On the Application of Formal Principles to Life Science Data: A Case Study in the Gene Ontology

Barry Smith,[1,2] Jacob Köhler[3] and Anand Kumar[2]

[1] Department of Philosophy, University at Buffalo
[2] Institute for Formal Ontology and Medical Information Science, University at Leipzig
[3] Department of Bioinformatics, University of Bielefeld

published in E. Rahm (ed.), *Database Integration in the Life Sciences* (DILS 2004), Berlin: Springer, 2004, 79-94.

Abstract. Formal principles governing best practices in classification and definition have for too long been neglected in the construction of biomedical ontologies, in ways which have important negative consequences for data integration and ontology alignment. We argue that the use of such principles in ontology construction can serve as a valuable tool in error-detection and also in supporting reliable manual curation. We argue also that such principles are a prerequisite for the successful application of advanced data integration techniques such as ontology-based multi-database querying, automated ontology alignment and ontology-based text-mining. These theses are illustrated by means of a case study of the Gene Ontology, a project of increasing importance within the field of biomedical data integration.

1 Introduction

In order to integrate databases, one has to overcome problems at several levels. These include *legal* problems related to access and redistribution, *technical* problems related to the employment of heterogeneous storage and access methods and query languages, and *housekeeping* problems relating to the management of corrections and updates. Here we are concerned with what are loosely called *semantic* obstacles to database integration [1], problems which manifest themselves both at the database schema level and at the level of data entry [2]. In simple terms, database systems should use language consistently, which means: they should use the same terms to refer to the same things. Semantic heterogeneities at the schema level arise where different names are used for equivalent database fields and tables. Systems to overcome schema level semantic heterogeneities are introduced in [3-8]. Semantic heterogeneities arise at the entry level where different terms are used for the same things, or the same terms are used for different things.

Ontologies have thus far played a role in the semantic integration of databases at the entry level by providing controlled vocabularies which have the goal of making it possible to search different databases secure in the knowledge that the same terms, wherever they are encountered, will always represent the same entities. One of the most impressive and influential developments in this respect is the Gene Ontology™

(GO) [9], which plays a central role in attempts to develop controlled vocabularies for shared use across different biological domains. The tremendous investment of time and effort by the GO Consortium has already brought considerable benefits to a range of different types of biological and biomedical research.

GO's December 2003 release contains some 16,658 terms divided into three networks whose topmost nodes are, respectively: *cellular component*, *molecular function* and *biological process*. In contradistinction to much existing usage, we shall refer to the nodes in such hierarchies not as *concepts* but rather on the one hand as *terms* (where we are interested in the hierarchies themselves, as syntactic structures), and on the other hand as *classes* (where we are interested in the biological entities to which these terms refer). It is classes, not concepts, which stand in *is a* and *part of* relations [10]. They do this in virtue of specific kinds of relations between the individual entities which instantiate these classes [10].

GO's three networks are structured by the relations of subsumption (*is a*) and of partonomic inclusion (*part of*). Crucially, GO treats its three structured networks as separate ontologies, which means that no ontological relations are defined between them. Thus for example GO does not record the fact that a given function is the *function of* a given component, or that a given process is the *exercise of* a given function. One reason for this is that the molecular functions and biological processes are standardly exercised by entities instantiating classes which fall outside the scope of GO. GO has an ontology of cellular components, not however of molecular structures or (with some few exceptions, such as *host* [GO:0018995]) of organism-level structures. Thus GO includes *behavior* [GO:0007610] and *adult behavior* [GO:0030534], defined as 'Behavior in a fully developed and mature organism', within its biological process hierarchies, but it has no room for *adult* or *organism*.

The decision of the GO Consortium to develop its three separate ontologies in this way has brought a variety of benefits, and we do not here recommend its abandonment. What we do recommend is that, in the process of constructing GO and of similar biomedical information resources in the future, special attention be paid to the problems which inevitably arise where standard rules of good classification are not respected – rules which have been tested in ontology research and in associated formal disciplines since the time of Aristotle and some examples of which are outlined below.

In the case of GO, these problems manifest themselves in characteristic shortfalls in formal integrity, turning on uncertainty as to the relations between the classes in GO's three ontologies. For example, there is no linkage between the term *taste* [GO:0007607], which is a biological process term, and the term *taste receptor activity* [GO:0008527], which is a molecular function term. This has an adverse impact on data integration when GO's data structure is used to retrieve or link related information from different data sources [11].

