# The classification of gene products in the molecular biology domain: realism, objectivity, and the limitations of the Gene Ontology


Charlie Mayor

PhD

City University London

Department of Information Science

June 2012

## Contents

# Figures

# Tables

## Acknowledgements

I thank my supervisors, Lyn Robinson and David Bawden.

And thanks to Steinunn, Halldór and Einar.

---

## Declaration

# Abstract

**Background:** Controlled vocabularies in the molecular biology domain exist to facilitate data integration across database resources. One such tool is the Gene Ontology (GO), a classification designed to act as a universal index for gene products from any species. The Gene Ontology is used extensively in annotating gene products and analysing gene expression data, yet very little research exists from a library and information science perspective exploring the design principles, philosophy and social role of ontologies in biology.

**Aim:** To explore how molecular biologists, in creating the Gene Ontology, devised guidelines and rules for determining which scientific concepts are included in the ontology, and the criteria for how these concepts are represented

**Methods:** A domain analysis approach was used to devise a mixed methodology to study the design of the Gene Ontology. Concept analysis of a GO term and a critical discourse analysis of GO developer mailing list texts were used to test whether ontological realism is a tenable basis for constructing objective ontologies. A comparison of the current GO vocabulary construction guidelines and a study of the reasons why GO terms are removed from the ontology further explored the justifications for the design of the Gene Ontology. Finally, a content analysis of published GO papers examined how authors use and cite GO data and terminology.

**Results:** Gene Ontology terms can be presented according to different epistemologies for concepts, indicating that ontological realism is not the only way objective ontologies can be designed. Social roles and the exercise of power were found to play an important role in determining ontology content, and poor synonym control, a lack of clear warrant for deciding terminology and arbitrary decisions to delete and invent new terms undermine the objectivity and universal applicability of the Gene Ontology. Authors exhibited poor compliance with GO data citation policies, and in re-wording and misquoting GO terminology, risk exacerbating the semantic problems this controlled vocabulary was designed to solve.

**Conclusions:** The failure of the Gene Ontology to define what is meant by a molecular function, the exercise of power by GO developers in clearing contentious concepts from the ontology, and the strict adherence to ontological realism, which marginalises social and subjective ways of classifying scientific concepts, limits the utility of the ontology as a tool to unify the molecular biology domain. These limitations to the Gene Ontology design could be overcome with the development of lighter, pluralistic, user-controlled ‘open ontologies’ for gene products that can work alongside more traditional, ‘top-down’ developed vocabularies.

# 1 Introduction

## 1.1 Research overview

In 1996 a small group of molecular biologist decided to create a controlled vocabulary for the purpose of indexing gene products from any species. Their aim was simple: to design a universal tool that would facilitate data integration across different databases containing similar types of genetic data. Traditionally gene products were indexed in these databases with idiosyncratic or bespoke vocabularies, and this precluded more sophisticated cross-database searches which might leverage the large quantities of genomic data rapidly being deposited into these resources by multiple research groups across the globe.

The solution to the problem devised by this small group of molecular biologists was the Gene Ontology, a controlled vocabulary now used extensively in the domain as a standard for the representation of gene product functions. The Gene Ontology is made up of over 35000 individual terms for describing what gene products do in biological contexts. It has vastly simplified the process of analysing gene functions in complex scenarios, and has enabled bioinformatics specialists to create sophisticated tools for molecular biologists to interrogate complex gene data.

Without the Gene Ontology, it would be difficult, if not impossible, to make sense of the huge volumes of scientific data produced by the molecular biology domain at an escalating rate. However very little research exists looking at how the Gene Ontology was constructed, questioning the rationale behind its design, and asking how alternative ways of classifying gene products might improve the ontology.

The vocabulary is built on a philosophical framework for understanding scientific concepts known as *ontological realism*, a commitment which places very strict criteria on the kinds of concepts that are permissible to be described in the Gene Ontology. Ontological realism claims to offer rules which render vocabularies like the Gene Ontology more scientific and more objective, yet some biologists have questioned their suitability for designing controlled vocabularies that represent what ordinary working biologists know about gene products.

Certainly there is no research on the Gene Ontology from a library and information science perspective looking at ontologies in the molecular biology domain, and therefore this thesis is motivated to explore the following research questions:

**Question 1.** Why is ontological realism a core tenet in the design of ontologies for the molecular biology domain?
**Question 2.** How have the Gene Ontology developers tried to make the vocabulary objective?
**Question 3.** What are the limitations of the Gene Ontology as a universal classification for gene product functions?
**Question 4.** How can the Gene Ontology be improved?

Further to these questions, the aims of this research are to:

**Aim 1.** Understand the rules for creating ontologies unique to the molecular biology domain
**Aim 2.** Explore how the Gene Ontology developers created their controlled vocabulary

**Aim 3.** Investigate how the Gene Ontology is used in analysing gene data

In order to achieve these aims and answer the key research questions, the thesis has the following objectives:

**Objective I.** Analyse the representation of a single GO term according to different epistemologies for concepts

**Objective II.** Test GO vocabulary guidelines against a major international standard for vocabulary construction

**Objective III.** Conduct a critical discourse analysis of GO developer discussions on a Gene Ontology mailing list

**Objective IV.** Analyse the reasons why GO terms are removed from the ontology

**Objective V.** Explore how Gene Ontology users cite and employ GO ontology data in their work

The Gene Ontology has revolutionised the way molecular biologists interrogate empirical information like gene expression data. However the rapid development of the ontology and its quick adoption by molecular biologists eager to have *any* kind of domain knowledge representation to support bioinformatic applications, has meant that the Gene Ontology now exists largely without competition. It is a tool lauded as the solution to numerous information management problems in the biosciences domain, a universal classification to unravel the mysteries of the genetic code.

This thesis questions the Gene Ontology's aspiration of creating a universal classification for gene products, challenges its commitment to ontological realism and questions whether in trying to be objective in representing scientific concepts, the developers may have created new problems for representing gene functions.

The research is of value because it may indicate ways to make the Gene Ontology better.

## 1.2 Classification for the growth of knowledge

### 1.2.1 A research assumption

This thesis is based on the assumption that the growth of knowledge in the biosciences domain requires the classification of objects in Nature into meaningful groups

Biology is the study of organic life. It seeks to explain the variety of forms and functions which exist in Nature, from the smallest string of biomolecules to the sophisticated behaviour patterns exhibited by the largest organisms. The reference to the 'biosciences domain' is intended to encompass the broad range of auxiliary sciences which are allied with traditional biologists, including biochemists, biophysicists and bioinformaticians who, although primarily interested in biology, frequently apply knowledge from other scientific subjects like chemistry, physics and computer science in order to support the growth of knowledge in biology.

The 'growth if knowledge', refers to the purpose of biology as a scientific subject being to provide better explanations for observable phenomena in Nature. To grow knowledge in biology is to make sense of these observations by positing relationships between the entities understood to exist in reality by biologists, and phrasing these theories into nomological statements such as 'Deciduous trees always lose their leaves in the Autumn'. Law-like statements and the theories they relate to can be tested by the empirical method. Thus by verifying or refuting statements, knowledge about Nature can develop and improve, and my starting assumption is that it is classifications and classificatory practices which are especially important to the growth of knowledge by the empirical method in the biosciences domain.

The biologist observes Nature in an effort to identify instances and classes in reality such as 'the diseased Oak tree in John's garden' (an instance), 'viral infection' (a class), 'mutated DNA sample 72 from last Tuesday' (an instance) and 'Duck-billed platypus' (a class). Knowledge in biology grows largely by the assumption that similarities exist between class members. When the biologist observes a new instance and places it within a class, implicit in this act of classification is the understanding that this new instance shares certain essential properties with the other members of that class. Thus, the biologist can infer by similarity. A very common example of this process in modern biology is the classification of proteins according to similarities in their primary amino acid sequence.

