Enhancing the Gene Ontology using text mining and design principles for formal ontologies

Jacob Köhler[1], Barry Smith[2,3], Katherine Munn[2], Alexander Rüegg[4], Andre Skusa[4]

[1] Rothamsted Research, Harpenden, UK
[2] Department of Philosophy, University at Buffalo
[3] Institute for Formal Ontology and Medical Information Science, University of the Saarland (as of August)
[4] Technical Faculty, Bielefeld University, Germany

Abstract

Motivation: The Gene Ontology (GO) is currently one of the most important computational resources for molecular biology and bioinformatics. Several recent papers have criticized certain shortcomings in GO, pointing in particular to examples of characteristic types of errors which flow from the failure to address basic ontological principles. The magnitude of such inconsistencies has however not yet been estimated, and no methods have thus far been proposed which would allow GO's curators to pinpoint flawed terms or definitions in a systematic way.

Results: By using computational methods based on ontology design principles we were able to isolate a significant subset of problematic GO terms. By aligning GO to other external ontologies we were able to propose alternative synonyms and definitions for some of these problematic terms, though we discovered that only in a very few cases do these other ontologies contain definitions of the corresponding terms which are superior to those supplied by GO.

Finally, we discuss how GO curators can use ontology design principles drawn from Basic Formal Ontology (BFO) to identify and avoid inconsistencies in GO.

Availability: The detailed results are available at
http://www.techfak.uni-bielefeld.de/~arueegg/go-evaluation/

Contact: Jacob Köhler, CPI, B60 (Centenary) 116D, Rothamsted Research, Harpenden, Hertfordshire, AL5 2JQ
jacob.koehler@bbsrc.ac.uk

Keywords: circularity index, intelligibility index, Gene Ontology, ontology alignment, formal ontology

1 Introduction

The Gene Ontology (Gene-Ontology-Consortium, 2001) has established itself as one of the most important computational resources for molecular biology and bioinformatics. GO has had a major impact on the annotation of genomes (Camon *et al.*, 2004) and is often used as a controlled vocabulary in database integration systems (Harris *et al.*, 2004). Recently more and more applications are exploiting the hierarchical data structure of ontologies like GO for such tasks as microarray analysis (Lee *et al.*, 2004; Zhang *et al.*, 2004), text mining (Nenadic *et al.*, 2002), database integration (Köhler, 2004), and measurement of the semantic similarity of ontological concepts (Van Buggenhout *et al.*, 2003). Such applications can take advantage of the data structure that evolves when ontologies are built following well established design principles as discussed in (Blázquez *et al.*, 1998; Ceusters, 2001; Ceusters *et al.*, 2003; Hovy, 2002; Noy *et al.*, 2001; Rosse *et al.*, 2003; Rosse *et al.*, 1998; Schulze-Kremer, 1997; Schulze-Kremer, 2002; Smith *et al.*, 2004a; Smith *et al.*, 2004c).

There are a number of controversial issues which affect the development of controlled vocabularies and ontologies, their formal notation, and how to implement them. For the purpose of this communication, we follow the account given in (Köhler *et al.*, 2003), without however claiming that this is the only way to define ontologies and controlled vocabularies. This means that we will consider a *controlled vocabulary* as a set of *nodes* each of which is associated with an *identifier, term, definition,* and an optional set of *synonyms.* In *ontologies* the nodes are linked by directed edges, thus forming a graph. This graph is then designed to represent a counterpart structure on the side of

entities (classes, universals) in reality. The edges of the ontology then represent the relations, e.g. *is-a* or *part-of*, which hold between these entities in reality. If a node has a parent node in the *is-a* hierarchy, then we say that it is *subsumed* by this parent node. A more elaborate definition of the data structure to which we refer when we speak of ontologies can be found in (Köhler *et al.*, 2003), and we note that GO itself is implemented in the way there described.

Several research programs (Ogren *et al.*, 2004; Wroe *et al.*, 2003) are using computational methods and ontology curation to overcome shortcomings in GO, and we have pointed to a variety of such shortcomings already in our papers (Kumar *et al.*, 2003; Smith *et al.*, 2004a; Smith *et al.*, 2004c), suggesting also ways in which it might be possible to overcome some of them by using computational methods.

According to (Smith *et al.*, 2004a), a well structured definition should not be *circular*; thus it should contain more information than the term itself. Further, a well-structured definition should be *intelligible*, i.e. the terms used in the definition should be simpler (more logically or ontologically basic) than the term to be defined.

In evaluating terms in the Gene Ontology, we thus introduced measures for these two principal parameters:

Circularity: Is the definition more than a reiteration or permutation of the GO term itself?

Intelligibility: Does the definition use sufficiently non-technical terminology to cover the meaning of the term?

In section 2, we develop methods for scoring circularity and intelligibility, and we set up a workflow suitable for drawing the attention of ontology curators to ill-defined terms (see Figure 1). This workflow also serves as a roadmap for the remainder of this communication.

