Classification in molecular biology: realism, objectivity and the Gene Ontology. 1. Structure, principles and concepts

Charlie Mayor
Centre for Information Science
City University London
Northampton Square
London
EC1V 0HB
United Kingdom
Tel: +44 2070408390
Fax: +44 2070408584
Email: charliemayor@email.com

Lyn Robinson **(corresponding author)**
Centre for Information Science
City University London
Northampton Square
London
EC1V 0HB
United Kingdom
Tel: +44 2070408390
Fax: +44 2070408584
Email: lyn@soi.city.ac.uk

Abstract

The Gene Ontology (GO), a scientific vocabulary widely used in molecular biology databases, is examined by an analysis of its structure, a comparison of its principles to those of traditional controlled vocabularies, and by a detailed analysis of a single concept within it. It is found that the GO deviates in some respects from its principles of ontological realism, and that the two forms of vocabulary could benefit from adopting good practice from the other. In a companion paper, the ways in which the GO is used and maintained are examined by bibliometric analysis, content analysis and discourse analysis.

Introduction

Molecular biology has been recognized for many years as an information-intensive, indeed information-based science (Cole and Bawden 1996, Brown 2003, MacMullen and Denn 2005). One of the most important features of this is the use of biological ontologies, which have been explained as "formal representations of areas of knowledge in which the essential terms are combined with structuring rules that describe the relationship between the terms" (Bard and Rhree 2004). The Gene Ontology (hereafter GO) is such a formal vocabulary, which is used for the indexing of gene products from any species. It has been widely used in molecular biology research for more than a decade; the PubMed database had more than 5,000 references to the GO, and its applications, at the beginning of 2013, and the initial description of the GO (Ashburner et al. 2000) had been cited nearly 8000 times. Without the GO, it would be difficult, if not impossible, to make sense of the great volumes of data produced in molecular biology.

This paper, and a companion paper (Mayor and Robinson 2013), report an examination of the GO from a variety of perspectives, focusing on two aspects in particular. First, the GO is explicitly rooted in *ontological realism*, which is to say that all its terms should relate to an objectively real biological entity or process. This study seeks to assess how valid this is, and whether elements of subjectivity and conceptual entities may be present. Second, the GO does not explicitly acknowledge any principles of controlled vocabularies in the library/information science (LIS) sense, although many similarities are evident. This study seeks to assess the extent of such similarities, and whether an explicit adoption of LIS principles could be advantageous for the GO. Theses two aspects constitute an examination of the relations between formal, objective ontology, and more subjective, conceptual vocabulary control.

The study uses a mixed methods approach, drawing on the domain analysis of Hjørland (2002). In this paper, we consider the nature of the GO *per se*, by analyzing its structure, and comparing its principles to traditional controlled vocabularies, and then by a detailed analysis of a single concept within it. In a companion paper (Mayor and Robinson 2013), we analyse the ways in which the GO is used and maintained, through a combination of bibliometrics, content analysis and discourse analysis. The empirical work of the study was carried out between late 2010 and early 2012. Fuller details of all aspects are given in Mayor (2012).

We now go on to briefly describe the gene ontology itself, and then to analyse its structure, based on a variety of sources. These include:

- the bio-informatics literature, and specifically the literature of the GO, as found in PubMed and Web of Science
- historical articles and texts on biological concepts
- other biological classifications and ontologies, particularly MeSH and examples in Open Biological and Biomedical Ontologies collection [http://www.obofoundry.org].

Sources relating to the GO itself include:

- the main GO website
  [http://www.geneontology.org]
- the GO wiki
  [http://wiki.geneontology.org]
- archived versions of the GO website
  [http://wayback.archive.org/web/*/http://www.geneontology.org]
- GO data files (including archived versions)
  [http://cvsweb.geneontology.org/cgi-bin/cvsweb.cgi/go/]
- GO annotation files
  [http://www.geneontology.org/GO.downloads.annotations.shtml]
- GO development trackers
  [http://sourceforge.net/tracker/?group_id=36855].

Additionally, detailed semi-structured interviews were carried out with two senior GO editors, to give additional understanding of the context.

The Gene Ontology

The Gene Ontology Consortium is "a community-based bioinformatics resource that classifies gene product function through the use of structured, controlled vocabularies" (Gene Ontology Consortium 2013, D530). The GO was first defined in 1998, and launched in 2000 and consisted of over 36,000 terms at the end of 2012 (Gene Ontology Consortium 2013). The terms, in essence, describe what gene products (the substances, usually RNA transcript sequences of proteins, produced by the operation of a gene) do in biological contexts. Examples of GO terms are 'GO:0007155 cell adhesion' and 'GO:0048513 organ development'.

For early accounts of the GO see Ashburner et al. (2000) and Gene Ontology Consortium (2001, 2004): for later developments see Blake and Harris (2008), Barrell, Dimmer, Huntley, O'Donovan and Apsweiler (2009), Leonelli (2013) and Gene Ontology Consortium (2011, 2013); and for descriptions of the GO in the more general bioinformatics context see Robinson and Bauer (2011), du Plessis, Škunca, and Dessimoz (2011) and Baclawski and Niu (2006). An account of the way changes are made in the ontology, and some of the problems encountered, is given by Leonelli, Diehl, Christie, Harris and Lomax (2011).

The aim of the GO from its inception has been to solve a problem relating to language, to the complex and often apparently inconsistent way in which different biologists use the specialised language of their science in their work. It was created to solve a special problem in molecular biology: how to classify gene products, when data resources in different species databases were described and classified using different keyword sets and indexes, which prevented effective data integration, such as cross-databases searches for gene functions. An ontology formalism was chosen because it gave the possibility of automatic reasoning across the relationships in the vocabulary. In organisational terms, a consortium was created to manage the interests and contributions from different sub-domains in biology which had never collaborated on creating such a vocabulary before.

The Gene Ontology comprises three independent controlled vocabularies covering *cell components*, *biological processes* and *molecular functions*. These vocabularies are independent in the sense that no relations are stated between these three vocabularies. The general structure of the ontology is known in mathematical terms as a diacylic graph, which means that a parent term can have multiple children. It is hierarchical, but there can be many different paths from a term and through its parents to the root of the ontology. The simple example in Figure 1 displays the term 'cytoplasm' from the cell component ontology and indicates several branch paths through to the root of the ontology.

