City Research Online

City, University of London Institutional Repository

Citation: Mayor, C. & Robinson, L. (2014). Ontological realism, concepts and classification in molecular biology: Development and application of the gene ontology. Journal of Documentation, 70(1), pp. 173-193. doi: 10.1108/JD-06-2013-0076

This is the unspecified version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/3102/

Link to published version: https://doi.org/10.1108/JD-06-2013-0076

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

City Research Online: http://openaccess.city.ac.uk/ publications@city.ac.uk

Ontological realism, concepts and classification in molecular biology: development and application of the Gene Ontology

Abstract

Purpose
The development and use of the Gene Ontology (GO), a scientific vocabulary widely used in molecular biology databases, is evaluated, with particular reference to the relation between the theoretical basis of the GO, and the pragmatics of its application.

Methodology
The study uses a combination of bibliometric analysis, content analysis and discourse analysis. These analyses focus on details of the ways in which the terms of the ontology are amended and deleted, and in which they are applied by users.

Findings
Although the GO is explicitly based on an objective realist epistemology, a considerable extent of subjectivity and social factors are evident in its development and use. It is concluded that bio-ontologies could beneficially be extended to be pluralist, while remaining objective, taking a view of concepts closer to that of more traditional controlled vocabularies.

Originality/value
This is one of very few studies which evaluate the development of a formal ontology in relation to its conceptual foundations, and the first to consider the GO in this way.

Keywords
Molecular biology; ontologies; controlled vocabularies; bibliometrics; content analysis; discourse analysis

Introduction

This paper reports aspects of a mixed-methods study of the Gene Ontology (hereafter GO), a formal biological vocabulary, widely used for the indexing of gene product information (Leonelli 2012, Gene Ontology Consortium 2013A). For full details of the study background, and of the GO and the Consortium which created and maintains it, see Mayor and Robinson (2013) and Mayor (2012).

In brief, 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 2013A, D530). It is operated by a group of volunteer editors, who are themselves bioinformatics specialists, with input from the wider bioscience research area. This is a largely virtual community, communicating through email exchange, wikis and ontology development tools. 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 2013A). 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'.

In an earlier paper (Mayor and Robinson 2013), we reported an examination of 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. That analysis considered two aspects in particular: the objective stance of ontological realism on which the GO is explicitly based, the difference of principle and practice between scientific ontologies such as the GO (Bard and Rhee 2004, MacMullen and Denn 2005) and traditional controlled vocabularies in the library/information science (LIS) area. This paper focuses on the ways in which the GO is used and maintained, through a combination of bibliometrics, content analysis and discourse analysis. Although there are numerous publications and copious documentation on the GO, summarised in Mayor and Robinson (2013), little attention has been paid to the details of the ways in which the system is developed. The only other significant study of this kind is a detailed examination of five changes in the GO, by interviews with the curators responsible (Leonelli, Diehl, Christie, Harris and Lomax 2011).

Four aspects of the study are reported here. First, a bibliometric analysis of literature referencing using the GO is made, to show its usage at a macro-level. Then a content analysis of a selection of this literature explores how users in the molecular biology discipline themselves report usage of the GO, and whether this usage corresponds to the Consortium's own idea of how the ontology should be applied. Thirdly, discourse analysis is used for a detailed examination of the way in which the GO is developed. Finally the process by which terms are removed from the GO is examined, this shedding further light on development and use of the ontology.

Bibliometric analysis

A simple bibliometric approach was adopted to analyse a sample of published articles from the biosciences domain; for a similar bibliometric study of the area focused on use of open data rather than use of ontology, see Xia and Liu (2013). Since the idea of a gene ontology was first mooted in 1996, the sample included material published from 1996 to the beginning of 2010, the searches being conducted in late 2010. Bibliometric analysis cannot reveal the motivations or reasons behind why biologists use ontologies, and one must bear in mind that bibliometrics rests upon a large number of assumptions, such as the correspondence between citations and a paper's importance, or that the number of papers published on a topic is positively correlated with that topic's popularity in a domain.

Nevertheless the significance of the GO in biological research can be assessed by using bibliometrics to investigate factors such as: trends in publication rates; extent of, and trends in, citation of ontology papers; scatter on ontology papers across the literature; and the nature of the papers being published on the topic.

Searches were restricted to Ovid *Medline* and ISI *Web of Knowledge* since these provide good coverage of published articles in peer-reviewed journals, in which GO-based material is most likely to appear. The bibliometric dataset presented below therefore offers a quantitative snapshot of GO research in the biosciences. The search strategy for constructing the dataset as follows:

- Ovid Medline

(gene adj ontology).mp. [mp=title, original title, abstract, name of substance word, subject heading word, unique identifier] plus date ranging

- Web of Science

Topic (gene AND ontology)

- Combine, and remove duplicate titles

The first articles were published in 2000, indicating that the research using GO is likely to still be in its infancy. Figure 1 shows the steady annual growth of the GO literature.

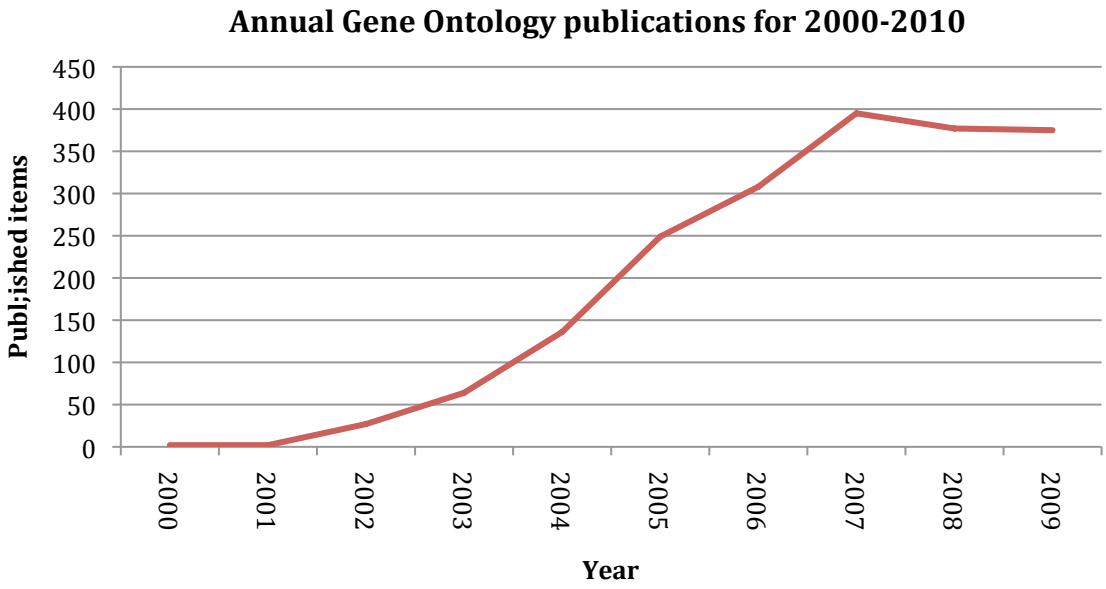

Figure 1

The G) project started in 1996 and in the early stages the nascent Consortium realised that it was essential to get published articles in top tier journals in order to promote the vocabulary and its benefits to the biosciences community, and to guarantee continued funding for the project in the future. Several papers were therefore published in 2000/2001, including the highly cited canonical paper by Ashburner *et al.* (2000), 'Gene Ontology: tool for the unification of biology', which were to be followed over the successive years by a growth in the publication rate of articles relating to the GO. Publications rates have stabilised in 2008/2009 and each year several hundred papers are published which use the GO.