Class membership implies common behaviours and explanations for structures. For a biologist to state that an elephant and a mouse both belong to the class 'mammals' is to claim that both organisms share essential properties which are mammal-like, and can be used to make predictions. The mouse and the elephant differ enormously in their physical appearance, and yet they share many physiological characteristics, such as being warm-blooded and producing live young. Equally, for a biologist to classify a newly identified gene as a tumour suppressor is to make a scientific statement about the structure and function of this new gene.

What is meant by the usage of the term 'classification' in this these? In a general sense, 'classification' is used to mean 'the act of classifying', where to classify is to arrange or organize by classes. A class is a group of things such as objects, processes or ideas which share common qualities or attributes. That which determines whether objects have qualities in common with one another is

an important question in this thesis, and in broad terms we can say that either reality, the human mind, or some combination of the two is the arbiter of what defines a class or category.

The term 'biological classification' is used to refer to the act of classifying biological entities – organisms, genes, functions– in the biosciences domain. Parallels exist between biological classification and what in the library and information science (LIS) domain is understood as 'bibliographical classification', or document classification. This application of LIS methods to the study of classificatory practices in the molecular biology domain is somewhat original, but one of the aims of the thesis is to demonstrate the validity of using LIS methods to investigate classifications in biology.

The word 'classification' is also used in this thesis to refer to the result of classifying. A classification can therefore be a representation, in words or pictures, of the groups of classes created by classifying. The familiar Linnaean hierarchy of plants and animals is an example of a classification in biology. The Library of Congress Classification [1] would be an example of a classification in the LIS domain. To illustrate, several examples of classificatory practices in biology and the classifications they produce are summarised in Table 1. Furthermore, the word 'taxonomy' is used specifically in the context of the biosciences domain to refer to the science of classifying organisms. Included in the act of classifying organisms are the processes of identifying organisms, describing organisms, and giving organisms names.

**Table 1: Examples of classifying and classifications in the biosciences domain**

| Biological entity being classified | Examples of the entity | Example of classification produced |
|---|---|---|
| **Organisms** | Beaver, snowdrop, tubercle bacillus | NCBI Taxonomy Browser [2] |
| **Genes and proteins** | p53, E-cadherin | PFAM database protein families [3] |
| **Enzymes** | amino-acid *N*-acetyltransferase | Enzyme Commission's numerical classification of enzymes [4] |
| **Functions** | Electron carrier activity | Gene Ontology classification of molecular functions [5] |
| **Phenotypes** | Bilateral club feet | Human Phenotype Ontology [6] |
| **Anatomical structure** | Heart, renal artery | e-Mouse Atlas Project [7] |
| **Medical concepts** | Diagnostic errors, genetic variation | Medical Subject Headings (MeSH) [8] |

To return to the starting assumption where it is claimed that the bioscience domain "...requires the classification of objects in Nature into meaningful groups", one might refute this claim by trying to identify examples of explanations in biology which do not rely in any sense upon classification as defined above. Basic knowledge claims about physical or chemical structures, for example the double helical solution for DNA, could perhaps be posited in a form which makes little or no reference to classification schema. However, true theories in biology are normally founded on classifications, and so to state 'The structure of DNA is a double-helix' is to make reference to a range of pre-existing classifications in biology relating to different kinds of biochemical properties, to different classes of organic molecules inside living cells, and to different types of macromolecular

structures. The double helical solution to DNA only makes sense and is explicable in the context of these biological classifications.

Furthermore if we take other biological theories, one can argue that classification always has a role to play in justifying theoretical statements. The theory that 'DNA is transcribed into a complementary RNA strand' requires the biologist to classify nucleic acids into different groups, and these classes determine the roles and behaviours of these molecules in biological processes. The Central Dogma in biology states that information may only flow from DNA to protein: proteins cannot change genomic information [9]. Here the biologists classifies biomolecules into different groups and articulates the hypothesis that is the Central Dogma through the properties of these groups, such as the privileged position of DNA as a source of biological information which cannot be altered by other classes of molecules like RNA and protein.

The Central Dogma in biology exists now in a modified form [10]. Processes such as reverse transcription (the conversion of modified RNA *back* into DNA), RNA replication (organisms which use RNA as an exclusive template for genetic information) and methylation (inheritable modifications to DNA performed by proteins) have undermined Crick's original version of Central Dogma in which information flow in life is unidirectional, from DNA to proteins.

This example illustrates an important theme in this thesis, which is that new knowledge inevitably changes how classifications are applied, and consequently how scientific theories are understood.

Equally, the statement 'Reptiles evolved from dinosaurs' indicates that members of the class of organisms biologists call 'reptiles' share common ancestors which themselves form a class known as 'dinosaurs'. Organisms outside this class evolved from different parental lineages and in a clear historical sense there is a distinct pathways of causality leading back in evolutionary time which, as verified from the fossil record and other evidence according to the best, current knowledge, connects reptiles today to the dinosaurs of yesteryears.

The work of many of the earliest biologists, labelled as naturalists, anatomists and microscopists in current language, were experts in the describing the *forms* of life, of articulating the classes of organisms in Nature and the kinds of parts these organisms possessed. The decoding and mapping of genes and structural studies of proteins in the twentieth century are similarly extensions of these studies of forms in Nature. The history of biology is the history of the biologist's efforts to explain relationships between the forms and functions of organic life [11]. In order to understand the origins of modern classification systems for gene products in the twentieth century, and the importance of these classifications for the growth of knowledge in the biosciences domain, it may prove useful to outline the history of several major classificatory practices in biology, in an effort to make the rationale for my starting assumption a little clearer.

### 1.2.2 Early essentialist classifications

The writings of Aristotle count amongst the earliest efforts at classifying organisms into meaningful groups based upon shared characteristics[12]. Most importantly, Aristotle's classifications for organisms are understood as being early examples of *scientific* classifications, since they are constructed with a view to explaining patterns in Nature by the membership of instances in these groups [13-15]. Aristotle toured the Mediterranean, collecting samples of different organisms,

examining their physiology, their anatomy, their behaviours, and attempted to create logical classes (what biologists might recognise today as taxa or species).

Aristotle's writings outline the subdivision of creatures and plants according to their how they nursed their young, their diets, number of limbs, methods of locomotion, morphology and flowering patterns [16, 17]. Aristotle arranged organisms into hierarchies from the simple to more complex creatures, and this reflected his commitment to what is known as the *scala naturae*, where God occupies the pinnacle of the hierarchy, angels lie beneath and so the arrangement continues down through the animals and plants to the simplest minerals at the base.

Aristotle's method does differ from folk classifications of Nature because his attempts were recognisably scientific. Aristotle was a realist; he believed that reality existed and could be observed objectively. Aristotle was therefore an objectivist, and believed in the capacity of Man to observe reality impartially and, in so doing, recognise order (an order presumed to be imparted by a divine creator).

A dispute exists in the history of philosophy as to whether members of classes in Aristotle's researches were deemed to be logically related because they shared *essential properties* [18-21]. The doctrine of essentials is that membership of a class is determined by essential properties, a concept influenced by the Platonic theory of forms. Each instance of a cow or a tortoise is related to a prototypical example of a cow or tortoise. There exist ideal forms, and these ideals – or *universals* – govern those essential properties that make a cow a cow. The class 'cow' forms what is known as a *natural kind*; natural kinds are logical classes in Nature, like the natural kinds 'dog', 'fish' or 'cactus'.