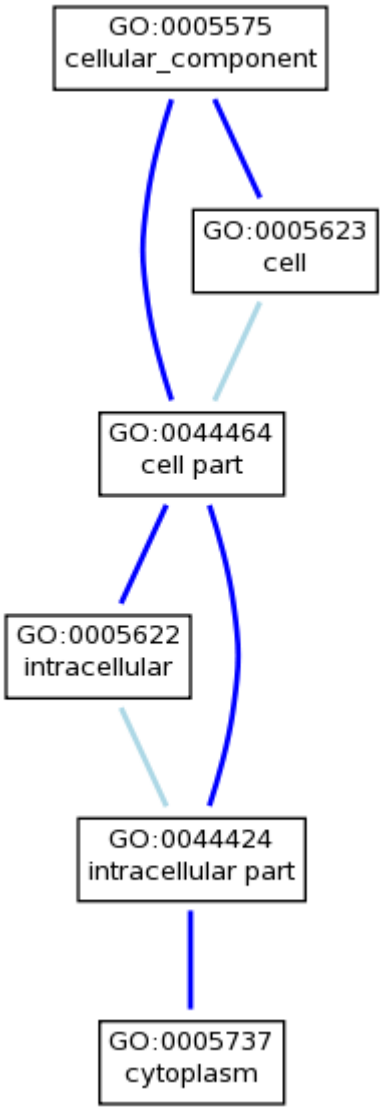

Figure 1 : GO term 'cytoplasm' and path to root of the Cell Component Ontology

The Gene Ontology often refers to each term as a node, the paths as arcs and the overall structure as a graph. Some examples of relations within the GO are shown below:

Relationship	**Example**
is_a	mitochondrion *is* intracellular organelle
part_of	mitochondrial membrane *part_of* mitochondrion
has_part	nucleus *has_part* chromosome
negatively_regulates	cell cycle checkpoint *negatively_regulates* cell cycle

Terms from the Gene Ontology are used to create *annotations* records for gene products in particular species databases (Hill, Smith, McAndrews-Hill and Blake 2008, Rhee, Wood, Dolinski and Draghici 2008). Although biologists refer to these associations between GO terms and gene products as annotations, from an LIS perspective it is recognizable as indexing or classification. Indeed, it has been shown that GO annotations follow a Zipfian distribution (Kalankesh, Stevens, and Brass 2012), as has long been known to be the case for conventional index terms (Nelson and Tague 1985). Annotations therefore serve a purpose of storing knowledge about gene products. This knowledge is based upon by evidence of different sorts, whether an inference made by an expert from a peer-reviewed journal article, or an automatic association created by a computer algorithm and based upon sequence similarities. Manual annotations created by subject specialists have been more highly valued by the biosciences community than have automatic annotations (Rogers and Ben-Hur 2009, Rhee, Wood, Dolinski and Draghici 2008), though the latter have been the subject of much research; see, for example, Seki and Mostafa (2008). At the end of 2012, there were over 350,000 manual annotations (Gene Ontology Consortium 2013).

Since inception, the Gene Ontology has developed in tandem with the evolving principles of Basic Formal Ontology. Biologists have been strongly motivated to follow the strictures of the BFO model because they aspire to create applications that will support automatic reasoning using ontologies (Smith et al. 2007). By combining an ontology like the Gene Ontology with high-quality annotation sets, it is hoped that biologists will be able discover new knowledge automatically from new or existing datasets.

However there may exist a tension between the aim of creating an objective ontology according to the precepts of realism to facilitate automated reasoning and data processing, and the aim of creating an ontology which reflects the best current knowledge of a diverse biosciences community.

The GO, despite having special features unique to ontologies, describes itself as a controlled vocabulary. A range of rules, procedures and standards have evolved over the course of the project determining the knowledge for inclusion in the ontology, and how it should be represented. The GO has never made any reference to standards for controlled vocabularies in a library/information sense and, since the project's inception, has always favoured the development of internal rules. These

rules relate to the philosophical principle of *ontological realism* acknowledged by GO developers as paramount to the creation of effective ontologies for biology.

Ontological realism

Ontologies are representations of domain knowledge, of concepts in a domain, captured in a structured language amenable to computer processing. LIS-style controlled vocabularies, such as thesauri and classifications, may be regarded as simple forms of ontology (Bawden and Robinson 2012, chapter 6).

Strict, formal ontology, which is the form commonly adopted in the biosciences, is realist in approach and denies that the concepts of a domain are mental objects, culturally relative and contingent to the thought processes of creative scientist. Rather concepts are objective and invariant. Reality has a structure and this structure can be captured within an ontology. The bio-ontologist aims to model these real-world classes as faithfully as possible, and resists any notion that concepts are plastic in their usage within languages and discourses, context-dependent, or personal and unique to the thinker (Smith 1995, Smith et al. 2005). This does not imply that the ontology model is perfect, and unchangeable; rather that it is a snapshot of the current consensus as to what the reality of the domain is (Leonelli 2012).

The philosophical principle of ontological realism has guided the development of the GO project, and indeed of most of the other ontology projects found in the biosciences domain. It has shaped the basic rules for all ontologies in the biosciences, and distinguishes the GO from the simpler thesauri, vocabularies or subject heading lists that might otherwise serve the purpose of indexing gene products. However, as Leonelli, Diehl, Christie, Harris and Lomax (2011, p6) note "its application has not always been straightforward"; we will return to the point later.

The Gene Ontology Consortium is one of many ontology projects collaborating together under the stewardship of the Open Biological and Biomedical Ontologies Foundry (OBO Foundry) [http://www.obofoundry.org], which seeks to overcome these barriers to data integration by creating standards for biomedical ontologies, and serving as 'hub' for ontology developers in biology. The OBO Foundry and its members adhere to a set of principles for ontology development (OBO Foundry 2011), grounded in the framework of Basic Formal Ontology (BFO) which offers an extended philosophical system for what an ontology is, and how an ontology ought to represent knowledge; see Leonelli (2012, 2013) for an analysis of the theoretical foundation, and alternatives to OBO for scientific ontologies. The OBO Foundry principles complement BFO to form a conceptual system for biomedical ontologies which any collaborator with the OBO Foundry, including the GO, are required to adhere to in order to participate; BFO is strongly realist in dogma (Smith 2006, Smith and Kumar 2004). Such ontological realism dismisses mental concepts and talk of concepts (which traditionally have been accepted as the object of knowledge representations) as anathema to the development of good, scientific ontologies.

Others argue that concepts in an ontology have their referents in the thinking minds of scientists rather than in reality; even some biologists resist the extreme notion that mental concepts have no place in ontologies for biology. For various perspectives on this hotly-debated issue, see Merrill (2010), Dumontier and Hoehndorf (2010), Lord and Stevens (2010) and Smith and Ceusters (2010).

Ontological realism is therefore the source for many of the structural features in the GO. The three sub-ontologies model distinct features of reality. Paths through the ontology must always be true, because reality does not err and every arc is a truth statement about what biologist understand to inhere in reality. Definitions for terms delimit structural features or activities in reality, and these correspond literally to protein structures encoded by genomic sequences. Terms in the GO never represent concepts in the mind of biologists, and nor do they ever stand as surrogates for linguistic functions in the language of biology. Term IDs and term names instead represent parts of a slice of reality, portioned off by the biologist in an effort to explain very specific life processes.

Structure of the gene ontology

A number of points in the structure of the GO should be emphasized, as they typify the distinctions between distinguish ontologies of this type and simpler vocabularies.

GO is a graph rather than a hierarchy

In presentation via web browsers like AmiGO, the Gene Ontology looks much like a traditional controlled vocabulary, with hierarchical, nested lists visualising paths from nodes back to the root of the ontology. However, any term have multiple locations in the ontology.

GO has created several different types of relations

The arcs between nodes in the Gene Ontology graphs represent one of several different types of relationships. The definitions for GO relationships can be quite complex, in part because they are created to satisfy the principles of formal ontology. However the power of ontology relationships is that they permit reasoning across the ontology. GO relationships act as logical restrictions on the connections between terms. Any relationship is valid if it describes a connection between two terms which is logically true.