Total citation rates indicate that papers published in 2005 account for the largest number of citations in the dataset. The strong peak for citations to papers published in the year 2000 is explained by the presence in that year of the Ashburner *et al*. paper, which the GO Consortium requires authors to cite in all instances where the GO is referenced. Citation rates show an annual decline in citations, as older papers have accumulated more citations; see Figure 2.

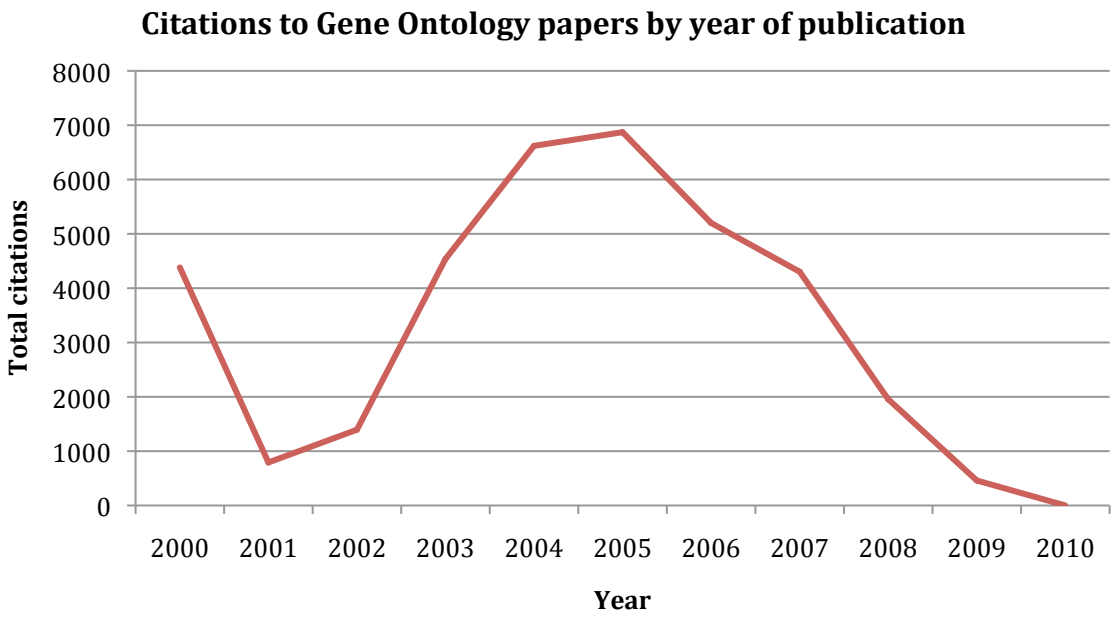

Figure 2

Average citations per year per article for years with ≥ 20 articles range from 70.9 citations per item for 2003 to 1.2 for 2009; see Figure 3. This pattern is typical, whereby older papers exhibit higher numbers of citations since they have been available to be cited for longer. No outliers were discovered, that is, younger publications which have been disproportionately popular as cited articles.

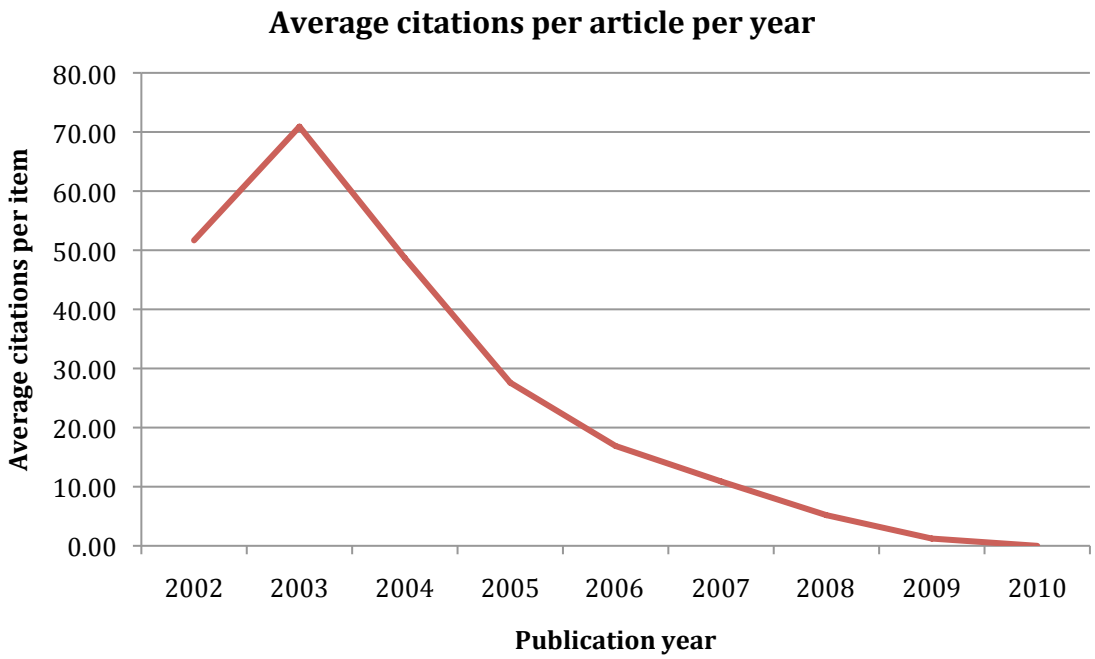

Figure 3

The geographical spread of authorship of GO research publications is shown in Figure 4. The preponderance of authors from the United States – 882 articles, with the next highest contributor being the UK with 158, might be expected, but the worldwide interest is clear, and matches the publication patterns of molecular biology and bioinformatics generally; see, for example, Song and Kim (2013).

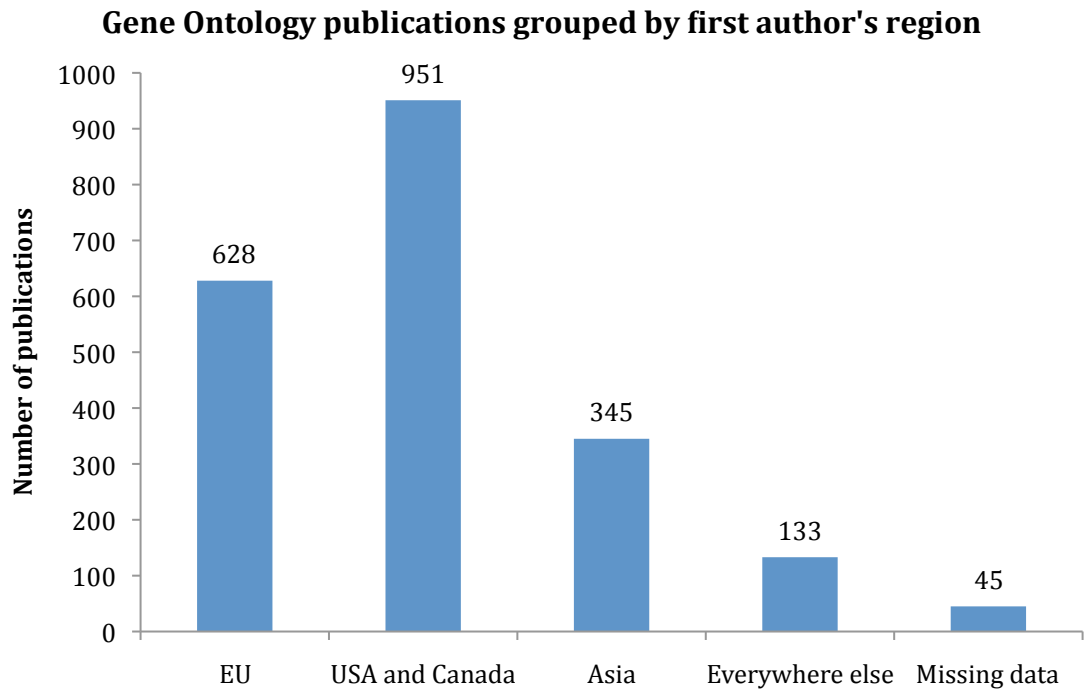
Gene Ontology publications grouped by first author's region
Number of publications
1000
900
800
700
600
500
400
300
200
100
0
628
951
345
133
45
EU
USA and Canada
Asia
Everywhere else
Missing data