Classification according to essentialism is to create classes and relationships between classes according to the natural kinds in reality. Natural kinds are determined by the essential properties of objects. Essential properties inhere in reality, within objects in reality. By observation and a scientific approach, the human mind may grasp the identity of these essential properties and classify accordingly. A classification created according to essentialism is therefore the mirroring of the natural order present in reality.

The Linnaean classification for plants and animals is another example of a classification which appears to be essentialist in origin.

However as mentioned previously, there is an argument that essentialism bears little relevance to the early history of biology [21]. Accusations are levelled at Ernst Mayr for creating a simplified version of the role of essentialism in biology [22-24] to support his interpretation of species in modern biology. Some historians deny that Aristotle or his successors like Linnaeus were essentialists at all. For example Winsor [20] argues that the classificatory approaches of post-Linnaean taxonomists are a typological practice. Individual animals and plants served as prototypical exemplars for broader categories. The ontological status of species, according to Winsor, is polytypic in that different sets of features are sufficient to determine membership in a group.

This is quite different to essentialism, where the natural order of plants and animals is governed by necessary and sufficient conditions.

Likewise Muller-Wille [25] argues that Linnaeus' methods were not guided by a commitment to essentialism. Much like Winsor's thesis, he argues that Linnaeus was a pragmatist using inductive

methods and observable characteristics to create categories that best matched reality: a biological concept of species, rather than essentialist. Winsor and Muller-Wille's offer a defence against the accusation that early biologists, and biologists today, are committed to an essentialism which tries to represent an idealised natural order, or a world of unchanging classes of species, created by God.

However biologists – like most scientists - are realists, and resist the idea that reality, and our knowledge about reality, is a psychological or social construct [26, 27]. Like Aristotle, they are committed to objectivism in some form. They believe they can observe reality from a neutral perspective and consequently believe their theories to be value-free.

Biologists *behave* as though they believe in essentialism. The molecular biologist gives names to genes and groups genes into families. All the genes in a family are considered to share common properties, properties governed by the form of those genes. This type of thinking is very close to a kind of essentialism. Essential properties explain regularities in Nature, and permit inference across logical classes.

Therefore the modern biologist may not believe in the *scala naturae*, yet would find little which is philosophically objectionable about essentialism.

Although the doctrine of essentialism is not universally accepted as a guiding principle in the history of biology, my argument is that the concept has relevance to the research programme in modern biology, and especially for the work of classifying genes and gene products. This contention is based on the work of Marc Ereshefsky [28-31] and his investigations into competing species classification principles. Ereshefsky criticises the essentialism inherent to the methods used in biological taxonomy, and demonstrates severe weaknesses in such a philosophical doctrine to creating scientific classifications in biology

Modern taxonomies for species are, in appearance and usage, essentialist. Living organisms are ascribed to single classes. Cluster concepts or polytypic rules may be used deep into taxonomies at the most detailed levels, yet biologists do not speak in terms of individuals being members of multiple species. If anything, molecular biology has reasserted the importance of essentialist doctrine to the success of the life sciences research programme. Genes vary between individuals, proteins can have multiple functions, but the purpose of biology is to model the fundamental properties of these units in biological systems. This is no oversimplification: molecular biology is committed to understanding the essential properties of genes and proteins at specific points in time in prototypical, idealised biological situations.

The modern, molecular biologist is a reductionist [32-35], seeking to explain biological phenomena at different levels of organisation: the molecular, the cellular, and the physiological. Explanations at each level are intended to be reducible to explanations at underlying levels. An example would be that reductionist biology seeks to explain the human capacity for humour in terms of the behaviour of the bio-molecules constituting the laughing person. Biology therefore aims to create coherent models of biological systems, and a method to achieve this by is by identifying instances and classes in Nature, defining logical relationships existing between them and the natural laws governing their behaviour.

This reasoning is founded on a form of essentialism which accepts that universals in reality can be mirrored in a computer [36]. This view is a simplification, and there are some biologists who argue for a realism which does not adhere to a doctrine of universals, for the metaphysics of universals do not rest comfortably with science [34, 37, 38].

Nevertheless, reductionism and attendant inclinations towards essentialism undergird much of modern biology.

## 1.3 Linnaean hierarchies and the *scala naturae*

In the period following Aristotle and his immediate successors like Theophrastus and Pliny the Elder until the beginning of the Renaissance period in Europe, there are many treatises and works that are recognisable as precursors to modern empirical biology [39]. The principles of anatomy, medicine, botany and zoology can be found in medieval European and the works of many Arab scholars, and by the 18th century much important progress had been made towards what biologists today would understand as the scientific process of taxonomy, or the classification of the natural world.

The classification most familiar to most non-biologists is what is commonly referred to as a Linnaean classification or a Linnaean hierarchy. Carl Linnaeus (1707-1778) is celebrated as the original taxonomist, adopting taxonomic principles that are still used by biologists today, with his hierarchical ranking of kingdoms, classes, orders, genera and species. Linnaeus however was but one of many 18th century naturalists, including Buffon, Lamarck and Cuvier [40], who made significant contributions to the systematization of the natural world. Their various almanacs, compendia and nomenclature systems classified cornucopias of new species brought back from newly discovered corners of the globe, and the names for many plants and animals date back to this fertile intellectual period.

Linnaeus and his peers were endeavouring to solve three broad problems in the classification of the natural world.

Firstly they sought a means to provide a methodology for *identifying* species. Naturalists collected different kinds of plants and animals but needed to distinguish between the different types. Classifications according to physical characteristics such as flowers offered simple means by which to place unknown organisms into family groups.

Secondly, they created nomenclatures for organisms. Polynomial naming systems created long, non-standardised and often confusing names for different species. Linnaeus' work was special because it created a simple, binomial names system (the genus name coupled to a species name) which biologists still use today.

Thirdly, the early taxonomists sought to faithfully represent the Divine order or *scala naturae* present in Nature. Early classifications tried to make sense of the world by representing the perfect harmony between different species which seemed to fall, by virtue of similar anatomical structures or biological functions, into transparently 'natural' groupings (such as types of fish, trees, or birds) and hierarchies from the simplest creatures and up via Man and the angels, to God Himself.

The early 18th century naturalists like Linnaeus were all men of faith, believing that God had created the Earth and all its myriad forms of Life. These men (and they were almost exclusively men) sought to observe patterns in Nature, and mirror these patterns in so-called 'natural' classification systems.

Yet their theological motivations were coupled to an empiricism and belief in scientific methods. In attempting to create comprehensive collections of species from different regions and organising these species into natural classifications, they assumed there to be a natural order in keeping with the precepts of essentialism described in the previous section.

Linnaeus' classification method established hierarchical ranks which subsumed lower-order classes, and in so doing satisfied the three problems indicated above. His justification for classes, based on the observation that interbreeding between individuals created classes of species, was controversial in 18th century Europe for its overt references to sex. Despite the novelty of this 'sexual' classification, Linnaeus created hierarchies which permitted the naive naturalist to simply navigate leaves of the schema tree (such as down through the leaves from 'quadraped') aiding enormously in the complex task of species identification. The binomial names formed the genus and species at the most detailed level of the classification, thus connecting in a logical manner identity and nomenclature. Linnaeus's system also respected the *scala naturae,* situating Man in his preeminent position in the hierarchy and the simpler organisms.

Is the Linnaean system natural or artificial – does it reflect universals in reality or is it contrived for simplicity? Ernst Mayr maintains that Linnaeus was philosophically grounded in a typological (or essentialist) species concept [11], and that consequently the Linnaean hierarchy is indeed an attempt at a natural classification system. However it is clear that many of the classification decisions Linnaeus made, especially those associated with his 'sexual system' for the classification of plants, are conveniences to facilitate the identification of species. Although Linnaeus may have aspired to a natural system, he was overwhelmingly pragmatic in his classification, and was clearly not averse to creating artificial classifications provided it aided in the practical task of identifying species in the field. He was after all primarily a teacher; essentialist leanings did not limit Linnaeus' success at designing useful species classifications, many features of which are still retained in modern taxonomies for species.