As we shall see, GO's three ontologies are not primarily used to support automated reasoning, but rather as a means by which biologists can easily and rapidly locate existing terms and relations within a structured vocabulary and so determine the proper treatment of new terms when the latter present themselves as candidates for incorporation in biomedical databases. To this end, GO has been highly successful in providing a controlled vocabulary that is used by different research groups in annotating genes and gene products. If, however, GO is to reap its full potential in

supporting data integration across the life sciences, then shortfalls from formal integrity of the sorts here isolated become important. Here we shall focus on isolating such shortfalls and on suggestions as how they might be avoided in the future. Such avoidance is, we shall argue, a prerequisite for the use of GO in conjunction with tools for advanced data integration of a sort which meet the standards of contemporary biomedical informatics, including tools for alignment of ontologies [12, 13], ontology based text-mining [14-16] and 'intelligent' multi-database querying.

GO has often been criticized for its inconsistencies and for its lack of clear guidance as to what the relations between its three ontologies are. However, the literature thus far has not provided practical solutions to these problems in terms of general rules as to how they can be avoided in the future. Such problems cannot generally be overcome in an automated way. Rather, they require that the process of design and manual curation follow the general rules of classification which are familiar from the literature of logic and philosophy. Certainly biologists can avoid some inconsistencies via the use of ontology editing tools such as Protégé-2000, which support some automated consistency checking. However, by using Protégé-2000 the authors of [17] were able to identify only some minor inconsistencies in GO. It is, however, possible to take advantage of the fact that the Gene Ontology, like most of the controlled vocabularies united under the Open Biological Ontologies (OBO) umbrella, is maintained using the relatively user-friendly tool DAG-Edit [13]. One message of what follows is that one step towards greater formal coherence can be achieved by building sound policy into DAG-Edit and similar tools in ways which ensure the avoidance of certain kinds of errors, and we are pleased to see that OBO (http://obo.sourceforge.net/list.shtml) has defined relationships in development.

2 Problems with *Part of*

GO seeks to establish a 'controlled vocabulary'. This means that it accepts the rule of *univocity*: terms should have the same meanings (and thus point to the same referents) on every occasion of use. Unfortunately GO breaks this rule already in its treatment of the *part of* relation, which is at the very center of its hierarchical organization and with which at least three different meanings are associated [18]:

P1. *A part of B* means: *A is sometimes part of B* in the sense that there is for each instance of *A* some time at which it is part of an instance of *B* (in the standard mereological sense of 'part' as a relation between particulars). Example: *replication fork* (is at some times during the cell cycle) *part of nucleoplasm*.

P2. *A part of B* means: *A can be part of B* in the sense of a time-independent inclusion relation between classes, which we can summarize as: class *A* is part of class *B* if and only if there is some sub-class *C* of *B* which is such that all instances of *A* are included as parts in instances of *C* and all instances of *C* have parts which are instances of *A*. Example: *flagellum part of cell* (some types of cells have flagella as parts).

P3. *A part of B* means: vocabulary *A* is included within vocabulary *B*. Example: *cellular component ontology part of gene ontology.*

GO's 'part of' violates not only *univocity* but also two other rules at the heart of good practice in the establishment of a formally rigorous framework of definitions: a term with an established use (inside and outside biomedical ontology) is used with a new, non-standard use; a lexically simple term is used to represent a lexically complex concept that is standardly expressed by means of phrases including the lexically simple term as part. To define 'part of' as meaning 'can be part of' is rather like defining 'sulphur' as meaning 'organo-sulphur compound.'

Solution: At present, DAG-Edit is shipped with only one built-in relation type, namely '*is a*', and even the latter can be deleted by the user at will. By including a fixed set of well-defined relation types that cannot be removed or modified by the user, biologists using DAG-Edit would become aware of the different relation types at their disposal. Whenever a user employs a relation type such as '*part of*', a menu should pop up which offers a list of alternative more specific relation types, such as *is localized in* or *is involved in.* This would go far towards solving the problems which currently arise when OBO ontologies built with DAG-Edit use different names and identifiers for the same relation types and associate different relation types with the same names and identifiers.

3 Problems with *Is a*

GO's three term hierarchies have some obvious relation to the species and genera of more traditional biology When we evaluate GO as a classification of biological phenomena, however, then we discover that GO often uses *is a*, too, in ways which violate *univocity.* This is in part because GO is confined to the two relations *is a* and *part of* and because it does not allow ontological relations between its three separate ontologies. Thus it is constrained to place too great a load on *is a* than this relation is capable of bearing.

Thus when GO postulates:

[1] cell differentiation *is a* cellular process

[2] cell differentiation *is a* development,

then it means two different things by '*is a*', and only in the former case do we have to deal with a true subsumption relation between biological classes. That there is a problem with the latter case can be seen by noting that, where the agent or subject of differentiation is the *cell*, the agent or subject of development is the whole organism. That one process or function class subsumes another process or function class, implies, however, that same subject or agent is involved in each. This implies, however, that the definition:

GO:0007275 **Development**

> **Definition:** *Biological processes specifically aimed at* the progression of an organism over time from an initial condition (e.g. a zygote, or a young adult) to a later condition (e.g. a multicellular animal or an aged adult)

should be modified by deleting the italicized portion, and that the relation in [2] should then more properly be expressed as: *contributes to*. Or alternatively, some way must be found to ensure that 'development' is used consistently in all its occurrences in GO, to refer either always to processes whose bearers are the whole organism, or to processes whose bearers may be individual cells, organs, or other organism constituents.