Figure 4

Much GO research is published in one of four journals: *BMS Bioinformatics*, *Nucleic Acids Research*, *Bioinformatics* or *BMC Genomics*. These four journals account for nearly one third of all the Gene Ontology articles published before 2010. The top ranking journals are shown below:

Source	Number of articles
BMC Bioinformatics	178
Nucleic Acids Research	175
Bioinformatics	174
BMC Genomics	131
Physiological Genomics	42
Genome Biology	41
PLoS One	31
Pacific Symposium of Biocomputing	29
Genome Research	28
Journal of Proteome Research	26

GO research publication across all journals follows a distinctive Pareto distribution, with a few journals publishing the majority of work and a 'long-tail' accounting for single articles published in many different journals; see Figure 5.

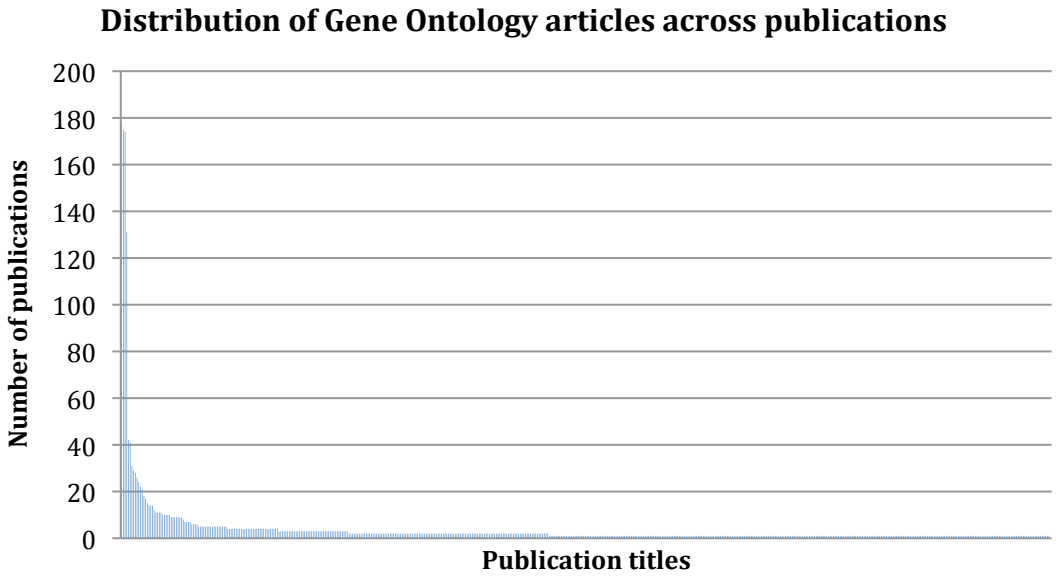

Figure 5

The bibliometric data above tells us very little about what authors are writing about when they reference the GO. In order to look in more detail at the subject of GO papers, the dataset was filtered for the MeSH heading *Oligonucleotide Array Sequence Analysis.* This relates to a method for the rapid and automated analysis of gene expression by use of short sequences of DNA (oligonucleotides) as probes. This created a subset of 720 papers which had been classified under the major MeSH heading dealing with microarray, and important gene expression analysis technique in biology. Each abstract for this subset was read and categorised according to whether authors were writing about a database, an analysis, a model, or some software.

A database paper was defined as principally reporting a community accessible database resource storing empirical biological data such as microarray data or sequence data, with this data being described using GO categories. An analysis paper was considered to use the GO to categorise and interpret empirical data, usually in what is known as a 'term enrichment analysis'. A model paper uses the GO to create computational models of biological systems. Software papers reported computer applications, normally available to the bioinformatics community, which incorporate GO files into the software (these are normally analytical tools for biologists to interpret datasets). The numbers of papers in these categories are shown below:

Category	**Count**
Analysis	522
Model	114
Software	66
Database	18

The preponderance of analysis papers is notable, and indicates that the GO literature is dominated by use of ontology in analysing scientific results, rather than self-referential ontology development or theory. While the use of a search term relating to an experimental technique may be expected to produce papers of this kind, the significant point is that they are there to be found.

These bibliometric results suggest that:

- publication rates on GO topics are numerous, stem from worldwide research, and are steadily increasing in number
- GO papers are being cited, indicating that the ontology matters
- GO papers are published in a wide range of bioscience journals, from various sub-domains, and primarily deal with analysis of experimental data, indicating that that it is a generally useful tool

This reinforces the importance of the GO in biological research, and also supports the idea that the GO is an appropriate lens for examining more general issues of classification and ontology in biology, so that that any conclusions drawn for the GO are likely to have general applicability.

Content analysis

This analysis was designed to examine publishing papers mentioning the GO, to ascertain how the GO is used and cited in practice. The set of literature used in the bibliometric analysis was the starting point. Only papers cited at least once, and hence of some significance, were considered; those published in 2009 were used, as being a set of tractable size, and a compromise between showing recent practice and being sufficiently old as to have had a chance of being cited. This gave a set of 163 papers. The abstracts of these were read in order to identify which papers definitely reported analysis of empirical data using Gene Ontology files, a technique normally dubbed in the domain as a 'GO analysis'. 113 papers published in 2009 were deemed to be GO analyses and the full texts of these papers form the core data for this content analysis. A smaller set of the most significant papers, defined as those cited 3 more times of which there were 32 in the set, was analysed separately.

The analysis included: reference to the GO ontology file versions used in the analysis, since the GO changes on an almost daily basis, and therefore authors are recommended to cite the ontology version they are using to ensure reproducibility of the results in the future; specific software tools used for GO analysis, as GO file version information is sometimes embedded in such tools; and whether analyses were restricted to a particular sub-ontology of the GO.

Finally, free text notes were made for papers where major deviations from recommended GO data citation policy were discovered. The data citation policy can be found on the GO Consortium website (Gene Ontology Consortium 2013B). In short authors are suggested to cite the canonical Gene Ontology paper by Ashburner *et al* (2000), to include date and/or version number for any ontology or annotation files, and are forbidden from editing the logical relationships or content of the ontology files.

Within the full set of 113 papers, a small core of journals was found to publish the greater proportion of GO-related publications, but beyond this core of traditional, bioinformatic titles, a very diverse complement of other journals, covering a broad range of sub-disciplines in molecular biology, were found to be publishing GO-related results. This would seem to suggest that the GO has succeeded in supporting a broad range of applications, especially in species or areas of interest not supported as a primary object of annotation with GO terms. These included papers investigating gene products in snails, mangroves, and tea plants. Also, GO analyses were found exploiting not just gene expression level, this being RNA transcript, but genome-wide association studies and protein-level investigations, implying that a systems biology approach, from the genomic to the transcriptomic to the proteomic, is facilitated by the GO.