GO is three separate ontologies

As noted above, the GO comprises three independent ontologies, for cell components, for biological processes and molecular functions. The assumption is that these are separate areas of knowledge. However, as the GO documentation acknowledges, it may be difficult to distinguish a biological process from a molecular functions, and cellular locations cannot be kept out of the other ontologies.

Cross products link to different ontology terms

Progress on linking the GO to other vocabularies has been relatively slow. Although work is in progress to provide formal cross products with other biological ontologies, by the end of 2012 none yet existed, not even between the three sub-ontologies of the GO.

On the other hand, there are numerous mappings between the GO and other vocabularies – for example, Enzyme Commission numbers UniProt Knowledgebase keywords, and the Unified Medical Language System (UMLS) (Lomax and McCray 2004, Barrell, Dimmer, Huntley, O'Donovan and Apsweiler 2009) – mapping GO keywords to their equivalents in the other vocabulary. Many of these, however, are between GO terms and keywords which would be inadmissible according to GO standards; if such mapping were considered synonyms, it would undermine many of the GO's formal principles.

Specific rules exist governing the contents of each of the three GO ontologies

There are more specific rules governing the structure of the three sub-ontologies, and the intended scope of these vocabularies. These are not entirely consistent across the whole GO. The Biological Process Ontology, for example, permits examples both of complete processes and collections of processes, where the others do not. The paucity of restrictions on the scope of Biological Process Ontology terms,

as compared to the other two sub-ontologies, is reflected in the number of terms in each vocabulary. The Biological Process Ontology consists of 23907 terms, whereas the Cellular Component Ontology and Molecular Function Ontology include 3050 and 9459 terms respectively (Gene Ontology Consortium 2013).

GO provides specific guidance on special topics

The Gene Ontology also provides guidance on special biological topics such as the cell cycle and metabolic process, so as to resolve contentious issues in the development of the ontology. For example, strictly speaking, GO only deals with single organisms, therefore terms describing processes occurring between a parasite and its host constitute an interaction between multiple organisms, and special rules govern how editors and annotators ought to approach these terms.

The structure of the GO shows the tension between the desire for a universally applicable and formally rigorous structure, and the need for a pragmatic solution to practical problems.

GO and LIS controlled vocabularies

Despite its commitment to ontological realism, the GO still identifies itself as a controlled vocabulary, representing biological knowledge in controlled terminology. It possesses features common to traditional controlled vocabularies, such as lists of terms with definitions and relationships. More importantly, the Gene Ontology is primarily used in the same way as an LIS controlled vocabulary, to index gene product entries in various species databases. It is therefore worthwhile to compare those standards used to construct the Gene Ontology against the existing, internationally recognised standards for vocabulary construction, ISO2788 and NISO Z39.19. Hereafter, we will refer to vocabularies constructed according to these standards as LIS vocabularies; see Bawden and Robinson (2012 chapter 6) and Broughton (2006) The criteria for such a comparison, referring to some core principles of controlled vocabulary, are:

- Ambiguity: How does GO ensure that terms have one and only one meaning?
- Synonymy: Does GO guarantee that each concept has one and only one preferred term?
- Using warrant to select terms: What warrant does GO use to select terms?
- Scope notes: How does GO explain the chosen meaning of terms?
- Compound terms: How does GO handle compound terms?
- Relationships: How does GO control relationships between terms in the vocabulary?

Ambiguity

To a large extent, the GO circumvents the problem of *homographs* – identical words having different meaning - by the limitation of its scope to the domain of molecular biology. Homographs in the ontology may have alternate meanings outside the domain, but these different meanings are not used to index gene products. The word ‘nucleus’ occurs throughout the GO, referring to the organelle in eukaryotic cells normally housing the chromosomes and replication machinery of the cell. There is no requirement in the ontology for this meaning to be differentiated from the term ‘nucleus’ as it is used in the physical sciences to refer to the positively charged central body in an atom.

Many GO terms are phrases containing homographs which are disambiguated implicitly by a standard phrase form, identifiable as a constituent in scientific language. GO:0060292, ‘long-term synaptic depression’ contains the homograph ‘depression’. No disambiguation is provided to distinguish the noun meaning ‘the act of lowering the activity’ from the noun referring to the psychiatric condition ‘depression’, because ‘synaptic depression’ is a phrase which is never decomposed into its constituents in the GO.

Polysemes differ from homographs in that the different meanings of a word or phrase are in some sense *related*. Instances abound in biology of words demonstrating polysemy such as ‘cell’, ‘drug’ or ‘bind’ which have multiple, related meanings in the English language. Although meanings are largely bounded by

convention in the molecular biology domain – biologists rarely apply the term 'binding' to mean the fastening within a cover, as in a book – efforts have been made by the GO to offer disambiguation for polysemes that could cause confusion for users.

For example, GO:0016020, 'membrane' is defined as follows: "Double layer of lipid molecules that encloses all cells, and, in eukaryotes, many organelles; may be a single or double lipid bilayer; also includes associated proteins". However the word 'membrane' has many more connotations in the biosciences domain. There are the mucosal membranes which line the cavities and canals of anatomical structures in the body, or the foetal membranes which surround and enclose the developing foetus. These types of membranes are not controlled for in the ontology since they fall out with the scope of the GO. However other types of membranes do need to be controlled for, and so 'basement membrane', 'plasma membrane' (and its exact synonym 'cell membrane') or 'photosynthetic membrane' terms relevant to plants are distinguished with unique qualifiers and definitions.

The problem for the GO is that in the Cell Component Ontology alone there are 404 GO terms which list the word 'membrane' in either the main term string or as some form of synonym for other GO terms. In one sense the Consortium has dealt with the problem of ambiguity here by providing a definition for every term. However the naive user, when presented with one of many hundreds of term containing the word 'membrane' is forced to refer to these definitions in order to understand the scope of each application.

The GO Consortium has also grappled with problems of ambiguity unique to biology. These special problems relating to the scope of specific terms is grounded in the ways different expert users understand words in the context of different species. In October 2007, almost 1000 terms and synonyms in the Gene Ontology files used what was described by the Consortium as a 'sensu qualifier'. Biologists working on different model species had failed to find agreement on common definitions for particular terms which, in different organisms, often carried subtly different meanings. For example, oogenesis, or the formations of egg cells in the females of species, has different meanings when mammals are compared to insects. These processes, although sharing the same name, were considered so different in their action that to use a single term 'oogenesis' to describe both processes would be false.

The solution devised by the Consortium was to introduce 'sensu qualifiers' whereby GO terms could be qualified by a species taxon to indicate to users the specific organisms or families a term ought to be applied to: 'sensu', meaning 'in the sense of', is commonly used in biological taxonomy. The GO term GO:0048477 'oogenesis' therefore had two related synonyms:

Oogenesis (sensu Mammalia)
Oogenesis (sense Insecta)

Some members of the GO Consortium regarded many GO terms as homographs, identical in wording yet different in meaning. For example, the biology of the

respiratory chains in eukaryotic cells versus bacterial cells was deemed to be so different that it would be misleading to annotate gene products with a single term; hence the terms 'respiratory chain complex I (sensu Eukaryota)' and 'respiratory chain complex I (sensu Bacteria)' were created to provide some means to distinguish between the two.