Another example of uncertainty regarding GO's *is a* relation is illustrated by:

[3] hexose biosynthesis *is a* monosaccharide biosynthesis

[4] hexose biosynthesis *is a* hexose metabolism,

Here the second *is a* seems more properly to amount to a *part of* relation, since *hexose biosynthesis* is just that part of hexose metabolism in which hexose is synthesized.

GO postulates:

[5] vacuole (sensu Fungi) *is a* storage vacuole

[6] vacuole (sensu Fungi) *is a* lytic vacuole,

where 'sensu' is the operator introduced by GO 'to cope with those cases where a word or phrase has different meanings when applied to different organisms,' (http://www.geneontology.org/doc/GO.usage.html#sensu).

Lytic vacuole is defined by GO as meaning: a vacuole that is maintained at an acidic pH and which contains degradative enzymes, including a wide variety of acid hydrolases. Inspection now reveals that '*is a*', here, stands in neither case for a genuine subsumption relation between biological classes. Rather, it signifies on the one hand the assignment of a *function* or *role*, and on the other hand the assignment of special features to the entities in question. Certainly there are, in some sense of the term 'class', classes of storage vacuoles and of lytic vacuoles; and certainly it is the case that all instances of *vacuole* (*sensu Fungi*) are instances of *storage vacuole* and of *lytic vacuole*. But that such relations obtain is not as yet sufficient for the existence of a genuine *is a* relation. A box used for storage is not (*ipso facto*) a special *kind* of box; rather it is a box which plays a special role in certain contexts. And similarly 'animal belonging to the emperor' does not designate a special *kind* of animal.

In Figure 1,

GO:0000327: **lytic vacuole within protein storage vacuole**

is recorded as standing in an *is a* relation to *protein storage vacuole*. In fact, however, we have to deal here with one of several ways in which GO expresses relations of *location*.

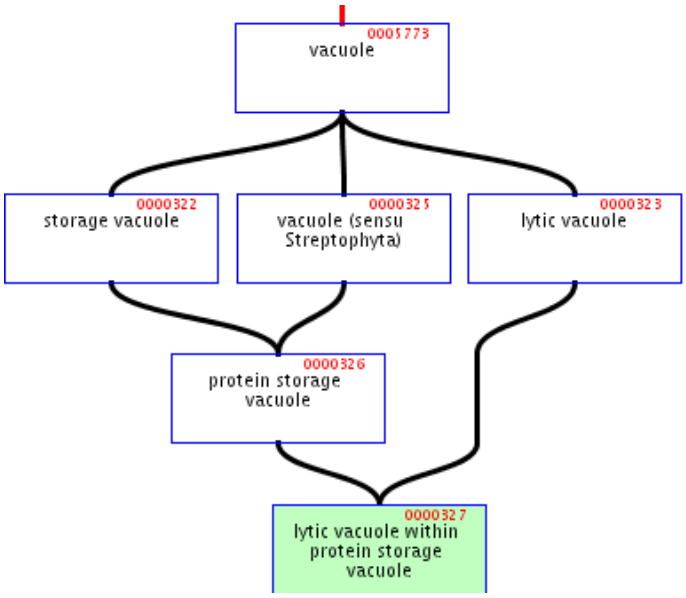

Figure 1: Treatment of *Vacuole* in GO
(taken from the QuickGO browser: http://www.ebi.ac.uk/ego)

Another such example is GO's term 'unlocalized,' which is, and was always intended to be, a temporary term, but the consideration of which is nonetheless illuminating:

> GO:0005941: **unlocalized**
> **Definition:** Used as a holding place for cellular components whose precise localization is, as yet, unknown, or has not been determined by GO (the latter is the major reason for nodes to have this parent); this term should not be used for annotation of gene products,

is used for example in statements such as:

[7] Holliday junction helicase complex *is a* unlocalized

[8] Unlocalized *is a* cellular component

[7] and [8] illustrate GO's shortfall from two further principles:

> *positivity*: complements of classes are not themselves classes. (Terms such as 'non-mammal' or 'non-membrane' do not designate natural kinds.)

> *objectivity*: which classes exist is not a function of the current state of our biological knowledge. (Terms incorporating expressions such as 'unclassified' or 'with unknown ligand' do not designate biological natural kinds.)

Shortfalls from *positivity* are illustrated also by terms like 'non-muscle myosin'; shortfalls from *objectivity* by terms like 'G-protein coupled receptor activity, unknown ligand.'