The abstracts of this of set of papers were further examined, as the abstract, more than the full paper, is likely to encapsulate the authors' approach to use of the GO. 29 abstracts made only a general reference to the GO with no further information regarding precisely how ontology or annotation files were used to inform the results.

84 abstracts had a clear methodological statement of how the GO was used in the interpretation of empirical data. This suggests that authors do make specific indications of how GO analysis has contributed to conclusions in the research work. A large proportion of abstracts however did make only a passing reference to the GO, and one interpretation of this is that mentioning GO and ontologies serves the purpose of meeting a norm in the molecular biology domain. Ontologies have garnered significant attention in recent years, and well cited gene expression analysis papers will routinely incorporate GO term enrichment analysis into results. A mention of the GO in an abstract highlights to readers that a research work has deployed this popular technique in some respect, although the paucity of details in many abstracts does not lend clarity to how an ontology has validated conclusions mentioned in an abstract.

Of the restricted set of 32 papers, only 2 papers gave enough citation information to ascertain the ontology file versions used in the GO analysis. A total of 34 software packages and data analysis tools were mentioned with in-house solutions ranked as the most popular. Authors provided very little information of precisely how they used in-house statistical methods, which is essence suggest those terms which are commonly appear together, in such analyses. Several other tools mentioned, such as Ingenuity Pathway Analysis and Genespring are commercial platforms requiring a license for usage. These products provide poor transparency for determining how GO enrichment is performed, or which GO version is used by the software.

Authors may restrict GO analysis to one or more of the three GO sub-ontologies; these are three independent controlled vocabularies covering *cell components*, *biological processes* and *molecular functions* (Mayor and Robinson 2013). The set of 32 full text papers were read for whether authors had used all three sub-ontologies in their GO analysis, or if specific vocabularies had been chosen. Results are shown below, and reveal that a majority of papers published GO analyses based on data drawn from all three sub-ontologies, although a significant number of authors chose to use only the Biological Process sub-ontology:

Sub-ontology used	Count
Biological Process only	6
Cell Component only	1
Function only	0
Function and Biological Process	2
Biological Process and Cell Component	1
Function and Cell Component	0
All three	22

For the 32 papers analysed, the average number of GO terms cited in the full text is 6.4 with 6 articles failing to give any information about the specific names of enriched GO terms. No relationships were found between the number of citations received by a paper and the number of GO terms mentioned.

In addition, the number of GO terms cited in tables and supplementary figures were also analysed. A total of 9 out of 32 papers provided no information about enriched GO terms in tables or figures embedded in the article. 10 articles provided lists of enriched GO terms in the form of supplementary materials available from the publishers' websites. The general trend in reporting GO terms noted was that longer lists were provided in tables than in the full text, and potentially very long lists of GO terms were made available in the supplementary materials (with greater than 100 GO terms mentioned).

Finally, each full text article was read for major deviations from recommended Gene Ontology Consortium protocols on using and reporting usage of ontology files. Examples of such errors and failings in consistency are shown below; the PMID numbers appended identify the PubMed record. A wider range of examples is given by Mayor (2012).

makes reference to a non-existent GO term ‘cytochrome C release’ [19148281]

Gene Ontology terms are presented alongside terms from alternative controlled vocabularies yet no distinction is made between the different sources [18814146]

Author-derived categories for functions of genes presented as authentic Gene Ontology terms [19049829]

Bespoke categorisations created to group several hundred Biological Process Ontology terms under 14 author-generated subject headings [19615732]

Failure to distinguish terms derived from the three GO sub-ontologies, with Functions, Processes and Cell Components merged into a single figure [19372578]

Source and identity of GO terms selected for an analysis not given by authors [19037624]

This analysis suggests that errors may be largely due to authors choosing those enriched GO terms which best fit the theoretical context of their research argument, and marginalising those GO terms which are suggested by the kind of statistical analysis noted above, but which appear to bear less resemblance to the systems under study. This selective presentation occurs in the text of results sections, and in the way enriched GO categories can be laid out in tables and diagrams.

Authors may use non-GO sanctioned terminology to summarise the meaning of a group of GO terms, re-word the term string name to match the language in the rest of the article, and even mistakenly present bespoke categories for groups of enriched gene products as officially sanctioned GO terms.

Errors were also noted in the text sections for GO analyses, with typographical mistakes relating to term names and confusion between GO terms and terms from other controlled vocabularies. Some papers make reference to terms which are now obsolete, or to terms using the ‘sensu’ qualifier for species-specific GO terminology, a system which was deprecated by the time of publication of the article in question (Mayor and Robinson 2013). The ontology changes, but published articles are not updated to reflect that the fact that elements of the GO analysis may have been rendered irrelevant.

We conclude that content analysis of GO analysis full text papers paints a less than positive picture of GO usage and reporting in the domain. Very few papers provided enough information on the sources and versions of ontology files to ascertain precise dates for when files were downloaded. This makes it impossible for the rest of the user community to appraise the validity of GO analysis results.

Standardisation for GO analysis techniques and software tools is largely absent. Worse still, many authors report using in-house methods or commercial software for which there are no methodological details. It is not possible to determine how one proceeds from a large list of differentially expressed genes to a small number of apparently enriched GO terms. Transparency in bioinformatic methodologies is vital if readers are to trust results. As the interviews with GO experts (Mayor and Robinson 2013) indicate, researchers choose whichever tools are most appropriate for GO analysis, and publishers or peer-reviewers make very few demands on authors to publish this information. Issues of how to present GO analysis methods, or the time and effort to make GO analysis data available may explain this feature of GO papers.

Selectivity in which of the three GO sub-ontologies are used in an analysis is similarly unexplained in the literature. Authors do confuse concepts from the Molecular Function Ontology and the Biological Process Ontology, and when results sections tend to conflate terms drawn from these two, independent vocabularies. One conclusion is that the GO has never confronted the issue of what the philosophical bases for a function in the biology actually is, and consequently GO users fail to distinguish processes from functions.

This content analysis discovered no conventions in the literature for how to present the results of a GO analysis. Often, very long lists of enriched GO terms are laid out in tables embedded in the article, or are provided in spreadsheets or PDF documents in 'Supplementary Materials' sections on publishers' websites. Several databases and open data initiatives do exist in the biology domain for sharing this kind of information. As yet though, there is no evidence to support the notion that researchers are trying to share GO analysis results data in any form that might support re-use or re-combination along the lines of the e-science vision.
The number of enriched GO terms reported in papers and the statistical significance levels of these terms varied wildly in the full texts analysed here. Some papers cited only one or two enriched GO terms in support of theoretical conclusions, whereas other authors listed tens or even hundreds of potentially important enriched GO terms informing statements behind the function of the biological system or pathology under consideration.

The problem here is one of complexity and presenting this complexity in a form which is comprehensible and informative to the reader. A GO analysis might take the gene expression levels of many thousands of genes, and reduce the meaning of these differential levels down to a handful of functionally important GO terms. Data visualization techniques in the ontology domain, especially those for navigating complex data-sets, are improving, and in the future there may be scope for improving the presentation of GO term enrichment analysis results in the literature, beyond the current lists of terms noted in this study.