In section 3, we then align GO terms to equivalent terms in other ontologies. In some cases it may be possible to replace problematic definitions in GO by borrowing definitions from other ontologies. However, the majority of problematic terms in GO are such that their shortcomings need to be removed by manual curation. To this end, we introduce in section 4 principles to be followed in improving definitions and in reshaping GO to overcome inconsistencies. These principles go beyond those introduced in (Blázquez *et al.*, 1998; Ceusters, 2001; Gruber, 1993; Hovy, 2002; Noy *et al.*, 2001; Schulze-Kremer, 1997; Schulze-Kremer, 2002; Smith *et al.*, 2004a) in that it relies upon a top-level ontology which has the potential to be used for automated consistency checking of the semantic content of the ontology.

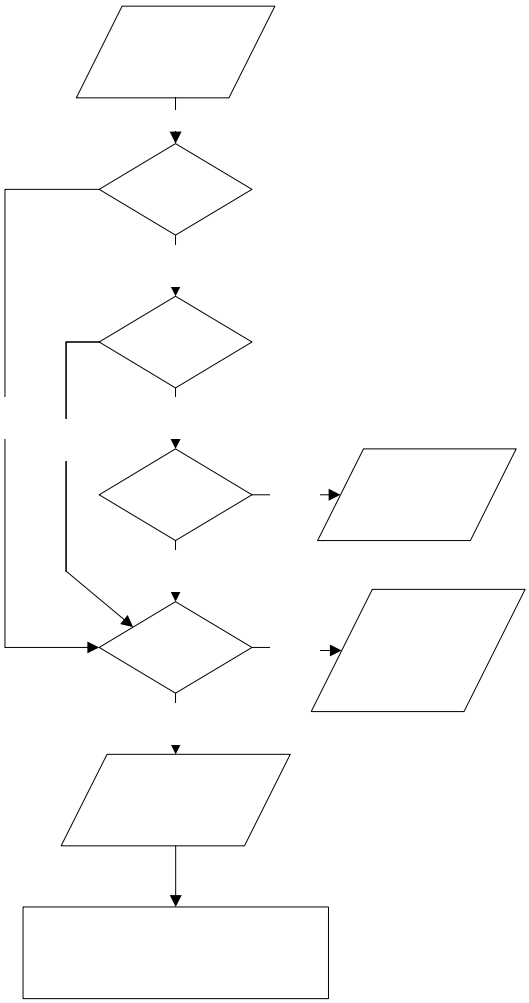

Figure 1: Workflow for the computational evaluation of the quality of terms and definitions. Definitions are considered to be circular if they have a circularity index $C \geq 0.5$ (see section 2.1) and as intelligible if they have an intelligibility index $I \leq 0.7$ (see section 2.2).

2 Evaluation of GO terms and definitions

In this section, methods for measuring circularity and intelligibility of GO definitions and terms are developed and applied.

All computations were carried out on GO's February 2004 release (revision 2.1707), within the text mining framework ONDEX (Köhler *et al.*, 2004), which is a system for automated ontology alignment and ontology based text indexing.

2.1 Circularity

According to the workflow outlined in Figure 1, the first step was to identify those terms that have no definition at all. Such terms were set to one side for the purposes of this analysis. Of the remainder,

we first identified GO terms with circular definitions. Consider for example:

> **id:** GO:0042270
> **term:** Protection from natural killer cell mediated cytolysis
> **definition:** The process of protecting a cell from cytolysis by natural killer cells.

As this example illustrates, the words used in a definition may differ syntactically in several respects from the words used in the term defined. Thus they may differ in flexion, form (singular versus plural), in capitalization; they may also contain stopwords such as "the", "of " "a", "from", "by", that contribute little to the definition. In our example, the only words in the definition that differ semantically from those in the term are "process" and "mediated". But even "process" is not informative, since the term in question is a term in GO's molecular process ontology, which means the fact that the entity in question is a process, is already reflected in GO's hierarchical structure.

We measured circularity by counting those words occurring in both the definition and the term, and related this number to the number of words in the definition. Thus we define the circularity index C as follows:

$$C := \frac{|\, s(def \setminus stop) \cap s((term \cup syns) \setminus stop)\,|}{|\, s(def \setminus stop)\,|}$$

Here:

> *s = the function that returns the set of all distinct lower case converted word stems from a set of words*
> *def = the set of all words used in the definition*
> *term = the set of all words used in the term*
> *syns = the set of all words used in the synonyms of the concept*
> *stop = the set of stopwords*

When a GO term has one or more synonyms, the circularity index compares the information contained in the synonyms to that contained in the term's definition. Thus the index compares the information contained in a term and all its synonyms to the information contained in the definition. For example:

> **id:** GO:0005105
> **term:** breathless binding
> **synononyms:** breathless ligand, FGFR1 binding, FGFR1 ligand, type 1 fibroblast growth factor receptor ligand
> **definition:** Interacting selectively with the type 1 fibroblast growth factor receptor (FGFR1).