Another rule of good classification familiar from the logico-philosophical literature is:

> *levels*: the terms in a classificatory hierarchy should be divided into pre-determined levels (analogous to the levels of kingdom, phylum, class, order, etc., in traditional biology).

(Similarly, terms in a *partonomic* hierarchy should be divided into predetermined granularity levels, for example: organism, organ, cell, molecule, etc. GO's distinction between *biological* processes*, cellular* components and *molecular* functions/activities reflects this partonomic principle of levels, and GO currently contemplates a move to distinguish 'cell physiological process' and 'organism physiological process' in this same spirit.)

Theorists of classification have recognized further that the terms on each such level should ideally satisfy the rules:

> *single inheritance*: no class in a classificatory hierarchy should have more than one parent on the immediate higher level

> *exhaustiveness*: the classes on any given level should exhaust the domain of the classificatory hierarchy

These two rules together constitute the so-called JEPD (for: jointly exhaustive and pairwise disjoint) criterion. *Exhaustiveness* is of course often difficult to satisfy in the realm of biological phenomena. Shortfalls from *single inheritance* are however easy to detect, and their acceptance thus amounts to the rejection of the JEPD ideal.

The acceptance of the latter in traditional classifications is no accident. Exhaustiveness is a clear positive trait for a classificatory hierarchy; its acceptance in some form is presupposed as a goal by every biological scientist. Single inheritance reflects the presumption that if a term in a classificatory hierarchy has two *is a* parents, then the hierarchy needs to be refined [19]. Nowadays, however, single inheritance is less commonly accepted as a positive trait because multiple inheritance is so useful a device in facilitating compactness and efficient maintenance of large-scale ontologies. This is because it allows one to make large changes to a portion of an ontology without the need to adjust each individual lower-level element. It also allows one to avoid certain kinds of combinatory explosion. On the other hand, as will become clear from many of the examples treated in this paper, shortfalls from *single inheritance* are often clues to bad coding. This is because such shortfalls mark often subtle deviations from a reading of '*is a*' as standing for a genuine class subsumption relation. Such deviations are difficult to communicate to curators in terms of generally

intelligible principles. But more importantly, they also serve as obstacles too ontology integration, since they amount to the conscious adoption of a policy according to which '*is a*' means different things in different contexts. [11]

We here leave open the question whether division into levels and single inheritance involving genuine *is a* relations can be achieved throughout the realm of classifications treated of by GO. We note only that, as Guarino and Welty [19] have shown, methods exist for systematically removing cases of multiple inheritance from class hierarchies by distinguishing *is a* relations from ontological relations of other sorts. Using these methods, well-structured classifications can be achieved by recognizing additional relations (for example: *has role*, *is dependent on*, *is involved in*, *contributes to*, *is located in*) and by allowing categories of entities of different sorts (for instance *sites*, *constituents*, *roles*, *functions*, *qualities*) within a single ontology. GO, however, does not have these alternatives at its disposal, not least because of its insistence that its three constituent vocabularies represent separate ontologies with no relations defined between them. At the same time, however, we will still need to find ways to represent in a formally coherent way those cases, such as *storage vacuole* or *immunological synapse*, where role and bearer together are referred to as forming a single entity.

4 Problems with Definitions

Like other biomedical ontologies, GO provides not only controlled vocabularies with hierarchical structures but also definitions of its terms. Indeed part of the goal of GO is to provide a source of strict definitions that can be communicated across people and applications. The definitions actually supplied by the GO Consortium, however, are affected by a number of characteristic problems which, while perhaps not affecting their usability by human biologists, raise severe obstacles at the point where the sort of formal rigor needed by computer applications is an issue, to the degree that much of the information content contained in GO's definitions is not accessible to software.

A well-structured definition should satisfy at least the rule:

> *intelligibility*: the terms used in a definition should be simpler (more intelligible, more logically or ontologically basic) than the term to be defined

– for otherwise the definition would provide no assistance to the understanding, and thus make no contribution to the GO Consortium's goal of formulating definitions which are usable by human beings.

That GO does not respect this rule is illustrated for example by:

> GO:0007512: **adult heart development**
> **Definition:** Generation and development of the heart of a fully developed and mature organism
>
> GO:0017145: **stem cell renewal**

> **Definition:** The self-renewing division of a stem cell. A stem cell is an undifferentiated cell, in the embryo or adult, that can undergo unlimited division and give rise to one or several different cell types.

Illustrates two features of GO's shortfall from standards of good practice in the formulation of its definitions: 1. the failure to draw a clear line between providing *definitions* of terms and *communicating extra knowledge*, 2. the inclusion within the definition of a GO of the definition of some second term, that is not present in GO. The solution to the latter problem is to provide a supplementary list of all non-GO terms treated in this fashion, together with their definitions. It would then be an interesting exercise to see what sort of ontology the terms included in this list would define.