Despite clear guidance from the GO Consortium on citing ontology file versions in results from GO enrichment analyses, authors rarely comply with even basic data

citation policies. This is a serious problem; the ontology files are constantly updated, and therefore without file versions in reported results, it is impossible to track and reproduce those results.

Detailed reading of results sections from GO enrichment analysis revealed further features of GO usage. In particular, authors demonstrate high selectivity in the GO terms they report as significantly enriched in analyses, demonstrate little consistency in the presentation of functional classes found to be affected in experimental systems, and most worryingly were found to frequently re-word and re-categorise GO terminology in discussion sections. This practice runs entirely counter to the aims of the GO project, which is to provide a vocabulary that unifies language across disparate sub-domains in biology.

This problem of consistency is mirrored in the difficulties experienced by the GO Consortium in trying to ensure that different annotators choose the same GO terms for the same gene product. The analogous problem of inter-indexer consistency with LIS controlled vocabularies has long been recognized as a problem; see, for example, Bawden and Robinson (2012, chapter 6), de Keyser (2012), Hughes and Rafferty (2011), Soler Monreal and Gil-Leva (2011) and White, Willis and Greenberg (2013). de Keyser (2012, p. 47) calls it a 'platitude' that indexers cannot agree with each other. It is notable that it persists as an issue in the seemingly well-defined and objective molecular biology domain. It might be thought that the creators of the GO were somewhat naïve in believing that consistency was a realistic objective, and that perhaps input from library/information specialists at an early stage would have been helpful [we are grateful to a referee for this thought].

In summary, these content analysis results show poor compliance with GO data citation policy, in both full texts and abstracts, weak transparency in the way GO analysis results are generated and frequent re-wording and erroneous quotation of GO terminology. The entire GO system is designed as an authoritative, objective and logically consistent source for indexing gene products, yet usage and reporting of the GO system in the literature presents a somewhat contrary picture.

Discourse analysis of ontology development

The GO mailing lists are texts documenting a part of the design process in creating a scientific classification. Message threads capture exchanges between the developers and users involved in the early development of the Gene Ontology, and these exchanges determined the ontology contents at a particular point in time. The messages do not capture the whole design process; yet they provide evidence of how a scientific classification is created, and to what extent it is socially constructed, by studying the discussions between developers, curators and users as the ontology developed, using discourse analysis.

Discourse analysis is best regarded as a specific form of content analysis, focusing on the way in which spoken or written language is used, with the aim of analysing in detail the way in which ideas are treated; for reviews and examples of discourse analysis in LIS, see Budd (2006), Nahl (2007) and Haider and Bawden (2007).

This part of the study applied discourse analysis to messages posted on a GO mailing list, as part of the development of the ontology, in order to examine the ways in which the content of the GO is decided, and specifically how disputes are resolved. The value of discourse analysis with this type of material is well-established; see, for example, Fayard and de Sanctis (2010), Chiluwa (2009) and Park (2007). The specific form of discourse analysis used is the form of Critical Discourse Analysis (CDA) devised by Fairclough (1989, 2001). CDA treats discourse as a social act, subject to the roles, beliefs and authority of the participants. Fairclough's variant recognises three values that parts of discourse may have: *experiential*, to do with content, knowedge and belief; *relational*, to do with social relationships; and *expressive*, to do with the way the participants evaluate the reality the text refers to. Fairclough gives a simple framework for analysing texts that accommodates these three values. The elements in the framework are not absolutely required for analysing a text and nor is the analyst obliged to attend to each feature which equal weight.

In tandem with Fairclough's framework, an adapted 'linguistic checklist' suggested by Fowler (1985) is employed. This is designed to aid in studying those parts of language which are strongly implicated in the practice of power in discourse. It is by no means exhaustive; however it did serve to structure readings of texts extracted from the GO mailing lists, and enabled the analysis of those special linguistic motifs which discourse analysis can show to be associated with the exercise of power in social activities.

The corpus of text for analysis was the set of all postings to the GO mailing list from 1999-2002; roughly 5,000 individual messages. The period was chosen as the time when all shared electronic discussions about the GO were conducting using this single forum; thereafter, a wider range of forums have been used, not all readily available, precluding a comprehensive analysis. Although this analysis therefore covers the early years of GO development, there is no reason to believe that the nature of such communications have changed; a point validated by checks of later communications, and by interviews with GO developers (Mayor and Robinson 2013).

An initial macro-level overview of the nature of these communications was followed by detailed analysis of a series of discourses, using Fairclough's framework and Fowler's checklist; more details are given in Mayor (2012).

Macro-level discourse analysis results

Participants in these discussions were the first generation of GO editors, annotators and partners. The style can be described as 'professional informal' with speakers dealing with technical topics related to biology and ontologies, whilst at the same time using humour, informal greetings, metaphors and slang to communicate. Four principal attributes of sources featured in the macro-level reading of mailing list posts: *author* – who is speaking; *theme* – what are they speaking about; *sequence* – what is the order of speech; and *observation* – what was meant, and specifically how was authority used to legitimize certain aspects of the GO.

Author attribute: A number of main types of speakers, based on their affiliations with the Gene Ontology Consortium, can be identified as participants in the GO mailing list discussion. These speakers can be thought of as composing particular social groups within the biosciences domain, based on their degrees of authority within the domain, and their roles in relation to the Gene Ontology project. They included: *Senior Developers*, having rights as editors over the ontology, and maintaining top-level control over how the ontology changes; *Curators*, partners from other biological databases who may or may not have ontology-editing rights, but are responsible for implementing GO annotations within the context of their own system environment; *Commercial Partners*, offering financial or technical backing; *Expert Users*, reporting errors, making suggestions or asking for advice in specific GO usage; and *naïve users*, typically expert in biology but not in ontology, asking about GO support for their work.

Theme attribute: Several general themes were evident: *announcements*, typically statements without any following discussion; *administration*, rather more discursive threads on a wide variety of issues; *problems*, with reports of errors, and discussions of issues in using the GO; and *ontology requests*, the most recurrent theme, with suggestions for new terms, for changes to ontology relationships, for addition of synonyms and definitions for existing terms, and for the obsoletion of redundant terms.

Sequence attribute: This records the chronological sequence of speakers in each discrete message thread. Message threads were identified according to their titles, and messages grouped chronologically under each unique title. For some types of messages, there is a clear sequence in turn-taking, or who gets to speak next. Messages categorised as announcements rarely illicit a response, whereas Administrative threads often request replies from particular users, such as attendees at a meeting. Problem Communications normally follow a sequence of turns between an Expert or Naive User, and a Developer on the GO project. Of particular interest in terms of a critical discourse approach is studying the sequence of speakers on Ontology Requests, which generally followed no pattern, involved many more speakers than other themes of discussion on the GO mailing list, and revealed

more about power positions in terms of who could close conversations, or interrupt with new comments.

Observations attribute: Observations at the macro-level of reading the GO mailing list record interpretations, analysis and results of linguistic features, syntax, speech acts and implicature relevant to understanding the role played by power and authority in the resolution of disputes between speakers. For example, in reading the mailing list, it is clear that the GO Consortium has always been open to unsolicited ontology requests. The vocabulary is designed to represent the best current knowledge in molecular biology, and therefore an ontology request from any quarter, be it Ontology Editors or Naive Users, is given consideration for admittance to the ontology files. What is perhaps surprising is the small number of speakers involved in ontology request discourses. Usually Senior Editors or Curators decided the substance of changes amongst themselves, with very little contribution from outside the GO Consortium. The development of the GO has been driven by the work of experts funded by the GO Consortium and a very small number of highly active users directly involved in annotation. Since the scope of the GO covers a huge, complex domain, GO terminology has not always been authored by those with expertise in the relevant area of molecular biology.