## 1.4 Darwin's species

### 1.4.1 Creatures according to their kinds

Charles Darwin (1809-1882) was a naturalist much like Carl Linnaeus and, like the other 18th century naturalists, his interests ranged across botany, zoology and geology [41, 42]. Darwin sought to explain how different species have come to live in different places – where did they come from and what could explain the similarities between different species?

His solution to this problem, which previously had been explained simply in theological terms ("And God said, "Let the land produce living creatures according to their kinds..." Genesis 1:24) was that new species evolve gradually from their ancestors through a process of change he called 'natural selection'.

The complexities of the theory of natural selection will be touched on only briefly here. After the manner of Ernst Mayr, Darwin's theory can be easily understood by splitting it into five different sub-theories, see Table 2.

Table 2: Darwin's Theory of Natural Selection, as simplified by Ernst Mayr [24]

| Darwin's Theory of Natural Selection | | |
|---|---|---|
| **Sub-theory 1** | Evolution | Nature is perpetually changing rather than existing in a constant state |
| **Sub-theory 2** | Common descent | Different groups of organisms can share the same ancestral species |
| **Sub-theory 3** | Multiplication of species | Diversity is created by species splitting into new species |
| **Sub-theory 4** | Gradualism | New species are created by a slow, incremental change process rather than abrupt, sudden jumps |
| **Sub-theory 5** | Natural selection | New species characteristics which confer a survival advantage over the alternatives are selected for |

Darwin's theory has proved pivotal to biology, and exists in a modified form today as the modern evolutionary synthesis which resolves the theory of natural selection with population genetics and what biologists now know about evolution. Yet for all the answers Darwin's theory offers to the question of the origins of species, so it has created difficulties, and possibly even intractable problems, for taxonomy.

### 1.4.2 Changing classes

Darwin's explanation for the origin of species was controversial because it undermined several very strong and long-standing philosophical commitments. Aside from convincing many people that species diversity in Nature was not the work of an intelligent creator, Darwin changed the paradigmatic assumption that species are atomic, unchanging units [43]. Evolution describes the slow transformation of one species into another, which implies that anything we consider to be a discrete class in Nature is in no way constant – it is derived from the transformation of an ancestral species, and itself may evolve into a new 'class' in the future. This is in contrast to the pre-Darwinian taxonomists like Linnaeus who created static, ahistorical classifications according to observable characteristics alone.

Change as a process to be accounted for in taxonomies undermines essentialism as a doctrine for species classification. Darwin's theory offered an explanation for the classes and structure of Linnaean classification and, like all good theories, improved upon it by proposing rearrangements according to the evolutionary history of a species. Linnaean taxa, seen through the lens of Darwin's theory, are determined by a common evolutionary ancestry. The natural kind 'birds' exists because all the different types of birds have evolved from common ancestors many millions of years ago (and as supported by the fossil record).

Evolution, being non-teleological, does not rely upon on any Divine plan, purpose or 'scale of perfection'. Man, though arguably a complex and in many ways unique species, does not occupy a special, privileged position in a species taxonomy, and hence the *scala naturae* is not sustainable. At the same time, the theory of natural selection, and especially the sub-theories of evolution and the multiplication of species, severely undermined long-standing commitments to species as natural kinds differentiated by their essential properties. Dogs and cats, rather than being discrete, atomic types, became related by continuous gradations of difference. The idea of universals or ideal 'forms'

for all God's creatures blurred and merged as Darwin's theory of natural selection became accepted as paradigmatic thought in biology.

What then was a species to Darwin? A species was not a typological concept because variation within species is not compatible with the assumption that definitive characteristics exist that determine class membership in any 'natural kind'-sense [24]. Species might be a nominalist concept, a mental construct we give a name to, although this is anti-realist and not consistent with normal science. This leaves us with an evolutionary species concept, whereby a species is defined as that which has evolved to be different. Such a concept for species cannot deal with gradualism or offer any means to distinguish species at present in Nature. Mayr argues that Darwin accepted the biological species concept: species as reproductively isolated organisms (akin to Linnaeus' sexual classification system).

Darwin himself never explicitly defined what he meant by 'species', and struggled to understand the term during the course of his researches. He drew at times on each of the species concepts mentioned above, and his theory of natural selection is in many ways anti-essentialist, based on populations, gradations of difference, and statistical measures determining memberships of classes. However he never explicitly argued against essentialism. Biologists today still do not agree upon what the concept of 'species' actually means, and consequently there are different hierarchies and taxonomies created according to different species concepts.

## 1.5 Classifying the products of the genomic revolution

### 1.5.1 The information organism

After Darwin's theory changed the way society understood the origins of diversity in the Natural World, there were to follow in the 20$^{th}$ century two further major domain shifts in biological thinking: biochemistry and genetics [44].

As JBS Haldane once wrote "...life is a pattern of chemical processes" [45] and the early 20$^{th}$ century saw a slew of new biochemical techniques designed to explore these chemical processes, such as the metabolism of sugars or the structure of enzymes. In parallel to the work of biochemists, the oftentimes abstract and theoretical work of geneticists demonstrated that the characteristics of organisms were inherited from one generation to the next, and inherited according to special rules, via units they dubbed 'genes'. As Morange argues [44], much of modern molecular biology is founded on a synthesis between the work of the biochemists and the insights of the geneticists. The work of both camps also had to be reconciled with Darwin's theories of gradual change generating new species. Not until the structure of DNA was solved was a definitive candidate identified for the source of an inheritable, chemical material that might store genetic information, and act as a raw material for evolution.

The discovery of the structure of DNA and the rapid insights it brought to biochemistry and genetics created a strong metaphor in biology which is the metaphor of 'genetic information'. The genetic code is the specific sequence of base-pairs in nucleic acids which is de-coded to produce specific proteins. This theory is the basis of the 'One Gene, One Protein' hypothesis. The genetic code is thought to act as an information storage device. DNA stores information which is interpreted in the context of a living cell into various biochemical and cellular processes. Evolution is the refinement of the meaning of this genetic information towards ever more successful ends via natural selection. If

the work of the biologist is to interpret genetic information, then the geneticist identifies 'words' in the genome whilst the tools of the biochemist reveal what these 'words' mean.

The metaphor of genetic information has reinforced reductionism in biology. It presupposes that genetic traits can be explained in terms of the molecular constitution of an organism. Through successive levels of complexity from the DNA sequence through the biochemical pathways in the cell, up through to cells interacting within tissues and within whole organisms, life can be seen through the genetic information metaphor as a flow of information. As such, life can be modelled in abstract, logical terms, on the page and in the computer. Genes contain information encoding proteins, and proteins have specific functions in biological systems which operate, under changing environmental conditions, to create cells, organisms, life.

Modern biology, driven by a commitment to a simple genetic information metaphor and reductionist thought, aims now to develop ever more sophisticated models of biological processes. Proteins acting through various pathways enact biological functions. Biologists seek to faithfully model these pathways, and if these models can be refined to the point of completion, an explanation for any biological process can be posited. I am attempting here to draw an analogy between the efforts of 18th century naturalists to produce complete 'mirror-images' of God's work in the world of plants and animals with the efforts of modern biology to create models of complex biological systems which are 'mirror-images' of precisely how Nature works.

