides proteins into the following top-level classes:

All alpha proteins;

[10] http://scop.mrc-lmb.cam.ac.uk/scop/

All beta proteins;
Alpha and beta proteins (a/b);
Alpha and beta proteins (a+b);
Multidomain proteins (alpha and beta);
Membrane and cell surface proteins and peptides;
Small proteins;
Coiled coil proteins;
Low resolution protein structures;
Peptides; and Designed proteins.

One of the main problems with this classification is that SCOP does not apply it consistently. Thus it classifies many membrane proteins not within the class *membrane and cell surface proteins and peptides*, but rather within classes comprehending protein groups on the basis of their structural features. Examples from the version of May 2004 are: Peroxisomal membrane protein, which is classified under *All beta proteins*. Another example is membrane penetration protein mu1, which is classified under *Multi-domain proteins (alpha and beta)*. While it is indeed true that for example a peroxisomal protein is a beta protein, such classifications nonetheless point to an incoherence in the underlying classificatory order. For they imply that we cannot infer from: *x is a membrane protein* to: *x is a membrane and cell surface protein and peptides*. When we move down the SCOP hierarchy, then we find 219 problematic membrane protein designations at the level of folds, 140 and 111 further problematic designations at the level of superfamilies and families, making 470 problematic cases in all.

The problem, for a purportedly *structural* classification of proteins like SCOP, is that not all membrane proteins have a common structure. For example, while 931 of the proteins classified within the axis *Membrane and cell surface proteins and peptides* have transmembrane helices, 158 of them do not. Indeed the question arises whether a class like *Membrane and cell surface proteins and peptides* should be present at all within a properly structural classification, or whether it would not be better to separate out a class labeled *Proteins with transmembrane helices* and relocate the other proteins classified as *Membrane and cell surface proteins and peptides* (some 14.5% of the total) elsewhere. The latter is a class that is based not on protein structure but rather on protein *location*, and it is this mixing of two aspects within a single classification – a mixing that is avoided for example in the purely structural classification which is the FMA – which leads to problems.

8. SWISS-PROT

We can illustrate a different set of problems which derive from failure to abide by consensus ontological principles if we examine SWISS-PROT[11] (now a part of UniProt[12]), a curated protein sequence database which provides descriptions of proteins together with annotations to a variety of further types of data and integration with over 60 proteomics and protein-related databases [Gasteiger *et al.* 2001].

There are two types of data present within SWISSPROT: *core data* and *annotation data*. The former itself consists of sequence data together with citation information and

[11] http://us.expasy.org/sprot/
[12] http://www.expasy.uniprot.org/index.shtml. Uniprot combines SWISS-PROT with the TrEMBL and PIRPSD databases.

taxonomic data (descriptions of the biological source of the protein). The latter consists of data pertaining to the functions of the protein, to its post-translational modifications (for example carbohydrates, phosphorylation, acetylation, GPI-anchor, etc.), domains and sites (for example calcium binding regions, ATP-binding sites, zinc fingers, homeobox, kringle, etc.); to secondary and quaternary structure (for example homodimer, heterotrimer, etc.), similarities to other proteins, diseases associated with deficiencies in the protein, sequence conflicts, variants, and so forth.

8.1 Problems with SWISS-PROT Annotation Data

SWISS-PROT affirms that its core data is well mapped, but "there are problems with some of the annotation data, especially those which are put within 'Comments' and are usually free text." In order to see what some of these problems might be, we investigated the comments included in SWISS-PROT which have some relations to GO. It is now interesting to investigate how far we can align SWISS-PROT in the organization of its annotation data with the more elaborate terminology and classification of GO. The headings used by SWISS-PROT to organize annotation data consist of: allergen, alternative products, biotechnology, catalytic activity, caution, cofactor, database, developmental stage, disease, domain, enzyme regulation, function, induction, mass spectrometry, miscellaneous, pathway, pharmaceutical, polymorphism, posttranslational modification, RNA editing, similarity, subcellular location and tissue specificity. Consider for example the term 'induction', which has 59 non-obsolete counterparts in the GO terminology including *induction of an organ* and *induction of apoptosis*. The problem is that of these 59 terms, 54 belong to GO's biological process ontology, 4 to molecular functions and 1 to cellular component.