Once again, the failure to follow a basic rule of classification and definition is a good clue as to the presence of coding errors. Thus consider:

> GO:0016326: **kinesin motor activity**
> **Definition:** The hydrolysis of ATP (and GTP) that drives the microtubular motor along microtubules,

The problem here is that hydrolysis is merely that which provides the energy for motor activity and not this activity as a whole. (Note that this term has been declared obsolete by GO, not however because it represents an error, but because it represents a gene product, and the latter do not fall within the scope of GO.)

> GO:0008527: **taste receptor activity**
> **Definition:** Combining with soluble compounds to initiate a change in cell activity. These receptors are responsible for the sense of taste.

fails to satisfy a further standard principle of good practice in definition, namely that of

> *substitutivity*: in all so-called extensional contexts a defined term must be substitutable by its definition in such a way that the result is both grammatically correct and has the same truth-value as the sentence with which we begin.

(where an 'extensional context' is a context not within the scope of mental verb phrases such as 'I believe that', 'She denies that', 'He knows that'). Adhering to the rule of substitutivity ensures *inter alia* that definitions have the same grammatical form as the terms defined. Failures to adhere to this rule are particularly egregious in connection with GO's molecular function hierarchy.

It is not possible, on pain of infinite regress, to define every term in accordance with the principle of *intelligibility*. Rather, some terms must be left undefined. Definitions should more generally satisfy the rule of *modularity*, which means that they should be organized into levels, with level 0 terms being picked out as undefined primitives and terms on levels $n + 1$, for each $n \geq 0$ being defined by appeal exclusively to logical and ontological constants (such as 'and', 'all', 'is a' and 'part

of ') together with already defined terms taken from levels ≤ *n*. Modular definitions are especially useful for the purposes of automatic reasoning.

Because the rules of *univocity, modularity* and *substitutivity* are not satisfied by GO's system of definitions, this means that substitution of the GO definitions for the corresponding defined terms appearing within other contexts can be achieved, at best, only with human intervention. A valuable source of automatic error-checking and of location of classificatory gaps is hereby sacrificed.

Consider the example:

> GO:0003673: **cell fate commitment**
> **Definition:** The commitment of cells to specific cell fates and their capacity to differentiate into particular kinds of cells.

The coarse logic of this definition is as follows:

$$x \text{ is a cell fate commitment} =_{\text{def}} x \text{ is a cell fate commitment and } p,$$

where *p* is, logically speaking, a second, extraneous condition. Here GO errs by providing in its definition at the same time too much and too little information, and the user does not know how to interpret the relation of the two clauses in the definition. (The first clause is marked, in addition, by the problem of circularity.)

If the rule of *modularity* is satisfied, then this means that in the case of compound terms the corresponding definitions should themselves be capable of being generated automatically. Thus for example the definition of 'garland cell differentiation' should be obtainable by taking the definition of 'cell differentiation' and substituting 'garland cell' for 'cell' throughout. What we find, however, is:

> GO:0030154: **cell differentiation**
> **Definition:** The process whereby relatively unspecialized cells, e.g. embryonic or regenerative cells, acquire specialized structural and/or functional features that characterize the cells, tissues, or organs of the mature organism or some other relatively stable phase of the organism's life history.

> GO:0007514: **garland cell differentiation**
> **Definition:** Development of garland cells, a small group of nephrocytes which take up waste materials from the hemolymph by endocytosis.

This leaves the user in a position where he does not know whether 'differentiation' as it occurs in the two contexts does or does not mean the same thing. GO's definition of 'garland cell differentiation' is marked further by the problem that it is in fact a definition of garland cell *development*, of which garland cell *differentiation* would in fact (given the definition of *development* provided earlier) be a proper subclass.

Solution: methods exist which can be used to control definitions for conformity with *modularity*, for example by highlighting words that occur in a definition even though they are located deeper in the *is a* hierarchy. Potential problems of circularity can also be indicated by highlighting non-logical words that occur in both a term and its definition. Compound terms can be recognized, and automatically generated

definitions can be suggested to the user in terms of already existing definitions of the component parts.

5 Problems with 'Sensu'

A related set of problems can be illustrated by examining GO's use of its 'sensu' operator, which is introduced to cope with those cases where a word or phrase has different meanings when applied to different organisms, as for example in the case of *cell wall.* (Cell walls for example in bacteria and fungi have a completely different composition.) 'Using the *sensu* reference makes the node available to other species that use the same process/function/component' (http://www.geneontology.org/doc-/GO.usage.html#sensu). If, however, 'sensu' is designed to indicate that the modified term refers to a different class from that to which the unmodified term refers, then in what sense *are* we still dealing with 'the same process/function/component'?