The metaphor of genetic information sustains an assumption that is still implicit in biological thinking today which is that the genetic code can contain all the information necessary to create an organism. Several philosophers of biology have criticised this persistent metaphor in biology [10, 46-49]. If the genetic code is considered as a part of an information system in the Shannon and Weaver communication sense, genes cannot be said to encode all the information needed to express even the simplest traits. To do so is to ignore the effects of the environment on the generation of characteristics, and it is the complex interplay of genes and the environment which create forms of life. A simple example would be the role diet plays in determining the weight of organism at any point in its life. Genetics partly determine the size of a human, but calorific intake, exercise and a host of other environmental factors must be taken into account if one were to try and predict how heavy a person will be at a specific point in the future.

### 1.5.2 Indexes for the Book of Life

The genomic revolution expanded the creation and application of classifications in biology. Biologists were quick to adopt computer technology to support their investigations into the genome, building large species databases of genomic information which were made freely available to scientists across the globe using nascent computer networks. As biology scaled-up to deal with the many hundreds of thousands of genes discovered in each species, and the many-fold proteins transcribed from these sequences, nomenclature problems proliferated. Each gene has many names and abbreviations and despite the best efforts of various committees and domain authorities, semantic web techniques are still being improvised today to cope with absence of name authority control in biology. The Swiss Institute of Bioinformatics for example lists over 120 different nomenclature references for naming genes and gene products [50].

As more and more information about genes and biochemical processes was discovered in the late 20th century, the need to group genes and proteins into meaningful 'families' based on shared

biological functions became ever more pertinent. For example, the Hox family of genes are all involved in developmental programming; to discover a new gene with strong sequence similarity to a Hox family member is a simple way to infer a function for an unknown gene, without labouring for many years in a laboratory trying to confirm the same result. Beyond sequence similarity, secondary and tertiary structures to proteins contributed to broader, non-homologous families and groupings for genes which were also important to the work of the molecular biologists ascribing functions to entities. All this knowledge is now embedded the species databases as metadata, metadata on sequences and homologies and higher-level level structural similarities, and these classifications led to numerous vocabularies describing different biological areas (see Table 3).

For example in biochemistry, enzymes are grouped into families based on the types of chemical reactions they catalyse. The shape and structure of enzymes determines their specificity, and so classifications exist for enzymes based on their shape and they way molecules can fold. These classifications allow biochemists to infer roles for newly discovered proteins. An example is the protein kinase domain family [51], members of which share the common function of transferring phosphates from nucleotides to amino acids. The biologist can infer that a newly discovered protein with this domain might have this function in a cell or, armed with many thousands of unknown transcript sequences, can postulate thousands of likely functions in a single analysis.

Homologous structures and functions permit the assignment of unknown entities to existing groups and families. Assessment of similarity is founded on sequence similarity. Much new biological knowledge is predicated on computer algorithms which measure how similar sequences are and then assign to a known class. The approach is a little like search algorithms in information science which attempt to assign meaning to a document by ranking important words. Modern biology takes much the same approach, and various classifications for genes and proteins form the indexing terms that are used to infer the function of an unknown sequence.

**Table 3: Example classifications in the bioscience domain**

| Classification | Object classified | Biosciences sub-domain |
|---|---|---|
| **Swiss-Prot keywords [52]** | Proteins | Proteomics |
| **The Institute for Genomic Research (TIGR) roles** | Genes | Genetics |
| **Commission on Plant Gene Nomenclature (CPGN)** | Genes | Plant science |
| **The What Is There (WIT) System [53]** | Genes | Metabolics |
| **Munich Information Center for Protein Sequences (MIPS) functional classification [54]** | Proteins | Proteomics |
| **TAMBIS ontology [55]** | Proteins and nucleic acids | Molecular biology |
| **Enzyme Classification [4]** | Enzymes | Biochemistry |
| **IUPHAR Database classifications [56]** | Receptors | Pharmacology |
| **Cluster of Orthologous Groups (COG) [57]** | Proteins | Evolutionary biology |
| **Transporter Classification database [58]** | Transporters | Molecular biology |
| **Monica Riley's classification for *E. coli* [59]** | Gene products | Microbiology |

### 1.5.3 Big data, big science, big problems

The late 1990s saw a slew of new genomic techniques designed to decipher the tens of thousands of genes found in the genome of major model organisms like yeast, mice and Man. The Human Genome Project was as much about the final draft of the Human Genome as it was about learning ways to decode large quantities of genomic information quickly and cheaply [60]. The average laboratory scientist in a biology department can now sequence genomic data at a fraction of the cost from only ten years ago. Consequently, the biosciences domain is overwhelmed by huge volumes of genomic data coupled with transcript and expression data, protein information and, naturally, empirical data from the experiments trying to make sense of what these entities do in a living cell.

Molecular biology today is driven by informatics, by databases warehousing massive quantities of data, by bioinformatics designed to process this data, by statistical and computational methods aimed at extracting knowledge from the information that accumulates, inexorably, on a daily basis. I say that biology is 'driven by informatics' because the automated processing of large datasets has become a requirement for the majority of biologists to do their work. In order to perform even simple experiments, the molecular biologist must be able to access domain-specific databases, manage large quantities of empirical data, and manipulate this data using oftentimes complex computational techniques. And to make sense of all this data using computers it must be classified.

At the completion of the Human Genome Projects, biologists turned to one another and asked: what next? Gene sequences must be named. They must be attributed to a species and a locus. Gene products must have functions, expression profiles, pathways. Proteins must be members of families, must be expressed in particular contexts and tissues. There quickly came the realisation that sequencing the genome was only the start and, if anything, would be the easiest task compared to actually making sense of what all this sequence data mean. Nascent standards for genomic data, the databases storing the information and early examples of systematic approaches to classifying gene functions were seen as but starting points for more complex classifications that might serve the purposes of an informatics driven biology. Big science or e-science are seen as the future, and this is reflected in current European Commission funding for e-infrastructure projects in the biomedical domain.

However scaling up data analysis in the biosciences can only be achieved by scaling up classifications and metadata standards to add rich, contextual information to the essentially meaningless strings of sequence data churned out by gene sequencers.

To return to the initial thesis assumption at the start of the introduction, the growth of knowledge in the present, informatics-driven biosciences domain requires the classification of sequence data, be it gene, transcript or protein data, into meaningful groups. These classifications may be otherwise understood as keywords, subject headings, families, indexes, tags, ontology terms or metadata. Yet these are all shorthand for classifications and serve the common purpose which is putting biological entities into meaningful groups. Knowledge cannot grow in the biosciences without first classifying the genes, the transcripts, the proteins that have been discovered, and using these classifications to make inferences and test theoretical propositions.

Classification in biology is therefore the object of my thesis, and the Gene Ontology, designed by biologists and for biologists to classify gene products, is the classification which is used to answer the main research questions in this research.

## 1.6 Why the Gene Ontology was created

The post-genomic revolution is marked by the hope that the various domains and activities in biology can be brought together and harmonised into a single grand project. Thus what is known about human physiology, cell cytology, eukaryotic molecular processes and the genetics of organisms can, it is hoped, be resolved into a single, coherent knowledge system. This aspiration in biology is explicitly reductionist and deterministic. If it is known that eating a Mars bar elevates circulating insulin levels in the bloodstream, the biologist intends that the explanation at the level of endocrinology can be explained by the synchronised expression of sets of genes at the genetic level.

One solution to this problem of harmonising the disparate and heterogeneous worlds of biological knowledge is through the application of ontologies [61], and this is the aim one of most sophisticated ontologies in biology, the Gene Ontology [5, 62, 63].