This term has 4 synonyms. 8 out of 9 non-stopwords in the definition also occur in at least one of the synonyms. Although this definition is an improvement in terms of circularity, it still does little more than reiterate the information contained in the term and synonyms.

Now consider, as an example of a non-circular definition:

> **id:** GO:0050919
> **term:** negative chemotaxis
> **definition:** The directed movement of a motile cell or organism towards a lower concentration in a concentration gradient of a specific chemical.

In this case the circularity index is 0, reflecting the fact that the definition and the term contain no words in common. This was the case for 2117 GO concepts.

We stipulate that terms with a circularity index of $C \geq 0.5$ are defined circularly. There are 1028 GO terms that meet or exceed this threshold. The detailed results of the evaluation of circularity can be found at http://www.techfak.uni-bielefeld.de/~arueegg/go-evaluation/.

By raising the threshold (as according to the workflow diagram in Figure 1 above), we can isolate terms with higher degrees of circularity and correspondingly isolate fewer concepts to be manually checked.

On our $C \geq 0.5$ threshold, 6.01 % of all GO definitions are circular. Such definitions are informationally redundant, since they contain no more information than do the corresponding terms themselves. They perform no service either for human users of GO or for those using GO for purposes of automatic information retrieval. However, if a GO term (or one of its synonyms) is intelligible, then it might be argued that the term itself serves also as the definition. Although we think that, apart from a small number of primitive terms (such as 'process' or 'component'), every term should have a definition which meets basic standards of adequacy (Michael *et al.*, 2001), a term's intelligibility rating can be used to narrow down further the list of problematic cases (see Figure 1). To this end, the following section introduces an index that can be used to quantify the *intelligibility* of both definitions and terms in an ontology like GO.

2.2 Intelligibility

Consider:

> **id:** GO:0050566
> **term:** asparaginyl-tRNA synthase (glutamine-hydrolyzing) activity.

definition: Catalysis Cyc:6.3.5.6-RXN

We believe that to most GO users neither the definition nor the term is here self-explanatory; rather, both require in-depth background knowledge drawn from a specific biological sub-discipline. We question also whether such terms and definitions are in any sense intelligible to computers programmed for automatic information extraction.

To isolate such cases we counted how many of the words that occur in a given GO definition are defined as terms in WordNet (Fellbaum, 1998), which is a lexical reference system that has basically the same underlying data structure as GO but with a much broader coverage. Its domain covers most areas of the common language used by non-experts.

The underlying assumption is that the terminology defined in WordNet represents a vocabulary shared in common by most GO users. WordNet contains a number of commonly used technical words, including words drawn from biomedical terminology, but they are terms whose level of technicality does not exceed that which most biologists and researchers in biomedicine can be expected to have mastered. We thus define the intelligibility index of a definition in an ontology like GO as follows:

$$I_{def} := \frac{|\, s(def \setminus stop) \cap s(wn) \,|}{|\, s(def \setminus stop) \,|}$$

Here,

s = the function that returns the set of all distinct lower case word stems from a set of words
def = the set of all words used in the definition
term = the set of all words used in the term
stop = the set of stopwords

We can also determine the Intelligibility Index of a term, I_{term} , by replacing *def* with *term* as follows:

$$I_{term} := \frac{|\, s(term \setminus stop) \cap s(wn) \,|}{|\, s(term \setminus stop) \,|}$$

The intelligibility index can take values between 0 (low intelligibility) and 1 (high intelligibility). The results of the evaluation of the intelligibility index can be found at http://www.techfak.uni-bielefeld.de/~arueegg/go-evaluation/.

The majority of high-scoring concepts in GO describe biochemical reactions. The reason for this is that these concepts are defined in terms of the biochemical reaction which they catalyze, and thus they use the names of chemical compounds, very few of which are contained in WordNet. It could however be argued that such concepts actually are intelligible: although most biologists will not know the names/formulas of the compounds involved, it is obvious that a biochemical reaction will in such case be specified in a systematic way that is at least in principle intelligible to most biologists. The human- and computer-readable representation of concepts related to metabolism and their linkage to external resources such as other ontologies and databases are in fact active fields of research within the Gene Ontology Next Generation Project (Wroe *et al.*, 2003).

Accordingly, we leave aside for present purposes those concepts that are related to metabolism and hence contain many names of chemical substances. Our method then seems to provide an accurate reflection of the intelligibility of the remaining terms and definitions, though it works best as a supplement to the circularity index. This is because by using the intelligibility index alone we cannot do justice to the fact that a given text string may be unintelligible even though it uses only familiar words. Thus the intelligibility index can reliably point to definitions and terms that use complicated terminology, but it will miss those cases where common terminology is *used* in an awkward way. Examples of such concepts will be discussed in section 4.

We stipulate that those terms and definitions are to be flagged for additional manual curation which have an intelligibility index (I_{def} or I_{term}) ≤ 0.7.