Since the primary goal of the GO Consortium is to provide an ontology of gene products applying to all species, *sensu* terms are intended to be used sparingly. In consequence, *sensu* terms, as in the case illustrated in Figure 2, are allowed to have non-*sensu* terms as children, as in

[9] R7 differentiation *is a* eye photoreceptor differentiation (sensu Drosophilia).

For again, there is R7 differentiation in species other than Drosophilia, for example in crustaceans [20].

Another problematic example, which also illustrates once more GO's multiple ways of handling the relation of localization, is GO's postulation of:

[10] bud tip *is a* site of polarized growth (sensu Saccharomyces)

from which we can infer that:

[11] every instance of bud tip has an instance of Saccaromyces polarized growth located therein.

But GO also has:

[12] site of polarized growth (sensu Saccharomyces) *is a* site of polarized growth (sensu Fungi)

from which we can infer that:

[13] bud tip *is a* site of polarized growth (sensu Fungi)

and from there to:

[14] every instance of bud tip has an instance of Fungus polarized growth located therein.

[11] is consistent with [14] only if either (a) every instance of non-Saccharomyces Fungus polarized growth is co-localized with an instance of Saccharomyces polarized group or (b) there is Fungus polarized growth only in Saccharomyces. (a) we take to be biologically false; but (b) implies that 'site of polarized growth (sensu Saccharomyces)' and 'site of polarized growth (sensu Fungi)' in fact refer, confusingly, to the same class, and thus that the latter should be removed from GO's cellular component ontology.

A further problem is caused by GO's use of 'sensu Invertebrata'. Whereas *vertebrate* is a well-defined biological taxon, biologists tend to disagree on what the definition of *invertebrate* should be, and thus apply the 'sensu Invertebrata' modifier to different taxa. The resultant errors are illustrated for example in the genes annotated to

GO:0006960 : **antimicrobial humoral response (sensu Invertebrata)**

in the browser AMIGO, many of which are not invertebrate genes but rather human genes (for example COPE HUMAN, PTGE HUMAN, PTE1 HUMAN, and so on). It is surely obvious that a gene with suffix 'HUMAN' should not be annotated to a biological process which is assigned to invertebrates.

Solution: These and other problems can be overcome by introducing 'sensu' as a relational expression with a well-defined meaning that references a systematic species nomenclature such as the TAXON database [21]. We understand that GO is contemplating taking steps along these lines. In addition measures can be taken to check automatically that all GO terms that occur in a given branch of an '*is a*' hierarchy use the same taxon. The problem with 'Invertebrata' is also overcome, in virtue of the fact that no standard systematic species nomenclature contains this term.

7 Problems with 'Function'

Recall GO's definition of 'toxin activity' as: 'Acts as to cause injury to other living organisms.' The problem here flows from GO's unstable view of how the classes in molecular function ontology should precisely be designated [22]. The same problem makes itself manifest also in cases such as:

GO:0005199: **structural constituent of cell wall**
Definition: The action of a molecule that contributes to the structural integrity of a cell wall,

where the definition confuses *constituents* (which ought properly to be included in GO's constituent ontology) with *activities*, which GO includes in its function ontology. Many other constituents are similarly defined as activities:

extracellular matrix structural constituent
puparial glue (sensu Diptera)
structural constituent of bone
structural constituent of chorion (sensu Insecta)
structural constituent of chromatin
structural constituent of cuticle
structural constituent of cytoskeleton
structural constituent of epidermis
structural constituent of eye lens
structural constituent of muscle
structural constituent of myelin sheath
structural constituent of nuclear pore
structural constituent of peritrophic membrane (sensu Insecta)
structural constituent of ribosome
structural constituent of tooth enamel
structural constituent of vitelline membrane (sensu Insecta)

8 An Alternative Regime of Definitions

As a brief illustration of a regime of definitions built up in such a way as to satisfy the principles listed above, we consider the Foundational Model of Anatomy (FMA), which is being developed at the University of Washington, Seattle as part of the Digital Anatomist Project. The term hierarchy of the FMA consists in a symbolic representation of the structural organization of the human body from the macromolecular to the macroscopic levels, with the goal of providing a robust and consistent scheme for classifying anatomical entities which can serve as a reference ontology in biomedical informatics [23].

FMA seeks to follow the formal rules for definitions laid down by Aristotle. A definition, on this account, is the specification of the essence (nature, invariant structure) shared by all the members of a class or natural kind. Definitions are specified by working through a classificatory hierarchy from the top down, with the relevant topmost node or nodes acting in every case as undefinable primitives. The definition of a class lower down in the hierarchy is then provided by specifying the parent of the class (which in a regime conforming to single inheritance is of course in every case unique) together with the relevant differentia, which tells us what marks out instances of the defined class or species within the wider parent class or genus, as in: *human = rational animal*, where *rational* is the differentia. An Aristotelian definition then satisfies the condition that an entity satisfies the defining condition if and only if it instantiates the corresponding class.