The Gene Ontology was launched in 2000 and as of July 2010 consists of 32063 terms. In its own words, the Gene Ontology declares its ambitions as follows:

> "Where once biochemists characterized proteins by their diverse activities and abundances, and geneticists characterized genes by the phenotypes of their mutations, all biologists now acknowledge that there is likely to be a single limited universe of genes and proteins, many of which are conserved in most or all living cells. This recognition has fuelled a grand unification of biology..." [5]

In a sense, the aim of the Gene Ontology from 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 biology in their work. From an early GO Consortium paper:

> "The original intent of the group was to construct a set of vocabularies comprising terms that we could share with a common understanding of the meaning of any term used, and that could support cross-database queries."[63]

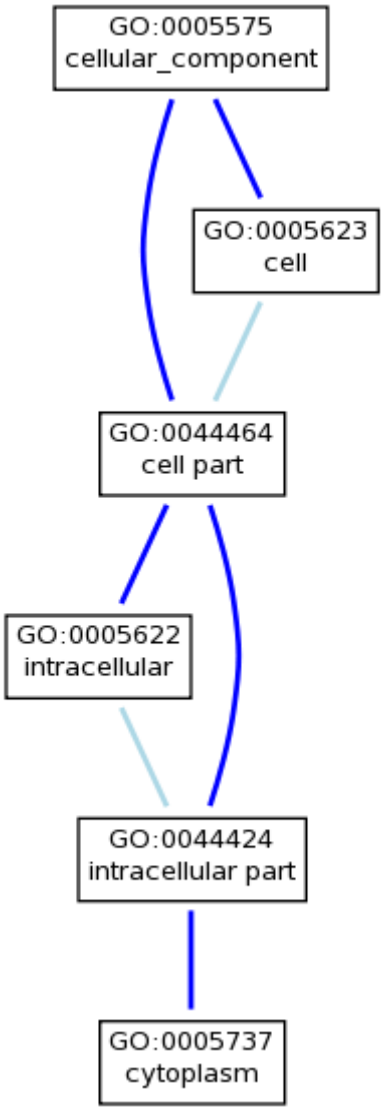
The Gene Ontology was conceived through the cooperation of three groups from traditionally quite distal domains in biology. Through their initial efforts, many other research groups representing the interests of other sub-domains in biology have been recruited to what is now known as the Gene Ontology Consortium. Specialists recruited by the Consortium and working in collaboration with special interest teams have developed the ontology. Contributions are invited from the wider biosciences community via various community wikis, Sourceforge development forms, expert meetings, mailing lists, and invitations for direct contact with the ontology developers. The design and quality of the Gene Ontology have been submitted to scrutiny in the literature [64-68] although very little work exists examining how the Gene Ontology is being used by biologists in their published work.

The Gene Ontology was therefore created to solve a special problem in molecular biology, which was how to classify gene products. This necessity existed because 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 was chosen because biologists wanted to be able to automatically reason 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 traditionally had never collaborated on creating a shared controlled vocabulary before.

The Gene Ontology comprises three independent controlled vocabularies covering cell components, biological processes and biological 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 relations within the Gene Ontology, and working examples, are shown in Table 1 [69]:

Table 4: Examples of relationships in the Gene Ontology

| Relationship | Example |
|---|---|
| **Is_a** | "mitochondrion *is_* intracellular organelle" |
| **Part_of** | "mitochondrial membrane *part_of* mitochondrion" |
| **Has_part** | "nucleus *have_part* chromosome" |
| **Negatively_regulates** | "cell cycle checkpoint *negatively_regulates* cell cycle" |

## 1.7 Formal ontology and ontological realism

### 1.7.1 Formal ontology

The tradition of classification in biology and the new wave of informatics-supported knowledge discovery for the biosciences (encapsulated by the 'e-science' banner) intersect one another in the field of ontology.

Ontologies are representations of domain knowledge, of concepts in a domain, captured in a structured language amenable to computer programming. In many senses the modern usage of the term 'ontology' touches on traditional representations of knowledge, such as controlled vocabularies, thesauri, classifications and the like. As Gruber writes in his seminal ontology paper 'A translation approach to portable ontology specifications':

> "A conceptualization is an abstract, simplified view of the world that we wish to represent for some purpose. Every knowledge base, knowledge-based system, or knowledge-level agent is committed to some conceptualization, explicitly or implicitly. An ontology is an explicit specification of a conceptualization." [70]

Formal ontology appeals to Aristotelian logic and is an effort at representing concepts and their relationships [71]. For each class in an ontology there exist instances in reality. Instances in reality can be grouped as natural kinds, with common essences. Formal ontologies articulate these relationships between these natural kinds. The ontologist encodes the logical relations between classes in an ontology as axioms to permit computer tractability [36]. The philosophy of formal ontology clearly complements the e-science ethos [72].

Ontology designers for the sciences do differ over their underlying philosophical commitments inherent in the act of creating an ontology. Subjectivist exponents argue that concepts in an ontology have their referents in the thinking minds of scientists rather than in reality [73]. Ontology is representation of these concepts and the role of these concepts in the cognitive processes of individuals [70].

Strict, formal ontology, which is a doctrine commonly adopted by ontologies for the biosciences, is realist in dogma [65, 73-76], and denies that concepts in a domain are uniquely mental objects, being culturally relative and contingent to the thought processes of creative scientist. Rather concepts are objective and invariant. There is a unique common-sense world [77], and the scientist attempts to identify invariants in this common-sense reality and to establish relationships between these real individuals [78, 79]. Reality has structure and this structure can be captured within an ontology [36, 80]. The ontologist aims to model these culturally stable, reality-grounded, 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.

Formal ontology can be considered in relation to the traditions of analytical philosophy, and in particular the writings of Quine [81].

Formal ontology is inevitably deterministic and reductionist. The aim of creating formal ontologies to support e-science applications in biology is to minimise the vagaries of subjectivity in the domain. Formal ontology aims to anchor the practice of knowledge discovery in biology in such a way as to make data processing, inferencing, hypothesis generation and error-checking amenable to computers, *sans* the puzzling, politicising, semantic sloppiness [82] and power-brokering of the human scientist.

A challenge is therefore to resolve 'hard' epistemological model in formal ontology and ontology design for e-science with contemporary 'soft' and social interpretations of science-as-practice. In many ways, contradictions between these two lines of thought are reproduced in the 'cognitive turn' in information science, and in the IS domain's efforts to resolve hard and soft dogmas in understanding information [83, 84].

Of interest to this thesis is the question of how the contents of an ontology in the biosciences are decided. What are the criteria determining which classes and relationships are to be included in an ontology? Some ontology designers claim that the philosophical principles undergirding any ontology are really only secondary to its primary, pragmatic purpose as a tool to get jobs done in information systems.

An aim of this thesis though is to understand those rules for creating ontologies which are unique to the molecular biology domain. Philosophical principles are difficult to ignore in ontology design. Philosophical principles, whether overtly stated or tacitly committed to, have important, practical consequences for how any classification is used, and the kind of new knowledge that might be discovered in an information environment supported by structures like ontologies.

### 1.7.2 Ontological realism

Thus far, the structure and organisation of the Gene Ontology has been broadly described. It is comprised of three sub-ontologies covering areas of knowledge which biologists consider to be distinct from one another. Being separated in this manner, each of the three sub-ontologies is covered by slightly different rules and conventions dictating the ontology content. There also exist special biological situations that the GO Consortium have discovered to be difficult to describe in terms of the ontologies, and these subjects are further explicated by additional rules and qualifications. Despite the differences between the sub-ontologies, they are all constructed as diacylic graphs, all the terms can have more than one parent in the hierarchy, and each sub-ontology draws on a limited number of logical relationships between terms, which aid in reasoning across the graphs.