Micro-level discourse analysis results

At the micro-level, detailed analyses were carried out, using the framework described above, of exchanges, on the problem and ontology request themes, between different types of speakers on the GO mailing list; space does not permit full description of these, for which see Mayor (2012). To give the flavour, we consider one discussion about proposed definitions of a term in some detail, and refer briefly to two others.

In the first example, a Senior Developer asks the list for contributions on a general definition for sexual reproduction. The definition is important, as it will affect the ontology content for all terms which are children of this node, and affect future annotations for any gene product related to reproduction.

What follows is an extended discussion between many speakers on the GO mailing list for several thousand words in which various contributors argue about their different conceptualisations of sex and sexual reproduction in different model organisms. There is little agreement between the discussion participants, and another Senior Developer eventually attempts to resolve the dispute.

In this case, senior developers reach consensus on a change to the ontology with relative ease; another example shows how conversations develop when there is not agreement between the different parties. In the course of trying to establish a common definition for reproduction, both sexual and otherwise, which would be acceptable to users working on a range of diverse species, GO mailing list participants offered a total of 23 different potential term definitions. This is a long and complex negotiation, and the full results of the discourse analysis are given in

Mayor (2012). However they key messages we can draw from these results are as follows:

- Speakers use personal pronouns when uncertain about a proposal or if suggesting a change which is likely to raise objections. As the discussion progresses and senior developers, conscious that the debate has hit an impasse, try to resolve the different points of view, more statements remove agency and present definitions as scientific, rather than subjective
- The definition becomes increasingly complex as speakers try to resolve different points of view on reproduction. Syntax becomes more complicated and elaborations in sentence structure render the arguments obtuse and resistant to criticism
- Opinions are divided into camps, based on model species organisms. This is indicated in the text by references to 'the worm people' and 'the plant people'
- Repeated modal adjustments either cast suggested definitions for reproduction as hesitant or, in the case of more senior developers, push suggested fixes as more definite or necessary
- Mailing list participants repeatedly present canonical definitions for reproduction from different sources, with implicature acting to show with each turn that previous suggestions were not acceptable. Direct confrontation and 'personalising' the discussion is thus largely avoided
- Eventually a comparably junior participant on the GO mailing list calls into question the entire procedure for determining new definitions and the validity of the debating process. This unusually provokes involvement from one of the chief architects of the GO project, who asserts control over the issue and abruptly determines a solution for the term definition

Modal adjustments, increasingly complex syntactical structures and the confident intervention of a senior developer show how the authoring of a definition for reproduction proved a considerable obstacle to the GO mailing list, and was eventually resolved only by the exercise of power by an authority figure in the GO Consortium. The definition for one term affects its relationship with every other node it connects to in the GO graph. In the case of reproduction, only an elaborate definition, one of the longest in the ontology, proved sufficiently scientific to meet the conflicting understandings presented for this concept during the course of the discussion.

In another discussion analysed, relating to terminology for the protein ubiquitin, it was found that, with minimal objection, four senior developers agreed to delete several major terms from the ontology, despite that fact that (a) they are recognised to act as a synonym for a recognised molecular function in biology and, (b) the terms refer to a major biological pathway, officially rewarded by the Nobel Prize Committee. This emphasizes the way in which formal ontology considerations predominate over any user warrant.

In a third example, a debate between a senior developer and a curator relating to a restructuring of an area of terminology, the fact that they felt it necessary to postpone any decision on the restructuring pending a face-to-face discussion

implies that not all terms are equal; some are considered more important than others. In this example the senior editor felt that an email discussion is an inferior communication method to sitting down for a meeting. The curator's language suggests deference to authority; uncertainty, irony, and the omission of agency with respect to who actually makes changes to the GO are indicative of a struggle to get changes made when confronted by a bureaucracy. The curator is appealing for email as a medium for resolving contentious issues because it enables all users to voice opinions, but the senior developer prefers a meeting for consensus among GO experts.

Discourse analysis conclusions

The results of the discourse analysis provide evidence that values, subjectivity and the exercise of power do in fact play a significant role in determining the content of the GO, notwithstanding its explicit commitment to objectivity.

Truth in the ontology is the result of a negotiation, constructed according to a consensus reached between a small group of developers and users. There is evidence that high-level definitions for terms in the ontology have been authored by senior editors in such a way as to close debate and marginalise alternate understandings of biological concepts. In this way, the GO has established a system of rules which creates ideological and institutional norms. These norms have helped the ontology to grow, and become integrated into different software applications for bioinformatics. On the one hand, the mailing list discourse presents individual, thinking biologists talking about their subjective, personal understandings of concepts in molecular biology. On the other hand, the discourse shows how the GO Consortium has necessarily exercised power to merge or exclude different points of view in order to sculpt a seemingly harmonious representation of knowledge in the domain. But with this comes some danger that the discourse primarily serves a small cadre of senior GO developers, rather than the wider molecular biology domain, which potentially understands any biological concept from all manner of plastic, imaginative ways. The GO developers have created a form of 'GO thinking', an ideology for knowledge in molecular biology which allows this knowledge to be structured as a classification.

The GO mailing list discourse therefore illustrates an interesting contradiction. Whereas ontologies in biology are presented as objective, value-free representations of reality, the process by which they are created is social. Subjectivity with respect to how concepts are understood is unavoidable, and power is necessarily exercised by the GO Consortium in order to resolve differences of opinion.

Obsoletion of terms

This part of the study aimed to categorise the reasons why several hundred GO terms in the Molecular Function Ontology were removed or 'obsoleted'. The aim of this approach is to understand why certain forms of knowledge are excluded from representation in the Gene Ontology, and to appraise the legitimacy of these exclusions.

The Gene Ontology Consortium has long had a system in place to remove terms from the ontology. In GO language this process is referred to as 'obsoletion' and the terms become known as 'obsolete terms'. Terms are never deleted from the ontology permanently. Future users may require historic versions of the ontology for data management purposes, such as repeating analyses conducted with previous versions of ontology files. Terms are therefore tagged as obsolete, and remain searchable in all subsequent iterations of the ontology files.

The principal reasons the GO Consortium cites for the obsoletion of ontology terms are: a term is outside the scope of GO and is no longer to be used to annotate gene products; or a term is redefined and changes meaning. GO curators propose terms for obsoletion through several different mechanisms, after which obsoletion occurs through a standard protocol, with a time period for objections, and then discussion between curators and editor.

The issue of term obsoletion, of what is removed from the GO, has been controversial from the GO's origins, and remains so. A particular problem is when terms familiar to molecular biologists are deleted because they do not meet the GO ontology standards. Term obsoletion is the process whereby the formal, logical precepts on ontology development and ontological realism confront the often messy ways biologists think and speak about theoretical problems. Ontological realism, the basis of the GO's structure, regards concepts are objective and invariant; reality has a structure and this structure can be captured within an ontology. The ontology model is not perfect, and unchangeable, rather it captures the current consensus as to what the reality of the domain is (see Mayor and Robinson 2013, also Leonelli 2012).