After proliferating sensu qualifiers through the ontologies, the strategy was eventually halted in 2007 and a project initiated to identify, merge and obsolete all terms which carried taxon information. Partly the reasons for this were ideological, because the aim of the Gene Ontology project had always been to provide a unified vocabulary to describe molecular biology and the gene products involved in molecular biology. Sensu qualifiers served to divide molecular processes by species, rather than accepting the assumption that there is an underlying unity to all molecular biology, irrespective of species. However, concessions to certain classes of species still persist in the ontologies. The Cell Component Ontology for example contains 81 terms referring to 'host cell' structures. The 'host cell' qualifier describes elements in the cell as they relate to the interaction between a parasite (or symbiotic organism) and the cell containing that parasite. So GO:0042025, 'host cell nucleus' describes the cell nucleus, but is only annotated to gene products originating from organisms such as parasites and targeted at the host cell nucleus.

We can see factors at play here that go well beyond the normal controlled vocabulary considerations of word meanings; also we can see the results of some contradiction in purpose, between the desires to describe biological reality, and to provide a pragmatically useful retrieval tool.

Synonymy

The Gene Ontology treats equivalent terms or synonyms as referring to the same class in reality. Therefore GO:0051169 carries the name tag 'nuclear transport' for a real-world class of process understood to be the directed substance transport "...into, out of, or within the nucleus". The name 'nuclear transport' is synonymous in the Gene Ontology with the name 'nucleus transport'. Whereas LIS controlled vocabularies would claim that these two names bear a semantic relationship and thus correspond to the same mental concept in the minds of users, the Gene Ontology approach is slightly different. In this example, usage of the names 'nuclear transport' and 'nucleus transport' is considered to refer to the same occurrents in reality, and these occurrents are members of the ontological class represented by GO:0051169. The OBO file format further expands on the ontology approach to synonyms, regarding a synonym as something used loosely for any kind of alternative label for a class. See Leonelli, Diehl, Christie, Harris and Lomax (2011) for discussion and examples of some issues with synonyms in the GO.

LIS vocabulary standards regard true synonyms as rare, with synonym use being highly dependent on context, such as between professional and lay terminology. The precept of ontological realism denies this contextual flexibility to synonyms since, as illustrated above, two different terms either refer to the same occurrent in reality, or

they do not. Therefore, the understanding of the nature of synonymy in the GO, and similar ontologies, is rather different from that in LIS controlled vocabularies.

Lexical variants are treated as exact synonyms in the GO, with American spellings generally used consistently, with British-English as alternatives.

Preferred terms are indicated in the GO by the first name given to any GO term ID. In a sense, the GO ID is primary to any term string used to refer to that ID. So GO:0009279 possesses the referent in reality which is the outer layer of any lipid bilayer, such as those found in bacteria, chloroplasts or mitochondria. The term string for this GO ID is 'cell outer membrane' but the GO selection of this string as the preferred form is usually at the behest of editors who seek to select a term which is self-explanatory or confuses as little as possible.

The Gene Ontology uses a system of synonyms to control for words or phrases with shared meanings. For example GO:0000272, 'polysaccharide catabolic process' is an exact synonym with 'polysaccharide catabolism'. This semantic relationship is further extrapolated through the Biological Process Ontology whereby all terms which are children of GO:0000272 share the same synonym structure. Therefore GO:0044239, 'salivary polysaccharide catabolic process' also has an exact synonym with the string 'salivary polysaccharide catabolism'.

Does GO ensure that each concept – or in GO philosophy, each instance in reality – is represented by a single preferred term? Largely yes, although flaws do exist in the GO approach to synonymy, and special problems are created by the GO's decision to maintain the three ontologies as separate entities. And there problems with new uses of terms, and with the same term used with different meanings in diverse fields of biology; Leonelli, Diehl, Christie, Harris and Lomax (2011) give examples of both.

There are also issues around the GO's status as an ongoing 'work-in progress'. This means that there are many ontology terms which, despite being implied, have yet to be formally added to the vocabulary. Some may be theoretical entities, like classes of enzymes for which evidence for their existence may be available, but has yet to be completely validated and accepted by the scientific community. Others are terms which might legitimately be added now, but which have not been added for practical reasons of time or added complexity to the ontology. For example, GO:0055006, 'cardiac cell development' is one of a number of 'cell development' terms. However development terms for every cell type in the body have yet to be added, partly because of the work involved, and partly because there has been a long-term aspiration that in the future, GO development terms may be cross-referenced to an existing cell-type ontology (Bard, Rhee and Ashburner 2005).

Warrant

Choice of, and justification for, terms in an LIS controlled vocabulary are generally based on of three, overlapping and related, types of warrant:

- literary warrant – the terms occurring in natural language texts in the domain of the vocabulary

- organizational warrant – the terminology used by formal groups and institutions associated with the domain
- user warrant – the terminology employed by, and understood by, the most likely users of the vocabulary

The Gene Ontology Consortium selectively ascribes *literary warrant* for term selection. At times it will follow canonical dictionary definitions for terms drawn from standard reference works. Yet, since the organisation considers the GO to be a solution for problems in cross-species indexing for gene products, editors will largely ignore literary warrant in an effort to 'tidy up' biological terminology. GO editors regularly devise new and non-standard terms to represent processes and functions for concepts which may be ambiguous, if literary warrant drawn from the domain were followed. Terms which have been selected according to literary warrant are readily identifiable in the GO according to ISBN references to textbooks and dictionaries, or PMID identifiers to source articles. A larger numbers of terms, however, will be attributed to groups or individuals within the Consortium.

Still less is *organizational warrant* a major factor. Since the project is a collaborative work between several different sub-domains in molecular biology, each of which is acutely aware of the differences in language between specialities, preferred terms are often the result of a negotiation and compromise. There is no single organization in the confederation of GO partners which has the authority to assert warrant over the vocabulary. Arguably the efforts of the GO to control the vagaries of biological language usage in the domain results in a distinct GO-style dialect for molecular biology which the Consortium imposes, with attendant benefits, on partner organizations.

While the Gene Ontology Consortium does place importance on contribution derived from the biosciences community, and will consider all term requests and ontology changes submitted by external users, *user warrant* is only applied in those situations where the suggestions from user groups meet the existing standards and rules for GO term inclusion. Wholesale acceptance of user warrant is strongly resisted, since the ontology is constructed according to strict philosophical rules and failure to adhere to these rules will often lead to rejection of new terms.