3 Ontology alignment

We identified an initial set of 6005 terms in GO which are either in themselves unintelligible or whose definitions are either suboptimal or missing. The next step was to see if it was possible to replace suboptimal or missing definitions with definitions of terms from other ontologies or controlled vocabularies already mapped to GO.

We performed these mappings using the ONDEX framework, by aligning GO pairwise to ontologies and controlled vocabularies such as MeSH (Lipscomb, 2000), WordNet 2.0 (Fellbaum, 1998), and the Enzyme Nomenclature (NC-IUBMB, 1992). To do this we used methods which, according to preliminary informal evaluations, achieve a precision of > 0.95. In addition, we imported 3371 manual mappings between GO and the Enzyme Nomenclature (see http://www.geneontology.org/external2go/). We also imported the mappings between the Enzyme Nomenclature and MeSH which are included in MeSH itself. We found a total of 14495 mappings between terms of these 4 ontologies, out of which 5284 mappings link GO terms to MeSH, WordNet or the Enzyme Nomenclature. We found in the

other ontologies counterparts to the 2046 undefined GO terms only for a subset of less then 50 GO terms. Of the 6005 cases where definitions were found to be circular, missing, or to have a low intelligibility index either for the definition or for the associated term, only 2831 had an equivalent term in one of the other ontologies. Although an equivalent term was found for almost half of the terms, the associated definitions were no better in terms of circularity or intelligibility than the definitions already existing in GO. For this reason manual curation of the GO terms will in most cases still be required, since only on a case by case basis can it be decided whether a GO definition should be replaced or supplemented or completely rewritten. In the next sections we discuss principles for such manual curation.

4 Basic Formal Ontology (BFO)

In the previous two sections we described the evaluation of the intelligibility and the degree of circularity of GO terms and definitions, and identified a subset of 6005 potentially problematic cases. We also found that in most cases equivalent terms in other ontologies and controlled vocabularies do not receive a superior treatment in terms of definitions, and we thus concluded that these terms can be improved only through manual curation.

In a series of papers (Ceusters *et al.*, 2003; Kumar *et al.*, 2003; Smith *et al.*, 2004a; Smith *et al.*, 2004c) we have attempted to show that formal ontology design principles can be used to support ontology curators in improving problematic terms and definitions in an ontology like GO. However, some changes to GO's high-level hierarchy are a prerequisite to making such improvements. Thus in this section we will provide the outlines of the Basic Formal Ontology (BFO) (Grenon *et al.*, 2004) which provides our framework for manual curation, and set forth the principles it prescribes for creating ontologically sound definitions.

In section 4.2 and 4.3 we provide examples of how GO definitions can be improved by using BFO principles. Of course there may be other ways to bring about equivalent improvements, but we suggest that applying the principles of BFO to GO bears considerable initial promise, not least because the methodology has already been applied successfully in fields such as medical ontologies (Ceusters *et al.*, 2003) and spatial informatics (Grenon *et al.*, 2004).

BFO is a top-level domain-independent formal ontology designed to serve as a general theory which can form the starting-point for a series of lower-level ontologies specific to given domains. Top-level ontologies such as BFO are designed to be used as controls on the results achieved by working applications rather than as substitutes for these working applications themselves (Borgo *et al.*, 2002). Our success thus far in the manual application of such controls suggests that we should explore the degree to which they can be used for automated consistency checking and conflict resolution in ontologies.

BFO conforms to establish ontological practice and employs the term 'entity' as general term designating everything that exists (all items, objects, beings, existents). Entities exist on the side of reality, terms exist on the side of the ontologies we build for purposes of representation of reality. Where informaticians standardly talk somewhat ambiguously of concepts, we distinguish carefully between the terms of an ontology and the entities in reality to which these terms correspond. For the general terms such as are found in an ontology like GO, these entities are universals (classes, natural kinds) which are *instantiated* by particular components, functions or processes.

BFO starts out from the thesis that the terms employed in an ontology are able to represent reality as it is at some appropriate level of granularity. Thus the theories put forward in the BFO framework are theories *about reality*. In contrast to philosophical and scientific theories however they are designed at the same time to serve as a basis for application ontologies designed to support information systems of various sorts. Here BFO is used as a means of formulating more carefully the relations among the terms of an application ontology such as GO.