Ceusters (2009) proposes that term deletions are one way in which ontology quality can be audited. The ontological realism to which Ceusters and the GO adhere interprets obsoletion to mean a change in reality, a change in current scientific understanding, or a correction of a mistake in the ontology. This analysis addresses the stated reasons for GO obsoletions, allowing for the possibility that there be additional social and subjective reasons.

All previously obsoleted terms are included in newer versions of the GO files. Therefore, each of the three sub-ontologies (Cell Component, Molecular Functions and Biological Process) can be searched for previously obsoleted terms.
A total of 1448 obsolete terms was identified: 8% in the Cell Component Ontology, 36% in the Biological Process Ontology and 56% in the Molecular Function Ontology.

The analysis therefore focused on the Molecular Function Ontology, which has a higher rate of obsoletion than the other two, so as to examine the largest set of directly comparable terms, without concern for differences in practice between sub-ontologies. 403 terms were examined, half the total available, chosen arbitrarily by GO number.

The ontology files give brief, one sentence explanations for a term's obsoletion such as 'Gene product' or 'Cell component'. These notes are in themselves not especially revealing, and relate to simple obsoletion categories familiar to GO editors. However, extended discussions for each obsoletion are sometimes available by searching the GO documentation.

For each obsoleted term examined, the reason for the term removal was categorised, along with notes behind the obsoletion and the dates for proposal and implementation. Categories for term obsoletion were developed by an inductive method, creating new categories as new justifications for obsoletions were discovered, and re-categorising as major themes and explanations became apparent through the course of the investigation. Eight categories were identified, as follows, with the percentage of obsoletions ascribed to them:

- Cell component term **4%**
  The term refers to a structural component in a cell
- By structure, not ligand **3%**
  The term defines a class according to a common structural feature, not the ligand a product may bind
- Complexes **1.5%**
  Term refers to a complex of two or more protein subunits
- Compound of multiple terms **8%**
  Term is a combination of more than one concept
- Gene product **66%**
  Term is the name of a gene product, or class of gene products
- Non-existent function **6.5%**
  Term is not a function
- Process term **10%**
  Term is a process
- Protein family **1%**
  Term describes a class of proteins

The preponderance of obsoletion on the grounds that the term is a gene product, rather than a function, is clear.

A more detailed understanding was gained by analysis of the documentation; details and examples are given by Mayor (2010). In summary, three general mechanisms for processing an obsoletion are evident.

First, individual terms may be suggested as candidates for obsoletion by GO curators or by users. These requests may be discussed by GO editors, users are given the opportunity to respond, and the ontology is edited or remains unchanged based on these discussions.

Second, larger groups of terms may be proposed for obsoletion, normally by GO editors or annotators . These groups will be formed of related terms, such as functions involved in the response to viruses, and an editor may suggest large-scale changes to meet the current GO standards.

Third, sets of terms may be obsoleted based on a GO meeting or GO developer special interest group. One such example of this procedure was the obsoletion of a large group of peptidase-related GO terms in 2008. These terms were previously acceptable, but a change to the GO standard meant that GO developers now considered them to represent gene products rather than true functions. The terms were obsoleted *en masse* and the reasoning recorded via a special Wiki page on the GO website.

The obsoletion categories in the table above delineate several different types of discussions about standards for the Gene Ontology, and relate to particular problems encountered by the ontology's architects. For each obsoletion category, similar kinds of arguments can be discovered by reading the detailed discussions.

A recurring justification was that an existing term was either undefined, or had no annotations. Limited history notes (see Mayor and Robinson 2013) often make the reason why a term was initially added to the GO difficult to determine. As the ontology developed, undefined terms were flagged for potential removal and replaced with alternatives. The problem of undefined terms originated from the early development of the ontology, in which there was no requirement for term definitions. Current versions of GO now have complete coverage for terms with definitions.

Some term obsoletions can be described as matters of taste. In 2004, all GO terms containing the word 'chaperone' were obsoleted because it was felt they were suggestive of protein transport, rather than the traditional function associated with chaperone proteins which is polypeptide binding and folding. The chaperone terms were purged from the ontology files, only for users to raise objections in 2008 that these did in fact represent appropriate molecular functions and ought to be reinstated. The call was resisted, but synonyms containing the string 'chaperone' persist in the ontology and their status with respect to the original reasoning behind their obsoletion as preferred names for terms is unclear. With no firm warrant to follow in selecting term names (see Mayor and Robinson 2013), the GO Consortium has seemingly obsoleted terms like 'chaperone activity' on what are essentially aesthetic grounds, rather than looking to the molecular biology literature corpus or appealing to expert contributors for a validation.

Bowker and Star (1999) describe the work of deciding the content of a scientific classification system as a kind of organizational forgetting. This means that as a classification is created, developers decide what is going to be excluded from a system. Without a means to describe an object or activity with a category or class name, an information system cannot record that object or activity. Things without names are forgotten by these systems, and Bowker and Star outline two key processes contributing to this forgetting: *clearance* and *erasure*. Clearance excludes knowledge from classification systems; it is the process by which the authors of an information system deliberately exclude certain knowledge representations, such as outmoded approaches and alternate forms of domain problems. Erasure contributes to organizational forgetting by the deletion or omission of content from an information system. The failure to add this content to the system means that a part of an organization's work ceases to exist.

In the case of the GO, it seems that ontological realism acts as a paradigm facilitating clearance in the construction of this controlled vocabulary for describing gene products. The GO has the form of a directed acylic 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. Old, unfashionable or disproven theories are not represented in the GO: they are deleted in order to render all paths in the graph as objective, as true. Obsoleted terms have no annotations, and therefore information systems cannot retrieve gene product data related to these concepts. The GO is designed to be ahistorical. Errors, dead-ends and discredited beliefs in the domain are purged from the ontology files by clearance. Term obsoletions are the work of clearance and act as the remnants of this work. Their remains in the ontology are a testament to old ways of thinking about the functions of gene products, ways of thinking which are deemed unscientific by the GO Consortium.

Most terms are obsoleted from the Gene Ontology because they contain gene product names; that is to say they represent assumptions about the role of a gene, rather than representing a known process or function. This represents a contradiction in GO Consortium thinking. On the one hand, term names are treated as simple tags or placeholders for nodes in the GO graph. They are phrases commonly recognised by the user community as surrogates for the process or function in reality represented by a particular GO node identification number. Yet the molecular biology community has long used gene product names as related synonyms for the molecular functions carried out by those gene products.

If the GO Consortium really were agnostic on the semantics of terms names which, according to formal ontology, have absolutely no bearing on reasoning across the ontology graphs, then why are gene product names cited so often as a criterion for eliminating GO terms? This is clearance in action. The Gene Ontology is trying to tidy up a somewhat sloppy habit in the biosciences, of failing to articulate a clear function for gene products and instead relying on gene product names to provide enough contextual information to imply a molecular activity.

The casual way in which obsolete term names are sometimes used as related synonyms for replacement nodes in the graph further confuses the issue for users. If an obsolete term name is retained as a synonym for valid nodes in the ontology, then what were the grounds for its obsoletion in the first place?

Other controlled vocabularies, such as MeSH and Library of Congress Subject Headings, have always allowed for term deletion, but invariably provide alternatives (Bawden and Robinson 2012 chapter 6, Humphrey 1984, Knowlton 2005).

The GO's procedure is different: when a term is obsoleted, it is not obligatory to provide a new node in the GO graph as an alternative. Many terms, especially orphan terms (those which have never been used in annotations), are obsoleted in this way because, according to the GO approach, they do not represent instances in reality. But as previously noted, obsoleted terms can be candidates for re-introduction into the ontology, and many obsoleted terms persist in the ontology files not as secondary IDs for other nodes, but as related synonyms.