Thus definitions in FMA look like this:

> **Cell** *is a* **anatomical structure** that *consists of* **cytoplasm** *surrounded by* a **plasma membrane** with or without a **cell nucleus**
>
> **Plasma membrane** *is a* **cell part** that *surrounds* the **cytoplasm,**

where terms picked out in bold are nodes within the FMA classification and italicized terms signify the formal-ontological relations – including *is a* – which obtain between the corresponding classes.

As the FMA points out, ontologies 'differ from dictionaries in both their nature and purpose' [24]. Dictionaries are prepared for human beings; their merely nominal definitions can employ the unregimented resources of natural language, can tolerate circularities and all manner of idiosyncrasy. In ontologies designed to be usable by computers, however, definitions must be formally regimented to a much higher degree. The specific type of regimentation chosen by the FMA has the advantage that each definition reflects the position in the hierarchy to which a defined term belongs. Indeed the position of a term within the hierarchy enriches its own definition by incorporating automatically the definitions of all the terms above it. The resultant system of definitions brings the benefit that the entire information content of the FMA's term hierarchy can be translated very cleanly into a computer representation, and brings also further benefits in terms of reliable curation, efficient error checking and information retrieval, and ease of alignment with neighboring ontologies. The FMA defines an ontology as a 'true inheritance hierarchy' and it thereby draws attention to the fact that one central reason for adopting the method of ontologies in supporting reasoning across large bodies of data is precisely the fact that this method allows the exploitation of the inheritance of properties along paths of *is a* relations. FMA's regime of definitions – unlike that of GO – gives due merit to this principle.

9 Conclusion

We are not proposing here that GO abandon all its current practices in structuring its ontologies and accompanying definitions. The world of biomedical research is clearly not concerned with all of those sorts of scrupulousness that are important in more formal disciplines. Rather, it is a world of difficult trade-offs, in which the benefits of formal (logical and ontological) rigor need to be balanced on the one hand against the constraints of computer tractability, and on the other hand against the needs of practicing biologists. All the formal rules presented above should therefore be conceived as rules of thumb, or as ideals to be borne in mind in practice, rather than as iron requirements.

Note, too, that we are not suggesting that the problems outlined in the above have led to concrete errors in the annotations of genes by third parties, for example in the construction of model organism databases. We hypothesize that the biologists who are responsible for such annotations are able to use their biological expertise in order to block the faulty inferences which would otherwise result. To the extent, however, that GO is pressed into service as a reference-platform for the *automatic* navigation between biomedical databases, then the issue of consistency with standard principles of classification and definition will come to be of increasing importance.

Some of the mentioned problems can be overcome via relatively minor modifications to DAG-Edit, which would have a significant impact on the design and reliability of GO's ontologies since they would sharpen the awareness of designers and users

in ways which can lead both to the avoidance of common pitfalls in the course of curation and to an ontologically more coherent structuring of the resultant data. The advantage in incorporating these changes into DAG-Edit would be also that it would not require that GO and the other OBO ontologies be rebuilt from scratch: actual and potential inconsistencies would be highlighted, and can be corrected on the fly.

Multiple inheritance, to repeat, is a particularly important cause of problems in the guise of both coding errors and obstacles to the coherent alignment of ontologies that will be needed in the future. This is because the success of such alignment depends crucially on the degree to which the basic ontological relations – above all relations such as *is a* and *part of* – can be relied on as having the same meanings in the different ontologies to be aligned. Cases of multiple inheritance go hand in hand, at least in many cases, with the assignment to the *is a* relation of a plurality of different meanings within a single ontology. The resultant mélange makes coherent integration across ontologies achievable (at best) only under the guidance of human beings with the sorts of biological knowledge which can override the mismatches which otherwise threaten to arise. This, however, is to defeat the very purpose of constructing bioinformatics ontologies as the basis for a new kind of biological and biomedical research designed to exploit the power of computers.

As Ogren *et al.* [25] have pointed out, almost two-thirds of all GO terms contain other GO terms as substrings, the including term being in many cases derived from the included term via operators such as 'regulation of ' or 'sensu'. Many of the latter recur consistently in certain kinds of subtrees of GO's three ontologies, and in ways which reflect ontologically significant relations between the corresponding classes. Ogren *et al.* propose that the presence of these operators be exploited 'to make the information in GO more computationally accessible, to construct a conceptually richer representation of the data encoded in the ontology, and to assist in the analysis of natural language texts.' We suggest taking this proposal still further by building the corresponding machinery for enforcing compositionality into the DAG-Edit tool and by exploiting analogous compositionality of information on the side of GO's definitions. Such proposals will, however, bear fruit only to the extent that GO's classifications and definitions satisfy the formal principles set forth above.