Behind these practical rules and conventions, there exists a philosophical principle guiding the development of the Gene Ontology project, and indeed of most of the other ontology projects found in the biosciences domain. This philosophical principle is known as *ontological realism* and in the following section, it will be explained how this philosophy has shaped the basic rules for ontologies

in the biosciences, and has served to distinguish the Gene Ontology from somewhat simpler thesauri, vocabularies or subject heading lists that might otherwise serve the purpose of indexing gene products.

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) [85]. The OBO Foundry describes its mission as "...establishing a set of principles for ontology development" with the aim of facilitating interoperability between large datasets produced by different sub-domains in biology.

Traditionally, data from diverse sources in biology – such as genetic, biochemical or physiological data – have been described by metadata standards that may be unique to that domain of study, and idiosyncratic enough to dissuade large-scale efforts to integrate these data into a coherent whole. For example, taking genomic data about the sequence and location of genes, biochemical data about the structure and function of these gene products under normal physiological conditions and combining this with medical data on various physiological markers in order to propose a hypothesis as to which biological pathways may be disrupted in a particular disease. The OBO Foundry 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 [86]. These principles were adopted by the organisation in 2006, and are listed in a short document on the OBO Foundry website. These core principles are supplemented by a set of additional principles the OBO Foundry adopted in 2008, and which are available on their wiki [87]. In addition, the OBO Foundry is 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. The OBO Foundry principles complement BFO form a conceptual system for biomedical ontologies which any collaborator with the OBO Foundry, including the Gene Ontology, are required to adhere to in order to participate

Of special interest to many ontology developers is the long-standing principle, officially adopted by the OBO Foundry in 2008, of what is termed 'instantiability'. Instantiability is the idea that all terms in biomedical ontologies must have instances in reality. This realist position, largely coordinated and defended by Barry Smith at the National Centre for Ontological Research in Buffalo, is an area of contention in the biomedical ontology domain. Heated discussion threads on the OBO Foundry message boards have developed into a number of peer-reviewed papers trying to dismiss or reinforce realist doctrine as a requirement for ontology development.

BFO is realist in dogma [65, 73-76], and denies that concepts in a domain are uniquely mental objects, being culturally relative and contingent to the thought processes of creative scientist. If anything, concepts are obstacles to the construction of useful ontologies, since they are not objective and invariant.

There is a unique common-sense world [77], and the scientist attempts to identify invariants in this common-sense reality and to establish relationships between these real individuals [78, 79]. Reality has structure and this structure can be captured within an ontology [36, 80]. The ontologist aims to model these culturally stable, reality-grounded, real-world classes as faithfully as possible. Mental

concepts are plastic in their manifestation within languages and discourse. Concepts are context-dependent, or personal and unique to the thinker. As such, ontological realists dismiss mental concepts and talk of concepts (which traditionally have been accepted as the object of knowledge representations like ontologies) as anathema to the development of good, scientific ontologies.

Subjectivist exponents argue that concepts in an ontology have their referents in the thinking minds of scientists rather than in reality [73]. Ontology is representation of these concepts and the role of these concepts in the cognitive processes of individuals [70]. Rather than being anti-realist, some biologists resist the extreme notion that mental concepts have no place in ontologies for biology.

Dumontier and Hoehndorf [88] argue that ontologies ought to represent theoretical entities which potentially have no instances in reality. Empirical science needs a place in ontologies for hypothetical entities, and a commitment to realism in BFO only limits the usefulness to the real work of real scientists

Merrill [89] targets the specific brand of realism expounded by Smith and Ceusters. Whilst not going so far as to defend any 'conceptualist' stance in ontology development, whereby ontology terms represent concepts as psychological states or perhaps as linguistic entities, grounded in a community of users, Merrill does develop the idea that biologists may adopt a more nuanced realism than the hard-line realism advocated as a necessity to quality ontologies. He criticises Smith and Ceusters opaque references to realism, arguing that theirs is a form of referentialism which is loaded with all manner of problems. Most importantly, Merrill claims that realism is in no way a requirement for ontologies which will facilitate data integration, and in fact all talk of realism can be dropped from Smith and Ceusters arguments without their approach losing anything; ontological realism is dispensable

Lord and Stevens [90] demonstrate via several case studies of knowledge representations using ontologies how realism itself can lead to ontologies which are not sufficiently complex to describe an experiment (whilst at the same time creating ontologies which themselves can be horribly complicated)

Smith and Ceusters respond [91] to these criticisms at length as they attempt to resist anything but the purest sort of realism as the philosophical basis for good ontologies. They clarify that ontological realism is not a philosophical doctrine scientists must adhere to, but a methodology for constructing good ontologies. Their argument, which is essentially the consensus even amongst their critics, is that conceptualism precludes data integration using ontologies in the sciences. The only alternative is therefore realism, and the acceptance that reality can be the only arbiter in deciding the classes which make up an ontology.

However, ontological realism has significant weaknesses. It cannot explain how 'types' are identified in a domain, other than through a Kuhnian appeal to norms and what Smith and Ceusters term 'settled science'. Arguably this sorting of universals by the consensus view of empiricists does nothing to eliminate conceptualist talk from ontology construction. Smith and Ceusters also fail to offer any evidence that an ontology constructed according to any other principle than ontological realism is demonstrably weaker than their own approach, other than in terms of facility to automate reasoning. The user is eliminated, the mirroring of reality by the ontology resisting all alternative methods for indexing biological resources.

Ontological realism is therefore the source for many of the structural features outlined in the Gene Ontology above. 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. Much like the basic forces in physics, there are a finite number of relationships between entities in the ontology. Definitions for terms delimit structural features or activities in reality, and these correspond literally to protein structures encoded by genomic sequences. Terms in the Gene Ontology 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.

## 1.8 Gene Ontology applications

If the Gene Ontology is a 'tool to get jobs done' in biology, then what kinds applications does this special classification have? Terms from the Gene Ontology are first and foremost used to create what biologists call *annotations* to specific electronic records for gene products (usually RNA transcript sequences of proteins) in particular species databases [92, 93]. Although biologists refer to these associations between GO terms and gene products as annotations, librarians and information scientists would recognise this indexing or classification. From a publication by the GOA Project:

> "A GO annotation is a specific association between a GO term identifier and a gene or protein and has a distinct evidence source that supports the association." [94]

Manual annotations created by subject specialists are valued by the biosciences community [93, 95] as compared to annotations generated by computational algorithms. As of July 2010, annotations to human gene products (that is, not including the many other annotations to gene products in other species) totalled over 190,000 separate associations to over 18,000 different gene products [96].

Annotations therefore serve a purpose of storing knowledge about gene products. As indicated, this knowledge is qualified by evidence of different sorts, such as an inference made from a peer-reviewed journal article, or an automatic association created by a computer algorithm and based upon sequence similarities. This knowledge about genes and their products is then put to use in various bioinformatics application, which in the words of the application developers themselves include.

- "...mapping biological knowledge on sets of genes" [97]
- "...annotating full-text articles" [98]
- Being used "...to facilitate data integration" [99]
- Supporting "...knowledge management tasks such as annotation (or indexing) of resources, information retrieval, access to information and mapping across resources." [100]
- Being "...exploited for decision support purposes" [100]
- Forming part of "...a tool [...] that dynamically links gene-expression data to the GO hierarchy" [101]
- Contributing to "...a gene-centric compendium of rich annotative information" [102]
- Annotating "...a gene-to-gene co-citation network" [103]
- "...the prediction of gene function based on patterns of annotation" [104]

- "...to provide a common gateway to access different model organism databases" [105]

Many of these applications are recognisable as areas of academic interest in the information science domain. Knowledge management, information management, data processing, information retrieval, text analysis, decision support systems: these are all traditional topics of long-standing interest to information scientists. Despite the technical language and specialised content of the biosciences domain, the design of the Gene Ontology and its application in the work of the biologists is open to analysis from the research perspective of information science.