An example of resistance to user warrant which has had a considerable effect on the look and feel of the GO vocabulary is the rejection of terms which are themselves homographs for gene products. A norm in molecular biology is the use of gene product names as terms representing classes of gene product functions. For example, the protein 'actin' is a protein involved in the contraction of muscle, rearrangements of cell shape during cell division, and the movement of molecules to different locations within a cell. The word 'actin' is used in the molecular biology literature to refer to a class of gene products which share common functional properties, and authors will normally talks about actins and the functions of actins. Early in the development of GO, the Consortium members discussed at length the implications of including gene product names like 'actin' in the GO. The feeling was that these kinds of names did not represent functions, for one of the functions of

actin is to bind calcium ions but its function could not be described as 'being an actin'. A fourth ontology was proposed, to represent classes of gene products like actins, and subdivide these families into, for example, 'cytoplasmic actins' and 'muscle actins'. It was suggested that in terms of cross-database searches, it would prove invaluable to retrieve all the gene products from different species indexed as 'actins' which would be difficult for users trying to achieve the same result using function terms.

The 'Fourth Ontology' idea was never developed further, and gene product names became an ongoing problem in the ontologies as curators repeatedly obsoleted terms which looked like gene product names ('Heat Shock Proteins') yet did not articulate a clear function according to GO requirements. User warrant would justify the inclusion of gene product names to represent classes of gene functions. GO organizational warrant contradicted this class of terms, and hence a very large set of terms that may have been meaningful and useful to users was excluded from the GO.

This shows some clear differences between ontologies of the GO kind, and the LIS style of controlled vocabulary. Again, we see the effects of the desire to 'enforce' a model of biological reality meeting the needs of a pragmatically useful organization; and also, a degree of subjectivity in decision making, somewhat at odds with the objective basis of the ontology.

Scope notes

The GO does not use scope notes in the usual sense, but every term has a definition. Definitions serve to restrict or expand the application of terms. Several years after the GO project was initiated, large numbers of terms were added to the ontology without definitions, and therefore a retroactive project was implemented to extend definitions to all nodes in the graphs. Additional information on the scope of terms can be included in a 'Comments' field in ontology files. Comments are added at the behest of individual ontology editors and there is no policy regarding their addition.

One major omission in the GO, given that the project is updated on an almost daily basis and terms are added or removed as required by ontology editors, is any system for systematically tracking changes and deletions from the ontology files. Terms which are removed are clearly denoted as 'obsolete', but changes to term definitions, term strings, synonyms and the establishment of new relationships for active GO nodes are not tracked in the ontology files.

The implications of this approach and the paucity of information for the naive user to aid in understanding how the scope of specific nodes in the GO graphs has developed over time discussed in an analysis of term obsoletion (Mayor and Robinson 2013).

Compound terms

LIS vocabulary standards offer detailed recommendations on handling compound terms in controlled vocabularies, and advises on different conditions under which it may be advisable to split compound terms. As noted above, the GO, like all ontologies under the auspices of the OBO Foundry, does not regards its terms as concepts, and hence does not follow any such recommendations. However, a large majority of terms in the GO are structured like pre-coordinated compound terms, with repeated foci and modifiers combined to create lists of terms and hierarchies which all look very similar. This repetition is a consequence of ontological realism, since the ontology aims to describe entities and processes in reality which naturally fall into common types of classes.

An example is GO:0008152, 'metabolic process', a high-level node in the Biological Process Ontology. It has numerous child terms, and all are compounds of the form [entity] plus the term string 'metabolic process', such as GO:006807, 'nitrogen compound metabolic process' or GO:0042440, 'pigment metabolic process'. Child terms in this branch of the ontology are themselves compound terms following the same broad, repetitive structure, and include biosynthetic processes, catabolic processes and terms for regulatory effects, with considerable repetition in compound term wording is quite clear. There are, for example, over 1600 terms and term synonyms in the ontology which contain the wording 'biosynthetic process' and over 1400 terms with the wording 'metabolic process'.

A consideration in the design of any controlled vocabulary is when to split compound terms. The LIS standards suggest use of literary warrant in support of keeping compound terms, yet in the GO case of 'metabolic process' for example, commonly used legitimate alternatives may include 'chemical reaction', 'chemical pathway' or even 'biochemistry' as in 'melanin biochemistry'. It is difficult to recommend one form over another, and GO wording for compound terms is an essentially arbitrary decision, based on what is comprehensible to the user, and what distinguishes particular classes of terms within the GO itself

Relationships

LIS controlled vocabulary standards typically recommend three forms of semantic linking to indicate relationships between vocabulary terms: hierarchical, equivalence and association.

As will be clear from what has already been said, considerable importance has been placed by the GO developers on hierarchical relationships, with instances and different forms of parthood outlined in detail in the tenets of Basic Formal Ontology, and specifically with the types of hierarchical relationships exhibited in the GO itself. However, Leonelli, Diehl, Christie, Harris and Lomax (2011) draw attention to problems stemming from an initial lack of rigour in definition of hierarchy relations.

With respect to equivalence relationships, the Gene Ontology represents exact, broad, narrow and related synonyms within the vocabulary. Information regarding preferred terms and USE FOR statements, as discussed above, is poor and mostly implied in the structure of the ontology by the main term string associated with each

GO ID. Main term strings are a preferred term for a concept, are not necessarily selected according to any systematic criteria, are subject to revision without history notes, and possess synonyms of various kinds added at the GO developers' discretion or at user suggestion. There is, in short, no rigorous or systematic attempt by the Gene Ontology to provide exhaustive equivalency relationships.; the term 'informal notion' used at various points in GO documentation emphasizes this point.

No precisely-defined associative relation is recognized in the GO. 'Related synonyms' are defined negatively as synonyms which are not exact, narrow or broad; this further extend the imprecision of the GO's synonym handling.

In concluding this comparison with LIS vocabulary standards, we find that the GO is a sophisticated, pre-coordination system of terms for indexing gene products across the entire domain of molecular biology in a species-neutral manner. The fact that its developers have never officially or unofficially recognised any standards for controlled vocabulary construction in the course of its development is less important than whether it has successfully achieved its desired outcome, which is clarity. The GO project's internal rules, sometimes labyrinthine, do achieve many of the goals the LIS vocabulary standards aim to facilitate.

Homographs and polysemes create minor problems of ambiguity across the three GO sub-ontologies, although larger scale problems of semantic confusion are avoided because the GO covers a very specific domain.

Synonym control is very poor in the GO, with term entries failing to provide comprehensive ranges of synonyms. Despite the fact the GO Consortium is seeking to solve what is a semantic problem across the molecular biology domain, it does not regard as a central part of the solution the adequate control of word or phrase variations for concepts. Rather, ontological realism offers a means to demote linguistic variation and its limitation via synonym control in the vocabulary as a subordinate issue to the modelling of knowledge about entities in reality. The structure of the ontology emblemises this commitment: term identification strings are surrogates for processes and functions whereas name strings and synonyms are seen as labels or placeholders for GO term IDs. However, it may be argued that, in order to standardize the representation of gene products, the vagaries of biological language must be confronted. Otherwise the Gene Ontology risks further compounding the challenges of creating useful semantic tools to leverage knowledge in the domain.