It seems that we must conclude that the nature of term obsoletion in the GO, justified by ontological realism and the exclusion of certain classes of concepts, makes it difficult for users to understand why terms have been deleted, and whether any alternative exists for indexing such concepts if needed.

Finally, in this section, it is worth mentioning the role of erasure in the Gene Ontology, as it is subtly different to this process of clearance described above. Erasure, in the manner described by Bowker and Star (1999), is the deletion or omission of content from an information system. In the context of the GO, it can be seen in the process of annotating gene products. In creating an annotation, i.e. choosing a GO term to associate with a protein or transcript entry in a database, all the rich and complex empirical data created in establishing an association is lost. The debates, contentions and theoretical context in which biological facts are discovered are deleted in the implementation of the GO. By this omission, a scientific classification like the GO drastically simplifies the scientific record, reducing it to a single association between a GO term and an entity in a database. With effort this scientific record can perhaps be re-created, but as an information system the GO model enables institutionalised forgetting via erasure.

Ontological realism has guided the GO developers to create a highly rarefied model of current knowledge in molecular biology. This model is ideally suited to aiding computer-based reasoning, as well as automatically creating new annotations to databases entities like putative proteins for which there is currently no functional information. However, in taking this approach, the GO necessarily ignores the rich and varied ways in which biologists understand theoretical problems and, more importantly, conceive new ideas. The GO's stance assumes that basic statements can be established as incontrovertible truths within the ontology and, based on a process of careful expansion, correction and elimination, an entirely true, objective model of knowledge in molecular biology can be constructed using the rules of ontology. This is contradicted by the occurrence within the GO of such issues as the

use of gene product names as ‘placeholders’ for complex concepts, the compounding of function and anatomical location, and a lack of clarity as to what is understood as a molecular function and what as a biological process.

The GO procedures for term obsoletion are intended to make the ontology ‘more true’. The idealistic intention is that, at some distant point in the future, reality will be perfectly mirrored in the networks of ontology nodes and relations. Yet as a vocabulary representing the knowledge of biologists, the ontology makes no effort to capture the idiosyncratic, the hypothetical, the varied ways in which working biologists understand concepts. The seemingly objective, empirical rules governing what ought to be excluded from the GO are tentative at best, and – as is seen from the discourse analysis and other aspects of this study – to a significant degree subjective, and socially constructed.

Conclusions

A preceding paper (Mayor and Robinson 2013) showed that, although the GO is developed on an explicitly objective basis of ontological realism, its structure and terms were in some respects similar to the more subjective structure and concepts found in LIS controlled vocabularies. This paper gives support to this idea, showing that subjectivity of viewpoint and social factors play a significant part in the way in which the ontology is developed and applied, so that the practice, particularly in terms consistency, does not always match the principles.

This should not be taken as a criticism of the GO, whose wide use testifies to its success. Rather it is to argue that its basis in ontological realism, and in turn in a realist and pragmatic epistemological understanding of concepts, may be too limiting. It fails to recognise the multiplicity of perspectives and viewpoints which are evident in a domain as formal and objective and molecular biology. In particular it severely limits the scope of the ontology to support innovation and conceptual novelty. That the GO works as well as it does is undoubtedly due to the expertise of its annotators and curators, using, as we have seen, a necessary degree of subjectivity and personal perspective. As Leonelli, Diehl, Christie, Harris and Lomax (2011, p.6) put it "expert judgement and manual intervention by curators appears key to the appropriate development of an ontology".

In envisaging how the GO, and similar scientific ontologies, might be developed to overcome such limitations, it is not at all necessary to abandon the objectivity at their root. Rather, ontologies might be extended so as to be pluralistic while remaining objective, accepting that there is no single way to describe the complexity of their domain; for the GO, this means no single way to classify all biological functions. This idea is developed further by Mayor (2012), and has similarities with the idea of an ontology as a kind of classificatory theory (Leonelli 2012).

Finally, we may note that any such development would lead to bio-ontologies treating concepts in a manner more akin to that in LIS controlled vocabularies. As suggested in the preceding paper (Mayor and Robinson 2013), it is likely both kinds of vocabulary have something to learn from the other.

References

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

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

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

Bowker, G.C. and Star, S.L. (1999), *Sorting things out: classification and its consequences*, Cambridge MA: MIT Press

Budd, J.M. (2006), Discourse analysis and the study of communication in LIS, *Library Trends*, 55(1), 65-82

Ceusters, W. (2009), Applying evolutionary terminology auditing to the Gene Ontology, *Journal of Biomedical Informatics*, 42(3), 518-529

Chiluwa, I (2009), The discourse of digital deceptions and '419' emails, *Discourse Studies*, 11(6), 635-660

de Keyser, P. (2012), *Indexing: from thesauri to the Semantic Web*, Oxford: Chandos

Fairclough, N. (1989), *Language and power*, London: Longman

Fairclough, N. (2001), Critical discourse analysis as a method in social science research, in Wodak, R. and Meyer, M. (eds.), *Methods of Critical Discourse Analysis*, London: Sage, pp 121-138

Fayard, A.L. and de Sanctis, G. (2010), Enacting language games: the development of a sense of 'we-ness' in online forums, *Information Systems Journal*, 20(4), 383-416

Fowler, R. (1985), Power, in van Dijk, T.A. (ed.), *Handbook of Discourse Analysis*, London: Academic Press

Gene Ontology Consortium (2013A), 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 (2013B), Gene Ontology Documentation [online] available at http://www.geneontology.org/GO.contents.doc.shtml, last accessed 23 July 2013

Haider, J. and Bawden, D. (2007), Conceptions of 'information poverty' in LIS: a discourse analysis, *Journal of Documentation*, 63(4), 534-557

Hughes, A.V. and Rafferty, P. (2011), Inter-indexer consistency in graphic materials indexing at the National Library of Wales, *Journal of Documentation*, 2011. 67(1), 9-32

Humphrey, S.M. (1984), File maintenance of MeSH headings in MEDLINE, *Journal of the American Society for Information Science*, 35(1), 34-44

Knowlton, S.A. (2005), Three decades since *Prejudices and Antipathies*: a study of changes in the Library of Congress Subject Headings, *Cataloging and Classification Quarterly*, 40(2), 123-145

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., 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

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), Ontological realism and classification: structures and concepts in the Gene Ontology, *Journal of the American Society for Information Science and Technology*, *in press*

Nahl, D. (2007), A discourse analysis technique for charting the flow of micro-information behaviour, *Journal of Documentation*, 63(3), 323-339

Park, J-R. (2007), Interpersonal and affective communication in synchronous online discourse, *Library Quarterly*, 77(2), 133-155

Soler Monreal, M.C. and Gil_Leva, I. (2011), Evaluation of controlled vocabularies by inter-indexer consistency, *information Research*, 16(4), paper 502 [online] available at http://InformationR.net/ir/16-4/paper502.html accessed 16 March 2013-03-16

Song, M. and Kim, S.Y. (2013), Detecting the knowledge structure of bioinformatics by mining full-text collections, *Scientometrics*, 96(1), 183-201

White, H., Willis, C. and Greenberg, J. (2013), HIVEing: the effect of a semantic web technology on inter-indexer consistency, *Journal of Documentation*, 70(3), in press

Xia, J. and Liu, Y. (2013), Usage patterns of open genomic data, *College and Research Libraries*, 74(2), 195-208