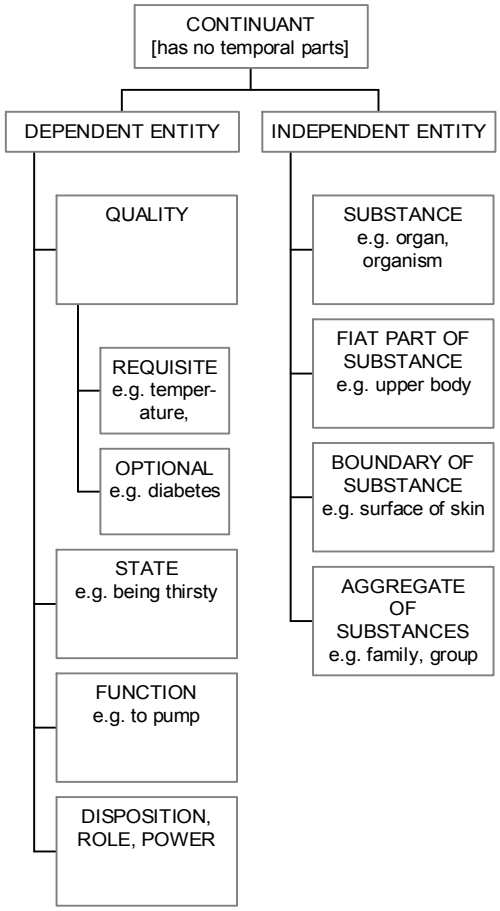

Figure 2: The ontology of continuant entities in BFO.

BFO first distinguishes *continuant* (Figure 2) and *occurrent* entities (Figure 3). Continuants are entities which *endure*, or *continue to exist*: they preserve their identity from one moment to the next even while undergoing changes. They are subject to a division between *independent* and *dependent* continuants. Independent continuants are physical objects, such as organs, cells, genes, and molecules, which do not require any other entities as their bearers or carriers in order to exist. Dependent continuants are entities such as shapes, qualities, functions, dispositions, states, and roles, all of which are distinguished by the fact that they depend for their existence on some independent continuant as bearer or carrier. A dependent continuant in the category of function – for example the function of a thermometer to measure temperatures – also exists self-identically from one moment to the next, and it exists even when it is not being exercised.

Occurrents, in contrast, are entities which *occur* in a given interval of time. Occurrent entities (processes, events, activities, changes) never exist in full in any single instant. Examples of occurrents are: the *exercise* of a function, the *execution* of a plan, the *application* of a therapy, the *realization* of a disposition.

At those levels of granularity relevant to bio-medicine, occurrents are always changes *of* or *in* some enduring entity or entities; thus they are dependent on continuants. The relationship continuants bear to occurrents is one of *participation:* occurrent entities depend for their existence on the participation of continuant entities. For example, the (continuant) *organism* as a whole participates in the process of *life*, so that if the organism did not exist the occurrent which is its life would not exist either.

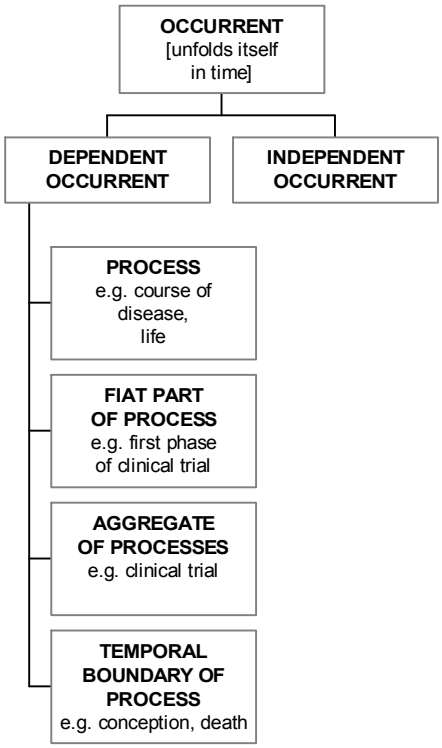

Figure 3 : The ontology of occurrent entities in BFO.

This tri-categorial system of independent continuants, dependent continuants, and occurrents provides the top-level architecture not only for BFO but also for the DOLCE ontology (Gangemi *et al.*, 2002) developed within the framework of the Semantic Web Initiative as the first module of the Wonderweb Foundational Ontologies Library (Masolo *et al.*, 2002). In addition, it underlies a number of other ontological systems currently in use, including LinKBase®, the large terminology-based medical ontology developed by the company L&C in Belgium (Ceusters, 2001; Verschelde *et al.*, 2004).

Drawing on this tri-part ontology we can now formulate already a series of principles which we believe should be respected by definitions and classifications in an ontology like GO:

P1. Such systems should employ a top-level hierarchy in which the above-mentioned three highest-level categories are clearly distinguished.

P2. These three categories should never overlap, i.e. all the direct and indirect parent nodes of each GO term should belong to the very same highest-level category as the term itself. Thus in particular a classification like GO should respect the factor of time: terms representing continuants should never subsume or be subsumed by terms representing occurrents.

P3. Terms designating entities in reality should never subsume or be subsumed by terms designating *knowledge* about reality, or features or outcomes of our processes of gaining such knowledge; thus *cardiac output* is not a *Laboratory or Test Result* or a *Diagnostic Procedure*, but rather something in the world. (Kumar *et al.*, 2003) *Molecular function unknown* is not a special kind of molecular function.

P4. Terms representing concrete entities (entities which exist in space and time and enter into causal relations) should never subsume or be subsumed by terms representing abstract entities (for example by units of measure, ideas, forms) or by terms representing terms or concepts.