The absence of defined warrant in choosing terms for inclusion and their preferred forms is a clear illustration of how the Gene Ontology is in fact manufacturing *non-standard* terms and definitions in the molecular biology domain. New compound terms are being created to represent nodes in the ontology which canonical literature in the domain has never used or addressed. This would not pose a major problem if the GO Consortium had received legitimacy from the biosciences community to create new compound terms and invent nodes to make the GO graphs more consistent. Yet the GO has been created by a small number of experts

representing the interests of the largest species databases, and the effort to expand the ontology and turn it into something useful has taken precedence over a normal standard in controlled vocabulary construction which is accepting a source for term warrant and identifying the authority from which concepts, term names and their relationships are derived.

Scope notes and history notes might serve to explicate to the naive user, or even the expert user presented with unfamiliar GO terms, as to how terms ought to be applied in annotating gene products. In the GO, these kinds of notes are brief, omitted, or copied from existing term notes, and are rarely informative. The absence of history notes for changes to GO terms also deletes a huge amount of contextual information about the source, and potential scope, for specific terms. The origin of GO terms can often be complex, as terms undergo repeated revisions yet none of this history is captured in the ontology files.

The GO uses compound terms extensively. Of particular interest is the way the ontology has created compound terms to manufacture logical divisions in the hierarchy. This has led to a source of confusion for users. The hierarchical levels in the ontology are not designed to have any significance, but in ontology applications, developers frequently interpret higher levels in the hierarchy to denote broader categories. Terms across different parts of the GO graphs which share hierarchical levels (such as three nodes down from the root) are not intended to share any common level of significance across otherwise unrelated biological systems. But by adding in compound terms to split arms of the ontology into logical ‘chunks’ the GO has lent the impression that hierarchical levels do matter.

This suggests that, were the GO to follow some of the LIS standards for vocabulary construction in concert with those structural features unique to ontologies, it would be clearer for annotators how to use terms for indexing gene products, and prove simpler for non-expert users to interpret the scope and application of nodes in the graph.

Concept analysis

To consider the nature of the GO in more detail, one of its concepts will now be analysed in detail.

There are different understandings of what a concept, in the context of a vocabulary or ontology, is, and systems for knowledge organization will differ accordingly; see, for example, Hjørland (2009) and Friedman and Smiraglia (2013). Concept analysis is here used to test the extent to which concepts in the GO are stable, context-free and value-free, as its declared basis of ontological realism would seem to require.

The GO term, 'cardiac cell differentiation' was chosen for analysis; the researcher had prior knowledge of cardiovascular biology, facilitating an understanding of the issues. The analysis procedure, guided by Hjørland's (2009) ideas was as follows:

- analyse the position of the term within the GO hierarchy, how it was created, and what justifications exist for these relationships
- identify equivalent or related terms in other biomedical vocabularies, and compare definitions; decompose the selected GO term into facets, and repeat the process for each facet
- search the biosciences literature for instances in which the chosen concept and its constituent facets are used; read and compare these meanings and contexts to the stated GO definition
- search through biology textbooks for canonical uses of the term, including index and glossary terms
- analyse the ways in which the term has been realised through the development of different theories in biology
- using this data, frame the term in the context of the theories of concepts established on empiricism, rationalism, historicism and pragmatism
- test whether term can be insulated from the criticism of context-dependency
- identify any terms, synonyms or definitions related to the term, but not excluded due to Gene Ontology design rules

Fuller details of the analysis, and in particular of the epistemological bases for concepts, are given in Mayor (2012).

Term, GO:0035051, or 'cardiac cell differentiation', was first added to the Biological Process Ontology in November 2003. Very few details are available regarding why it was added, as is quite normal in the GO. Its definition since addition is unchanged and reads:

> "The process whereby a relatively unspecialized cell acquires the specializedostructural and/or functional features of a cell that will form part of the cardiac organ of an individual."

The term is annotated to over 460 unique gene products across several different species databases, including gene product databases for mouse and human genomes.

The GO term 'cardiac cell differentiation' exists only in the Biological Process Ontology, and is related to two major parent processes in this ontology. Firstly, several 'is_a' relations mark it as a type of cell differentiation, and this focus on its role as a cellular process is bridged to the ontology root by the process term 'cellular developmental process'. Secondly, 'cardiac cell differentiation' maintains a 'part_of' relation to organ development processes and more specifically, the development of the heart. The path to the ontology root in this arm of the ontology passes via multicellular organism terms (since only multicellular organisms have organs) and also through 'anatomical structure development', as the heart is an important element in the functional anatomy of many organisms.

The GO term 'cardiac cell differentiation' is itself a parent term to a number of more specific children, examples of which are:

GO:0003348, 'cardiac endothelial cell differentiation'
GO: 0060935, 'cardiac fibroblast cell differentiation'
GO:0060950, 'cardiac glial cell differentiation'
GO:0055007, 'cardiac muscle cell differentiation'
GO:0060945, 'cardiac neuron differentiation'
GO:0003292, 'cardiac septum cell differentiation'
GO:0060947, 'cardiac vascular smooth muscle cell differentiation'
GO:0010002, 'cardioblast differentiation'
GO:00039293, 'heart valve cell differentiation'
GO:0007513, 'pericardial cell differentiation'

All the above are 'is_a' relations to GO:0035051 and although some carry annotations to known gene products, several are 'orphan term', never having been indexed to any gene products. Orphan children of 'cardiac cell differentiation' are 'cardiac fibroblast cell differentiation', 'cardiac glial cell differentiation', 'cardiac septum cell differentiation' and 'heart valve cell differentiation'.

The GO is an enumerative system of classification; however, one can identify different concepts or *facets* which comprise the classes found in the Gene Ontology. The GO developers have never explicitly stated that there are recurring facets which order the content GO yet it may be that the identification of facets in the GO classification has both value for the extension of the existing system, and for the development of potential alternative systems for ordering biological knowledge.

Applying Ranganathan's familiar PMEST formula (Bawden and Robinson 2012, chaoter 6), we may say that:

- personality is the character of a subject which distinguishes it from other subjects; GO:0035051 is marked as dealing with the facet of the biological process *differentiation*

- matter is the physical substance of which the subject is composed; GO:0035051 deals solely with the matter type ‘cells’
- energy is the action which occurs with respect to the subject; in the case of a biological process like differentiation, energy and personality are one and the same, the personality of the subject being an action carried out within a cell.
- space deals with the positional component of a subject; here, the heart or heart-like structure of multicellular organisms.
- time describes a temporal period associated with the subject; here the time facet for ‘cardiac cell differentiation’ could be the differentiation of cardiac cells in the embryo. Potentially, this time facet may occur at a later point in the life of an organism, for example if cardiac cells were to differentiate in the mature adult.

Vickery (1960) expanded Ranganathan’s original PMEST formula to include further facets appropriate to the description of entities in science and technology. These included Substance (product), Organ, Constituent, Structure, Shape, Property, Object of action (patient, raw material), Action, Operation, Process, Agent, Space, and Time. Vickery’s addition of new facets to the PMEST formula offers a means to solve problems such as the distinction between personality and energy in the case of ‘cardiac cell differentiation’. The product of this process is a cardiac cell, and the object of action would the less specialized precursor cell type. The organ in which the process occurs would be the heart or developing heart structure, and its structure, shape and properties can be defined according to existing standards in anatomical science.