Acknowledgements
This paper was written under the auspices of the Wolfgang Paul Program of the Alexander von Humboldt Foundation. Thanks are due also to Stuart Aitken, Bert R. E. Klagges, Steffen Schulze-Kremer, Jane Lomax, Cornelius Rosse and Jean-Luc Verschelde.

References

[1] Köhler, J.: Integration of Life Science Databases. BioSilico March (2004)
[2] Kim, W., Seo, J.: Classifying Schematic and Data Heterogeneity in Multidatabase Systems. IEEE COMPUTER 24 (1991) 12-18
[3] Madhavan, J., Bernstein, P. A., Rahm, E.: Generic Schema Matching with Cupid. In: Proc. 27th Int. Conf. on Very Large Data Bases (VLDB) (2001)
[4] Köhler, J., Philippi, S., Lange, M.: SEMEDA: Ontology Based Semantic Integration of Biological Databases. Bioinformatics 19 (2003)

[5] Köhler, J., Lange, M., Hofestädt, R., Schulze-Kremer, S.: Logical and Semantic Database Integration. In: Proc. Bioinformatics and Biomedical Engineering (2000) 77-80
[6] Baker, P. G., Brass, A., Bechhofer, S., Goble, C., Paton, N., Stevens, R.: TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. An Overview. In: Proc. sixth International Conference on Intelligent Systems for Molecular Biology (1998)
[7] Stevens, R., Baker, P., Bechhofer, S., Ng, G., Jacoby, A., Paton, N. W., Goble, C. A., Brass, A.: TAMBIS: transparent access to multiple bioinformatics information sources. Bioinformatics 16 (2000) 184-5
[8] Ludäscher, B., Gupta, A., Martone, M. E.: Model-Based Mediation with Domain Maps. In: Proc. 17th Intl. Conference on Data Engineering (ICDE), (2001)
[9] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., Sherlock, G.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25 (2000) 25-9.
[10] 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, 2003. (2003)
[11] Lord, P. W., Stevens, R. D., Brass, A., Goble, C. A.: Semantic similarity measures as tools for exploring the gene ontology. Pac Symp Biocomput (2003) 601-12
[12] Knight, K., Luk, S.: Building a Large-Scale Knowledge Base for Machine Translation. In: Proc. National Conference on Artificial Intelligence - AAAI (1994)
[13] Lambrix, P., Habbouche, M., Perez, M.: Evaluation of ontology development tools for bioinformatics. Bioinformatics 19 (2003) 1564-71
[14] Kim, J. D., Ohta, T., Tateisi, Y., Tsujii, J.: GENIA corpus-a semantically annotated corpus for bio-textmining. Bioinformatics 19 Suppl 1 (2003) I180-I182
[15] Nenadic, G., Mima, H., Spasic, I., Ananiadou, S., Tsujii, J.: Terminology-driven literature mining and knowledge acquisition in biomedicine. Int J Med Inf 67 (2002) 33-48
[16] Chiang, J. H., Yu, H. C.: MeKE: discovering the functions of gene products from biomedical literature via sentence alignment. Bioinformatics 19 (2003) 1417-22
[17] Yeh, I., Karp, P. D., Noy, N. F., Altman, R. B.: Knowledge acquisition, consistency checking and concurrency control for Gene Ontology (GO). Bioinformatics 19 (2003) 241-8
[18] Smith, B., Rosse, C.: The Role of Foundational Relations in Biomedical Ontology Alignment. (under review)
[19] Guarino, N., Welty, C.: Identity and subsumption. In: R. Green, C. A. Bean, and S. Hyon Myaeng, (eds.): The Semantics of Relationships: An Interdisciplinary Perspective. Kluwer Academic Publishers (2002) 111–126
[20] Hafner, G. S., Tokarski, T. R.: Retinal development in the lobster Homarus americanus. Comparison with compound eyes of insects and other crustaceans. Cell Tissue Res 305 (2001) 147-58
[21] Wheeler, D. L., Church, D. M., Federhen, S., Lash, A. E., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Sequeira, E., Tatusova, T. A., Wagner, L.: Database resources of the National Center for Biotechnology. Nucleic Acids Res 31 (2003) 28-33
[22] 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
[23] Rosse, C., Mejino Jr., J. L. V.: A Reference Ontology for Bioinformatics: The Foundational Model of Anatomy. Journal of Biomedical Informatics in press (2003)
[24] Michael, J., Mejino, J. L., Jr., Rosse, C.: The role of definitions in biomedical concept representation. Proc AMIA Symp (2001) 463-7

[25] Ogren, P. V., Cohen, K. B., Acquaah-Mensah, G. K., Eberlein, J., Hunter, L. T.: The Compositional Structure of Gene Ontology Terms. In: Proc. Proceedings of the Pacific Symposium on Biocomputing - PSB (2004)