Failure to abide by these and a variety of similar principles – including analogous principles applying to the relations between terms and their definitions – not only leads to characteristic coding errors; it also implies suboptimal reasoning capabilities: valid inferences will be blocked and invalid inferences will be admitted.

To superimpose the highest-level categories onto the GO structure in accordance with P2 requires relatively little initial work, since GO's three ontologies are already structured in such a way as to be disjoint. It requires moving GO's existing top-level terms below the corresponding highest-level concepts of BFO. One problem which needs to be solved however (a problem recognized also within the GO community itself) is how to classify GO's function terms: as continuants or as occurrents. GO does not define *function*, but it refers to functions in such a way that they designate activities, which are occurrents. For example, it defines molecular function (GO:0003674) as

> Elemental activities, such as catalysis or binding, describing the actions of a gene product at the molecular level. A given gene product may exhibit one or more molecular functions.

Thus failure to distinguish between functions and activities has the unfortunate consequence that a function that is not being exercised is not capable of being acknowledged within the GO framework. (Thus a heart that is not pumping cannot be said to have the function of pumping blood.)

More problems arise when we move down to lower level terms in the GO ontology. We are developing computational methods that check whether the principles (P1–P4) are also satisfied on these lower levels, for example by using tree-processing methods (Rozenshtein *et al.*, 1995), designed to check whether a given term is subsumed by two or more higher-level terms in a way which causes conflicts.

4.1 Applying Basic Formal Ontology to GO

We have used the formal ontological principles provided by BFO to classify some sample entities which GO terms describe. That GO leaves so many terms undefined is of course a major obstacle in performing this task: many of the undefined terms in GO form parts of terms that GO does define, or they form parts of the definitions of such terms. For example GO contains the term *response to blue light* but it does not contain the terms *response*, *blue*, or *light*. It defines *adult feeding behavior* (GO:0008343), but its definition: 'feeding behavior in a fully developed and mature organism,' is circular.

In order to classify the entity described by this term using BFO categories, it is necessary to search for clues about the way in which the term is used by GO; we thus looked up its component words in GO. The term *behavior* (GO:0007610) is defined by GO as follows:

> The specific actions or reactions of an organism in response to external or internal stimuli. Patterned activity of a whole organism in a manner dependent upon some combination of that organism's internal state and external conditions.

Unfortunately neither *adult* nor *feeding* receives a corresponding definition. The closest GO comes to providing a definition for *adult* is in the context of its definition of the term *adult behavior* (GO:0030534): 'behavior in a fully developed and mature organism' – a definition which is unfortunately circular. The closest GO comes to providing a definition of *feeding* is in the context of its definition of the term *feeding behavior* (GO:0007631): 'behavior associated with the intake of food' (also circular). It should be noted that GO is not clear as to how *feeding behavior* differs from *eating behavior* (GO:0042755), which it defines as:

> The specific actions or reactions of an organism relating to the intake of food, any substance (usually solid) that can be metabolized by an organism to give energy and build tissue

Note that this definition also contains an embedded definition of *food*.

Another major obstacle to applying BFO to GO is that many GO definitions, even where they are not circular, are too unspecific for the purposes of top-level domain-independent ontology. For example, it is unclear what relationship GO intends to describe in using the terms *relating to* or *associated with* in definitions such as those mentioned above: what does it mean for a behavior to be *related to* the intake of food? Definitions in GO leave unanswered many of those questions that need to be answered for purposes of constructing a reference ontology that meets basic standards of internal coherence and external validity.

It may be argued that a lack of precision in the definition of a term can be advantageous to a system of knowledge representation (not least in the area of biology), and thus that it is inadvisable to draw sharp distinctions where language itself allots a series of different but overlapping meanings to single terms. From the point of view of reference ontology, however, what is most important is that reality be represented *as it is.* If there truly *are* multiple aspects of reality which can be expressed by a single term in language, then it is the job of reference ontology precisely to distinguish what these aspects are.

Bearing in mind such issues we have applied the reference ontology BFO to some entities described by sample terms in GO.

4.2 BFO applied to *adult feeding behavior*

GO places *adult feeding behavior* in the category *biological process*. We must start our analysis by classifying *adult feeding behavior* under one of the three top-level BFO categories: as an occurrent. Relying on the GO definition alone, however, helps us determine neither the temporal extension of this occurrent, nor which specific continuants it depends upon.

There are many independent continuants which might participate in the occurrent *adult feeding behavior*, for example: the adult organism itself, food, and some environment (including some source of food). There are also many dependent continuants which, by virtue of their dependence upon the continuants that bear them, participate in *adult feeding behavior.* For example, those dependent continuants in virtue of which an independent continuant is edible, such as the *quality* of being organic; those which make feeding behavior possible, such as the *functions* of various organs; those which take part in causing the organism to engage in feeding behavior, such as the *disposition* of *hunger*; and whichever are the features that render the organism an *adult* (for example that it reaches a stage of fertility.