As a single, illustrative example, it is clear that ‘cardiac cell differentiation’ is a compound term composed of a number of different facets, these facets being dependent on the type of faceted classification formula one may choose to use. One may apply the process of differentiation to different types of cells beyond cardiac cells, and indeed this is partially accomplished in the GO through sections of the graph dealing with differentiation of specialized cell types in the blood, nervous and respiratory systems.

Considering equivalent and related terms, the only such in the GO is the related term is ‘heart cell differentiation’. Results of detailed searches of biomedical vocabularies for terms related to the concept of cardiac cells and the concept of differentiation are given by Mayor (2012). We may summarise the results as showing that there is little evidence to support the idea that alternative controlled vocabularies in the biosciences domain include a general concept for a ‘cardiac cell’ as it is used in GO terms like ‘cardiac cell differentiation’. Much more common is reference to the specialized cell types such as cardiac muscle cells, cardiac nerve cells, cardiac cells found in embryo and cardiac cells in non-mammalian species.

Canonical textbooks in the domain can be an indicator of established knowledge and a source of literary warrant for controlled vocabularies. An examination of several

such suggests that again, much more complex and detailed terms are used; GO concepts represent an idealised form of the molecular biology of a cell.

What do these kinds of searches tell us about the GO term for 'cardiac cell differentiation'? This term possesses large numbers of arguably related terms in commonly used biological ontologies. A range of more specific terms describe different kinds of cardiac cells found in the heart organs of individuals, and the GO creation of a facet 'cardiac cell' to group together these different kinds of cells is not reproduced in comparable ontologies. Lexical synonyms for 'cardiac cell differentiation' are not dealt with by the GO, perhaps because the GO Consortium does not consider its product to be a solution for the problems of semantics and linguistic variation in the biosciences domain. However in even this simple search of alternative vocabularies, one can see the challenges of lexical variation and synonyms reproduced in these controlled vocabularies designed in part to circumvent these persistent problems.

The concept of cell differentiation is very common in biology, and vocabularies beyond the GO frequently describe cell differentiation and its regulation in terms of the different cell types in which it occurs. Concepts for the biological process of cell differentiation therefore extend to as many different cell types as one may care to consider, and may be further subtended by reference to these processes in specific anatomical regions of organs, such as the differentiation of cardiomyocytes in the ventricular or atrial regions of the heart. Since biological ontologies do not rely upon faceted approaches to term creation, every time a new cell differentiation term is needed to index a group of gene products, an entirely new compound term must be authored and defined.

A final point to consider is the GO Consortium's decision to create a node and children in the differentiation arc of the Biological Process graph devoted to cardiac cells and cardiac cell sub-types. This design choice is in conflict with the GO's own standards which aim to be species neutral. Not all species possess a heart and so arms of the ontology dedicated to kinds of biological processes occurring in specific organs have no relevance to particular taxa. The GO is not a universal classification. It incorporates classes of biological process which are irrelevant to large segments of the molecular biology domain. This replicates the very problem which the ontology aims to solve.

This analysis suggests that there is a convincing argument for accepting that every single GO term in the GO can potentially be considered as a *concept* rather than an abstraction of pure instances in reality. Even though a GO node may satisfy the condition of instantiability, it cannot exhibit context-independence and value-independence. The basis of the GO ontology requires its terms to be established on the basis of pragmatism; other approaches to concept formulation may be valid and useful (Mayor 2012). It is also worth noting that GO concepts, and indeed arguably the GO in its entirely, act as boundary objects (Star and Griesemer 1989, Star 2010, Huvila 2011): bridging disparate communities of practice, by facilitating

communication. It is not clear that this role is accepted, still less prioritized, in current GO processes.

The GO term ‘cardiac cell differentiation’ acts as a compromise, a shared idealisation of a complex network of ideas and theories about hearts, their cellular structure, and their functioning in living systems. Genetic, biochemical, physiological and pathological perspectives on this concept, as with all nodes in the Gene Ontology, are contrasting and potentially conflicting. In this sense, the GO’s commitment to ontological realism becomes something of a hindrance. The GO does not permit the incorporation of hypothetical concepts to describe gene products into the vocabulary, and this is a potentially serious drawback, if a purpose of the ontology is to facilitate new discoveries. This may be seen as a failure of imagination on the part of the Gene Ontology's designers; while it may serve its intended purpose, its scope is inevitably limited.

Conclusions

The GO has without doubt been highly successful in its own terms; this very success, as well as its structure, have led it to become large, and to a degree unwieldy. However, the analyses reported illustrate some, perhaps unexpected, features, which apply to all the scientific ontologies of which it is an exemplar.

Although the GO is explicitly based in a realistic and objective world-view, its terms exhibit the characteristics of concepts, failing to show complete independence of context and opinion. Elements of subjectivity are also present in the way the terms are determined. And, in as much as its terms do relate exclusively to objectively identified real-world entities, this limits its value as an aid to discovery and innovation. Indeed, it may be argued that the GO project risks marginalising or excluding important conceptual and linguistic forms in the domain, solely on the basis that they do not correspond to the designers' ontological commitments. The GO is logically consistent according to the precepts of formal ontology, but models a very limited representation of knowledge in molecular biology. This is further shown by the limitations of the mapping of the GO to other vocabularies, the inconsistencies between the three sub-ontologies, and the need for special rules.

The GO has never attempted to conform to the rules of traditional LIS controlled vocabularies, although its processes have gone a long way to emulating their more essential features. However, in numerous respects, including particularly synonym control, lack of identifiable warrant for choices, and absence of vital history information, it suffers by comparison with other biological vocabularies. Adoption of some of the principles of more traditional, concept-based vocabularies, would be beneficial for both annotators and users of the GO.

Conversely, of course, aspects of the GO and similar vocabularies, most particularly its amenability for automatic processing, could well be considered as improvements to LIS style controlled vocabularies. Studies of this sort may, we hope, lead to the improvement of both.

References

Ashburner, M. et al. (2000), Gene ontology: tool for the unification of biology, *Nature Genetics*, 25(1), 25-29

Baclawski, K. and Niu, T. (2006), *Ontologies for bioinformatics*, Cambridge MA: MIT Press

Bard, J.B.L. and Rhee, S.Y. (2004), Ontologies in biology: design, applications and future challenges, *Nature Reviews Genetics*, 5(3), 213-222

Bard, J, Rhee, S.Y. and Ashburner, M. (2005), An ontology for cell types, *Genome Biology*, 6(2), R21 [online] available at http://genomebiology.com/2005/6/2/r21, accessed 13 March 2013

Barrell, D, Dimmer, E., Huntley, R.P., O'Donovan, C. and Apsweiler, R. (2009), The GOA database in 2009 – an integrated Gene Ontology Annotation resource, *Nucleic Acids Research*, 37(D1), D396-D403 [online] available at http://nar.oxfordjournals.org/content/37/suppl_1/D396, accessed 13 March 2013

Bawden, D. and Robinson, L. (2012), *Introduction to Information Science*, London: Facet

Blake, J.A. and Harris, M.A. (2008), The Gene Ontology (GO) project: structured vocabularies for molecular biology and their application to genome and expression analysis, *Current Protocols in Bioinformatics*, 23: 7.2.1-7.2.9 [online] available at http://www.currentprotocols.com/WileyCDA/CPUnit/refId-bi0702.html accessed 14 March 2013

Broughton, V. (2006), *Essential thesaurus construction*, London: Facet

Brown, C. (2003), The changing face of scientific discourse: analysis of genomic and proteomic database usage and acceptance, *Journal of the American Society for Information Science and Technolog*y, 54(10), 926-938

Cole, N.J. and Bawden, D. (1996), Bioinformatics in the pharmaceutical industry, *Journal of Documentation*, 52(1), 51-68.