The catalytic activity is classified within the molecular function axis. 37 non-obsolete GO terms related to SWISS-PROT's 'catalytic activity' (terms containing 'catalysis' and related grammatical forms) are similarly divided between molecular functions (17), biological processes (13) and cellular components (7). SWISS-PROT's term 'pathway' is associated with 199 GO terms containing 'pathway' or related forms 192 denoting biological processes, 6 denoting molecular function and 1 denoting a cellular component. In any case, therefore, a mapping from SWISSPROT annotations to the corresponding GO terms will be tangled indeed.

8.2 Problems Related to Subcellular Location Annotations

Both SWISS-PROT and GO provide annotations for subcellular locations. The Gene Ontology Annotation (GOA) project[13] [Camon *et al*. 2004] aims to apply GO's vocabulary to a non-redundant set of proteins described in the UniProt and Ensembl databases which together provide complete proteomes for Homo sapiens and certain other organisms. However SWISS-PROT's annotation data pertaining to subcellular locations for its proteins have not been synchronized with this GOA data. In light of the problems with GO's treatment of location, however, as noted in [Smith *et al.* 2004] such a synchronization could at best be only a first step towards representation of location for proteins. For example, SWISS-PROT annotates the protein CCHL_HUMAN P53701: cytochrome c-type heme lyase to the subcellular level: mitochondrial inner membrane. GO, on the other hand, links the same protein to: mitochondrial intermembrane space

[13] http://www.ebi.ac.uk/GOA/

(defined as 'the space between the mitochondrial outer membrane and inner membrane') and also with GO: mitochondrial inner membrane, both of which are described by GO as standing in a part-of relation to: mitochondrion. Comparing the two annotations, we find SWISSPROT has a less compete coverage than GOA. To take matters further, [Lill et al 1992] have specified that the given protein is 'bound to mitochondrial inner membrane and located within the intermembrane space'. Unfortunately we will face obstacles if we try to use the GO framework to do justice to such distinctions since GO does not have relations of the type *is-bound-to* or *is-located-in*. This means that knowledge within the medical texts for example where boundedness to or location at a given subcellular location has important implications regarding a protein's 3D structure and posttranslational modification. Yet – as is well-documented within SWISSPROT itself – such knowledge remains unclaimed where GO is used as annotation framework.

9. Conclusion

The above discussion of the application of philosophical and formal-ontological principles as a means of overcoming certain systematic shortcomings of GO and other biomedical information resources should, we believe, be of interest to information scientists in general. Indeed it can be shown that other ontologies and terminology systems, including lexical databases such as WordNet, suffer from similar shortcomings. Our methodology can thus be generalized to serve as one basis for the quality assurance of information systems in general.

Acknowledgments

Work on this paper was carried out under the auspices of the Wolfgang Paul Program of the Humboldt Foundation and also of the EU Network of Excellence in Semantic Datamining and the project "Forms of Life" sponsored by the Volkswagen Foundation.

Bibliography

Bittner, T and Donnelly, M. The Mereology of Stages and Persistent Entities. In: Proceedings of ECAI'04 (2004).

Bittner, T. and Smith, B. A Theory of Granular Partitions", Foundations of Geographic Information Science, M. Duckham *et al.*, eds., London: Taylor & Francis, 2003, 117–151.

Bittner, T. and Smith, B. Granular Spatio-Temporal Ontologies. 2003 AAAI Symposium: Foundations and Applications of Spatio-Temporal Reasoning (FASTR).

Bittner, T. Axioms for parthood and containment relations in bio-ontologies, Proceedings of KR-Med, 2004.

Bittner, T., Donnelly, M., and Smith, B. Endurants and Perdurants in Directly Depicting Ontologies, AICOM, 2004b.

Bittner, T., Donnelly, M., and Smith, B., Individuals, Universals, Collections: On the Foundational Relations of Ontology. IFOMIS Technical Report, University of Leipzig, 2004.