BFO also makes it possible to draw a tripartite distinction among the ways in which GO uses the term *behavior*: first, as this specific case of behavior here and now, as in: 'the behavior in which this organism is currently engaged'; second, as the universal or natural kind *behavior*, which is instantiated by any given instance of behavior but which does not depend for its existence on any one specific instantiation; third, as a generic or prototypical kind of instantiated behavior, such as 'the typical feeding behavior of a species of vertebrate organism.' Similarly we can distinguish for function terms such as carbohydrate metabolism (GO:005975) first the specific function of this mitochondrion to metabolize carbohydrates; second, the general function, to metabolize carbohydrates, which is instantiated thereby; third, a generic or prototypical function, to metabolize carbohydrates, which does not refer to any specific instance but rather idealizes therefrom. Distinctions such as this are currently rarely drawn in the discipline of bioinformatics.

4.3 Applying BFO to circular and unintelligible definitions

Using our circularity rankings of GO definitions we determined that the GO term whose definition received the highest possible score for circularity GO's is: *urogenital system development*, defined as: 'the development of the urogenital system'. In the following we illustrate how this definition could be improved once the BFO ontology is superimposed on GO.

First, we classify *development* as an occurrent entity. GO leaves *system* undefined, but BFO would rectify this by applying a definition it has already developed for *bodily system* (Smith *et al.*, 2004b), which is applicable to organic systems in general (for example to the circulatory system, the immune system, and so forth). An improved definition of *urogenital system development*, against this background, would incorporate the occurrent *development*, together with the continuant *urogenital system* upon which it depends.

term: urogential system development
definition: A biological process in which those elements whose function is to contribute to the processes of urination and reproduction change from an initial condition to a later condition in which the given elements are able to carry out their functions in a way

that contributes more effectively than before the change to the organism's survival.

This improved definition of course presupposes concepts such as *function*, *survival, element* and the relation of *contributing* which a function bears to survival; GO would have to define these terms also, which it can do by applying the definitions offered in (Smith *et al.*, 2004b).

Our intelligibility rankings tell us that GO's definition of *hepoxilin-epoxide hydrolase activity* (GO: 0047977):

> Catalysis of the reaction: (5Z,9E,14Z)-(8x,11R,12S)-11, 12-epoxy-8-hydroxyicosa-5,9,14-trienoate + H2O = (5Z,9E,14Z)-(8x,11x,12S)-8,11,12-trihydroxyicosa-5,9, 14-trienoate

has a very low degree of intelligibility ($I_{def} = 0.24$).

BFO would allow us to make this definition more intelligible first by categorizing both *catalysis* and *reaction* (neither of which GO defines) as occurrent entities. It would then categorize each of the components in the reaction as independent continuants of certain specific types, thus making it no longer necessary to know in advance the chemical name of a given continuant in order to have some grasp of what sort of entity it is. Finally BFO would delineate those dependent continuants which have these independent continuants as their bearers, including those roles, qualities, etc. in virtue of which the independent continuant contributes to the given reaction.

5 Discussion

Our indices of circularity and intelligibility can be applied also to other ontologies and controlled vocabularies. We have used these criteria to rank all GO definitions and terms and we showed by ontology alignment that only in some cases are equivalent concepts in other ontologies defined in a better way than they are in GO. This means that improving definitions will require a good deal of manual curation not only in GO but also elsewhere. However, methods such as those introduced in this publication, and text mining approaches such as those described in (Blaschke *et al.*, 2002; Chiang *et al.*, 2003; Ding, 2001; Gkoutos *et al.*, 2004; Sanderson *et al.*, 1999), can provide some support to ontology curators in building ontologies and improving definitions.

The previous section discusses several terms that are missing from GO. There are computational methods for identifying such missing concepts (Ogren *et al.*, 2004), based on calculating word frequencies in definitions and analyzing the compositional structure of GO terms (63.5% of all GO terms contain other GO terms as substrings).

GO definitions could be improved significantly by being re-written in the BFO framework. Although part of what is needed to improve GO is a term-by-term rewriting of many definitions, this will be of only nominal help unless all of GO is re-worked to incorporate the BFO framework. Further, introducing BFO's highest-level architecture to GO would allow inconsistent relations in GO's 'is-a' hierarchy to be detected automatically. It would also allow GO to be aligned more easily with other ontologies.

Thus given the computational methods introduced in the above for detecting trouble spots in GO definitions, and given the BFO reference ontology for re-constructing these definitions within a common framework, myriad inconsistencies and problems in GO can be resolved.

Acknowledgements

This paper was written under the auspices of the Wolfgang Paul Program of the Alexander von Humboldt Foundation and the project "Forms of Life" sponsored by the Volkswagen Foundation.

References

Blaschke, C. and Valencia, A. (2002) Automatic Ontology Construction from the Literature. *Genome Informatics*, **13**, 201–213.

Blázquez, M., Fernández, M., García-Pinar, J.M. and Gomez-Perez, A. (1998) Building Ontologies at the Knowledge Level using the Ontology Design Environment. In *Eleventh Workshop on Knowledge Acquisition, Modeling and Management*. Banff, Canada.