Dumontier, M. and Hoehndorf, R. (2010), Realism for scientific ontologies, in Galton, A. and Mizoguchi, R. (eds.), *Formal ontology in information systems*, Vol. 209 of *Frontiers in Artificial Intelligence and Applications*, Toronto: IOS Press, pp 387-399

du Plessis, L., Škunca, N. and Dessimoz, C. (2011), The what, where, how and why of gene ontology – a primer for bioinformaticians, *Briefings in Bioinformatics*, 12(6), 723-735

Friedman, A. and Smiraglia, R.P (2013), Nodes and arcs: concept map, semiotics and knowledge organization, *Journal of Documentation*, 69(1), 27-48

Gene Ontology Consortium (2013), Gene Ontology annotations and resources, *Nucleic Acids Research*, 41(D1), D530-D535 [online] available at http://nar.oxfordjournals.org/content/41/D1/D530, accessed 11 March 2013

Gene Ontology Consortium (2011), The Gene Ontology: enhancements for 2011, *Nucleic Acids Research*, 40(D1), D559-D564 [online] available at http://nar.oxfordjournals.org/content/40/D1/D559, accessed 12 March 2013

Gene Ontology Consortium (2004), The Gene Ontology (GO) database and informatics resource, *Nucleic Acids Research*, 32 (D1), D258-D261 [online] available at http://nar.oxfordjournals.org/content/32/suppl_1/D258, accessed 11 March 2013

Gene Ontology Consortium (2001), Creating the gene ontology resource: design and implementation, *Genome Research*, 11(8), 1425-1433

Hill, D.P., Smith, B., McAndrews-Hill, M.S. and Blake, J.A. (2008), Gene Ontology annotations: what they mean and where they come from, *BMC Bioinformatics*, 9(Supplement 5) S2 [online] available at http://www.biomedcentral.com/1471-2105/9/S5/S2

Hjørland, B. (2002), Domain analysis in information science. Eleven approaches – traditional as well as innovative, *Journal of Documentation*, 58(4), 422-464

Hjørland, B. (2009), Concept theory, *Journal of the American Society for Information Science and Technology*, 60(8), 1519-1536

Huvila, I. (2011), The politics of boundary objects: hegemonic interventions and the making of a document, *Journal of the American Society for Information Science and Technology*, 62(12), 2528-2539

Kalankesh, L.R., Stevens, R. and Brass, A. (2012), The language of gene ontology: a Zipf's law analysis, *BMC Bioinformatics*, 13:127 [online] available at http://www.biomedcentral.com/1471-2105/13/127 accessed 13 March 2013

Leonelli, S. (2012), Classificatory theory in data-intensive science: the case of Open Biomedical Ontologies, *International Studies in the Philosophy of Science*, 26(1), 47-65

Leonelli, S. (2013), Classificatory theory in data-intensive science: the case of Open Biomedical Ontologies, *International Studies in the Philosophy of Science*, 2691), 47-65

Leonelli, S., Diehl, A.D., Christie, K.R., Harris, M.A. and Lomax, J. (2011), How the gene ontology evolves, *BMC Bioinformatics*, 12: 325 [online] available at

http://www.biomedcentral.com/content/pdf/1471-2105-12-325.pdf accessed 13 March 2013

Lomax J. and McCray, A.T. (2004), Mapping the gene ontology into the Unified Medical Language System, *Comparative and Functional Genomics*, 5(4), 354-361

Lord, P. and Stevens, R. (2010), Adding a little reality to building ontologies for biology, *PLoS ONE*, 5(9), e12258 [online] available at http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0012258 (accessed 12 March 2013)

MacMullen, W.J. and Denn, S.O. (2005), Information problems in molecular biology and bioinformatics. *Journal of the American Society for Information Science and Technology*, 56(5), 447-456

Mayor, C. (2012), *The classification of gene products in the molecular biology domain: realism, objectivity, and the limitations of the Gene Ontology*, PhD thesis, City University London, available from the University repository at http://openaccess.city.ac.uk

Mayor, C. and Robinson, L. (2013), Classification in molecular biology: realism, objectivity and the Gene Ontology. 2. Development and application, *submitted for publication*

Merrill, G.H. (2010), Ontological realism: methodology or misdirection, *Applied Ontology*, 5(2), 79-108

Nelson, M.J. and Tague, J.M. (1985), Split size-rank models for the distribution of index terms, *Journal of the American Society for Information Science*, 36(5), 283-296

OBO Foundry (2011), OBO Foundry Principles [online], available at http://www.obofoundry.org/crit.shtml (accessed 12 March 2013)

Rhee, S.Y., Wood, V., Dolinski, K. and Draghici, S. (2008), use and misuse of the gene ontology annotations, *Nature Reviews Genetics*, 9(70, 509-515

Robinson, P.N. and Bauer, S. (2011), *Introduction to bio-ontologies*, Boca Raton FL: Chapman and Hall / CRC Press

Rogers, M. and Ben-Hur, A. (2009), The use of gene ontology evidence codes in preventing classifier assessment bias, *Bioinformatics*, 25(9), 1173-1177

Seki, K. and Mostafa, J. (2008), Gene Ontology Annotation as Text Categorization: An Empirical Study, Information Processing and Management 44(5), 1754-1770

Smith, B. (1995), Formal ontology, common sense and cognitive science, *International Journal of Human-Computer Studies*, 43(5-6), 641-667

Smith, B. (2006), From concepts to clinical reality: an essay on the benchmarking of biomedical terminologies, *Journal of Biomedical Informatics*, 39(3), 288-298

Smith, B. and Ceusters, W. (2010), Ontological realism: a methodology for coordinated evolution of scientific ontologies, *Applied Ontology*, 5(3), 139-188

Smith, B. and Kumar, A. (2004), Controlled vocabularies in bioinformatics: a case study in the gene ontology, *Drug Discovery Today: BIOSILICO*, 2(6), 246-252

Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A.L. and Rosse, C. (2005), Relations in biomedical ontologies, *Genome Biology*, 6(5):r46 [online] available at http://genomebiology.com/content/pdf/gb-2005-6-5-r46.pdf (accessed 12 March 2013)

Smith, B. et al. (2007), The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration, *Nature Biotechnology*, 25911), 1251-1255

Star, S.L. (2010), This is not a boundary object: reflections on the origins of a concept, *Science, Technology and Human Values*, 35(5), 601-617

Star, S.L. and Griesemer, J.R. (1989), Institutional ecology, 'translations' and boundary objects: amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39, *Social Studies of Science*, 19(3), 387-420

Vickery, B.C. (1960), *Faceted classification: a guide to construction and use of special schemes*, London: Aslib