Camon, E., Magrane, M., Barrell, D., Lee V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., Apweiler R. (2004) The Gene Ontology Annotation (GOA) Database: Sharing Knowledge in Uniprot with Gene Ontology. Nucleic Acids Research Jan 1 32(1): D262-D266 (2004).

Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database, (1998), MIT Press.

Gasteiger E, Jung E, Bairoch A. Swiss-Prot: Connecting Biomolecular Knowledge via a Protein Database. Curr Issues Mol Biol. 2001 Jul;3(3):47-55.

Gene Ontology Consortium. Creating the Gene Ontology Resource: Design and Implementation. Genome Res. 2001; 11: 1425-1433.

Grenon, Pierre, Barry Smith: SNAP and SPAN: Prolegomenon to Geodynamic Ontology. in: Spatial Cognition and Computation, 4: 1 (March 2004), 69–103.

Gruber, T. A Translation Approach to Portable Ontology Specification. Knowledge Acquisition, 1993, 199-220.

Lill R, Stuart RA, Drygas ME, Nargang FE, Neupert W. Import of Cytochrome c Heme Lyase into Mitochondria: A Novel Pathway into the Intermembrane Space. EMBO J. 1992 Feb;11(2):449-56.

Lo Conte L, Brenner SE, Hubbard TJP, Chothia C, Murzin A. (2002). SCOP Database in 2002: Refinements Accommodate Structural Genomics. Nucl. Acid Res. 30(1), 264-267.

Rosse, C. and Mejino, J. L. V. A Reference Ontology for Bioinformatics: The Foundational Model of Anatomy. Journal of Biomedical Informatics. 2003;36:478-500.

Rosse, C. and Shapiro, L. G. and Brinkley, J. F. (1998) The Digital Anatomist Foundational Model: Principles for Defining and Structuring Its Concept Domain. In Proceedings, American Medical Informatics Association Fall Symposium, pages 820-824.

Smith B and Fellbaum, C. Medical WordNet: A New Methodology for the Construction of Information Resources in Consumer Health, *Proceedings of Coling* 2004.

Smith B, Köhler J, Kumar A. On the Application of Formal Principles to Life Science Data: A Case Study in the Gene Ontology. DILS 2004, Leipzig. Lecture Notes in Bioinformatics. 2994; 79-94, 2004.

Smith B, Rosse C. The Role of Foundational Relations in Biomedical Ontology Alignment. (In press) Proc. Medinfo 2004.

Smith B, Williams J, Schulze-Kremer S. The Ontology of the Gene Ontology. In: Proc. Annual Symposium of the American Medical Informatics Association (2003), 609- 613.

Smith B. Basic formal ontology. Technical report, Institute for Formal Ontology and Medical Information Science, University of Leipzig, (2003a)

Smith B. Ontology. In L. Floridi (ed.), Blackwell Companion to Philosophy, Information and Computers, Oxford, 2003.

Smith, B, and Brogaard, B. A Unified Theory of Truth and Reference. Loqique et Analyse: 43 (2003) 49-93.

Smith, B. and Grenon, P. The Cornucopia of Formal-Ontological Relations. *Dialectica*. 2004 (in press)

Smith, B.: The Logic of Biological Classification and the Foundations of Biomedical Ontology (Invited paper). In: Proc. 10th International Conference in Logic Methodology and Philosophy of Science, Oviedo, Spain (2003b)

The Gene Ontology Consortium. Gene Ontology: Tool for the Unification of Biology. Nature Genet. 2000; 25: 25-29.

Wroe CJ, Stevens R, Goble CA. A Methodology to Migrate the Gene Ontology to a Description Logic Environment using DAML+OIL. Pacific Symposium on Biocomputing 2003; 8: 624-635.

Yeh I., Karp P. D., Noy, N. F., Altman, R. B. Knowledge Acquisition, Consistency Checking and Concurrency Control for Gene Ontology (GO). Bioinformatics 2003; 19 (2): 241-248.