Borgo, S., Gangemi, A., Guarino, N., Masolo, C. and Oltramari, A. (2002) Ontology Road Map: Ontology infrastructure for the Semantic Web, pp. 16: National Research Council, Institute of Cognitive Sciences and Technology, Italy.

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R. and Apweiler, R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res*, **32 Database issue**, D262-6.

Ceusters, W. (2001) Formal Terminology Management for Language Based Knowledge Systems: Resistance is Futile. In *Trends in Special Language Technology* (eds Temmerman, R. and Lutjeharms, M.), pp. 135-153.

Ceusters, W., Smith, B., Kumar, A. and Dhaen, D. (2003) Mistakes in Medical Ontologies: Where Do They Come From and How Can They Be Detected? In *Workshop on Medical Ontologies* (ed. Pisanelli, D.M.). Rome, Italy: IOS Press, Amsterdam.

Chiang, J.H. and Yu, H.C. (2003) MeKE: discovering the functions of gene products from biomedical literature via sentence alignment. *Bioinformatics*, **19**, 1417-22.

Ding, Y. (2001) IR and AI: Using Co-Occurrence Theory to Generate Lightweight Ontologies. In *12th International Workshop on Database and Expert Systems Applications (DEXA)*, pp. 961-966. Munich, Germany.

Fellbaum, C. (1998) *WordNet : an electronic lexical database*. Language, speech, and communication, pp. xxii, 423. Cambridge, Mass: MIT Press.

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A. and Schneider, L. (2002) Sweetening Ontologies with DOLCE. In *EKAW*. Siguenza, Spain.

Gene-Ontology-Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res*, **11**, 1425-33.

Gkoutos, G.V., Green, E.C., Mallon, A.M., Hancock, J.M. and Davidson, D. (2004) Building mouse phenotype ontologies. *Pac Symp Biocomput*, 178-89.

Grenon, P. and Smith, B. (2004) SNAP and SPAN: Towards Dynamic Spatial Ontology. *Spatial Cognition and Computation*, **4**.

Gruber, T.R. (1993) Toward principles for the design of ontologies used for knowledge sharing. In *International Workshop on Formal Ontology*, vol. Formal Ontology in Conceptual Analysis and Knowledge Representation (ed. Poli, N.G.a.R.): Kluwer Academic.

Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G.M., Blake, J.A., Bult, C., Dolan, M., Drabkin, H., Eppig, J.T., Hill, D.P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J.M., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S., Fisk, D.G., Hirschman, J.E., Hong, E.L., Nash, R.S., Sethuraman, A., Theesfeld, C.L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S.Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E.M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T. and White, R. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*, **32 Database issue**, D258-61.

Hovy, E.H. (2002) Comparing Sets of Semantic Relations in Ontologies. In *The semantics of relationships : an interdisciplinary perspective* (eds Green, R., Bean, C.A. and Myaeng, S.H.), pp. cm. Boston: Kluwer Academic Publishers.

Köhler, J. (2004) Integration of Life Science Databases. *Drugs Discovery Today: BioSilico*, **2**, 61-69.

Köhler, J., Boeck, L., Everding, R., Schiller, R., Specht, M., Wolf, O. and Rüegg, A. (2004) ONDEX - Ontology based text indexing and querying. *under review*.

Köhler, J., Philippi, S. and Lange, M. (2003) SEMEDA: Ontology Based Semantic Integration of Biological Databases. *Bioinformatics*, **19**.

Kumar, A. and Smith, B. (2003) The Unified Medical Language System and the Gene Ontology: Some Critical Reflections. In *KI 2003 – 26th German Conference on Artificial Intelligence*, vol. 2821, pp. 135–148. Berlin, Germany: Springer.

Lee, S.G., Hur, J.U. and Kim, Y.S. (2004) A graph-theoretic modeling on GO space for biological interpretation of gene clusters. *Bioinformatics*, **20**, 381-8.

Lipscomb, C.E. (2000) Medical Subject Headings (MeSH). *Bull Med Libr Assoc*, **88**, 265-6.

Masolo, C., Gangemi, A., Guarino, N., Oltramari, A. and Schneider, L. (2002) WonderWeb Deliverable D17: The WonderWeb Library of Foundational Ontologies.

Michael, J., Mejino, J.L., Jr. and Rosse, C. (2001) The role of definitions in biomedical concept representation. *Proc AMIA Symp*, 463-7.

NC-IUBMB (1992) *Enzyme nomenclature 1992 : recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes*, pp. xiii, 862. San Diego: Published for the International Union of Biochemistry and Molecular Biology by Academic Press.

Nenadic, G., Mima, H., Spasic, I., Ananiadou, S. and Tsuj2638 0(anisand)-4.3( )7.1(Tsud)-4.l.3(i)7Dn6(1).3(ry5.7(Tc0.01(, C)6.10(99)-d)-4.erm-d)-4itbicv