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 [65, 76, 106]. 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 exists 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. To return to the research questions posed above, how have the Gene Ontology developers tried to make the vocabulary objective? What are the limitations of the Gene Ontology as a universal classification for gene product functions? And how can the Gene Ontology be improved?

# 2 Methodology

This thesis aims to understand the rules for creating ontologies which are unique to the molecular biology domain. Furthermore, the research is intended to explore how the Gene Ontology developers created their controlled vocabulary.

Much of the existing Gene Ontology research is technical in its application. It is the application of the Gene Ontology to special problems in the biosciences domain, such as how best to index gene products or analyse, by classification, gene expression datasets. Most published research on the Gene Ontology assumes the premises upon which the classification was designed are beyond dispute for the ontology models, in natural language, the very structure of reality. Much ontology research in the biosciences therefore describes the integration of these classifications into species databases, their application in bioinformatics algorithms, or their utility in the analysis of biological data. Very little work exists examining the philosophical principles upon which the Gene Ontology is founded, since to the working biologist such principles as realism and objectivity are rarely contended.

The thesis questions why ontological realism is a core tenet in the design of ontologies for the molecular biology domain. Are there limitations to the Gene Ontology as a universal classification for gene product functions, created by these realist commitments? And what other features of knowledge discovery in biology, what other competing epistemologies might improve the Gene Ontology or inform alternative classifications for gene products?

No research exists considering the historical, cultural and sociological implications of the Gene Ontology project. Some authors have touched upon the technical implications of ontologies for biomedical information systems [61, 100, 107]. Others have concentrated on the philosophical structure of ontologies [106, 108, 109], and their potential role in knowledge discovery in the biology domain [76, 88]. Such qualitative research can provide a natural complement to semi-quantitative techniques in the analysis of the standards and rules governing the Gene Ontology [110-113], or in exploring how biologists use this classification. For example, discourse analysis or content analysis can reveal features of the way biologists classify gene products in their work which are not satisfied by the Gene Ontology.

A mixed methodology can offer a powerful way to approach a research problem from several angles at once, in an effort to provide a fuller picture and a richer explanation behind said research problem. This thesis therefore applies several different methodologies in researching the Gene Ontology as a special classification in a specific domain, that of molecular biology.

In a broad sense this is an investigation of molecular biology as a knowledge domain, as a community of users creating and sharing information under special conditions they have created to suit their purposes. The information the molecular biology domain creates - for example gene sequences, special languages describing molecular functions or complex empirical datasets - can best be understood in terms of the communication chain [114, 115]. Information is created, disseminated, organized, indexed, stored and used by the molecular biology domain. The Gene Ontology is but an element in this communication chain, and the chain is the object of study of the LIS professional.

In a specific sense this research is a study of a single scientific classification within the molecular biology domain. The Gene Ontology acts as a classification system to aid molecular biologists in the organization and indexing of gene product information. In focussing on this single structure in the communication chain, in applying several different, specific research techniques to Gene Ontology data sources, the research is engaging in the process of understanding the molecular biology domain as a whole, in accepting there is an underlying unity to the way molecular biologists use and understand their domain information.

The broad problem of the communication chain molecular biology and the specific task of understanding the role of the Gene Ontology as a special classification within that chain is unified by adopting the overarching research approach known as *domain analysis*.

## 2.1 The research approach: domain analysis

### 2.1.1 The social, the functional, and the realist in domain analysis

According to Birger Hjørland's original vision, domain analysis is intended as a multidisciplinary, multiple methodology approach to information science which aims to understand "...knowledge-domains as thought or discourse communities" [116]. As a research approach, discourse analysis can be understood as being composed of three different elements in that it is part social science, functionalist in perspective, and realist in philosophy.

As a social science, domain analysis approaches information from psychological, sociological, linguistic and epistemological viewpoints which are all primarily social in nature. Information in production and use is grounded in the social, and from a domain analytical approach is primarily understood by accepting the important role social relations play in creating knowledge. Arencibia-Jorge *et al*. [117] use the phrase 'socialized meaning' to describe the social features of knowledge in special communities, and their work, which is a form of domain analysis, uses citation and linguistic analysis to explore conceptualisations of theories and problems in the cancer research domain.

The second theme in domain analysis is that the approach is functionalist in outlook. The role of information in communication and behaviour can be understood in terms of functions and mechanisms, and the domain analyst seeks to reveal how these functions and mechanisms operate within a domain. Jeong and Kim demonstrate the functionalist reasoning in their approach to domain analysis by a semantic analysis of conference proceeding, workshop and seminar abstracts [118]. These types of scholarly events are assumed to serve specific purposes in the biomedical informatics domain, and their purpose within the communication chain would be the dissemination of information to communities of users. To assume a functionalist perspective is to assume that the occurrence of terms within abstracts serve as surrogates for concepts in the minds of users in the domain of biomedical informatics.

Domain analysis is realist in philosophy. It is committed to the doctrine that information science can study variables and factors which inhere in a reality which is external to the subjective and cognitive world of the individual human mind. In this sense, domain analysis offers a counter-point to the 'cognitive turn' in information science [83, 119-122]. Both articles from Arencibia-Jorge *et al*. and Jeong and Kim demonstrate the realist commitment in domain analysis. Citation analysis accepts that when an author makes a citation to another article, that citation is related in a complex yet law-like fashion to social relationships between individuals working on common research problems.

Equally, the analysis of natural language in a corpus composed of abstracts similarly accepts the realist presumption that the words an author uses in describing a work has consistent, objective referents in the minds of thinking experts in the biomedical informatics domain. Realism is important to domain analysis as it grounds the discipline in the study of observable phenomena in reality, rather than lending primacy to subjectivity and the psychologies of knowledge.

### 2.1.2 11 ways to study a domain

Hjørland and Albrechtsen define the meaning of 'domain' in their original 1995 paper [116]. A domain (or knowledge-domain) is a community that works together towards common goals. The community shares a common language, reasons in a shared way, and tackles problems from the same viewpoint. A domain is the manifestation of a discourse in which the people working and talking in that community create a coherent, social group. Defined as such, many communities of information users may be accommodated within the definition of a domain. For example, museum curators [123], graphic designers [124], hobbyist cooks [125] and social workers [126] have all been the subject of domain analysis in an effort to better understand how these communities create, organise and share information.

By their activities as part of a domain, people construct knowledge; this knowledge may be explicit or tacit. Structuralist or cognitivist perspectives on meaning cannot account for everything we know about how individuals understand the world through concepts. Rather, social roles, even within the seemingly objective scientific communities, are vital to understanding how domains interpret reality and synthesise knowledge about that reality.

Theories of knowledge play an important role in domain analysis. For example, scientific communities largely subscribe to the positivist or rationalist epistemologies by which knowledge is pre-instantiated, waiting for the observer to measure, understand and discover. Yet to fully grasp the intricacies of the information environment in a domain, and open these intricacies to empirical analysis, one must accommodate wider epistemologies, such as historicist or pragmatist interpretations of the sources of knowledge. For Hjørland, it boils down to how humans engage with reality which "... cannot be understood naively by the unprepared and isolated subject. It is the knowing subject, who is formed by history and culture, including the concrete development in specific knowledge-domains, who has the possibility to perceive the reality." [116]

Hjørland's domain analysis [116, 127] offers an overarching research approach for my work on the Gene Ontology. The Gene Ontology is one part of the complex information environment in which modern molecular biologists participate in as they tackle research problems in the biosciences. The domain which is the object of my study is therefore the domain of molecular biology. The community of users who make up this domain include not only molecular biologists, but also the bioinformatics specialists, statisticians, computer experts, and allied researchers who all contribute to the study of molecular biology.

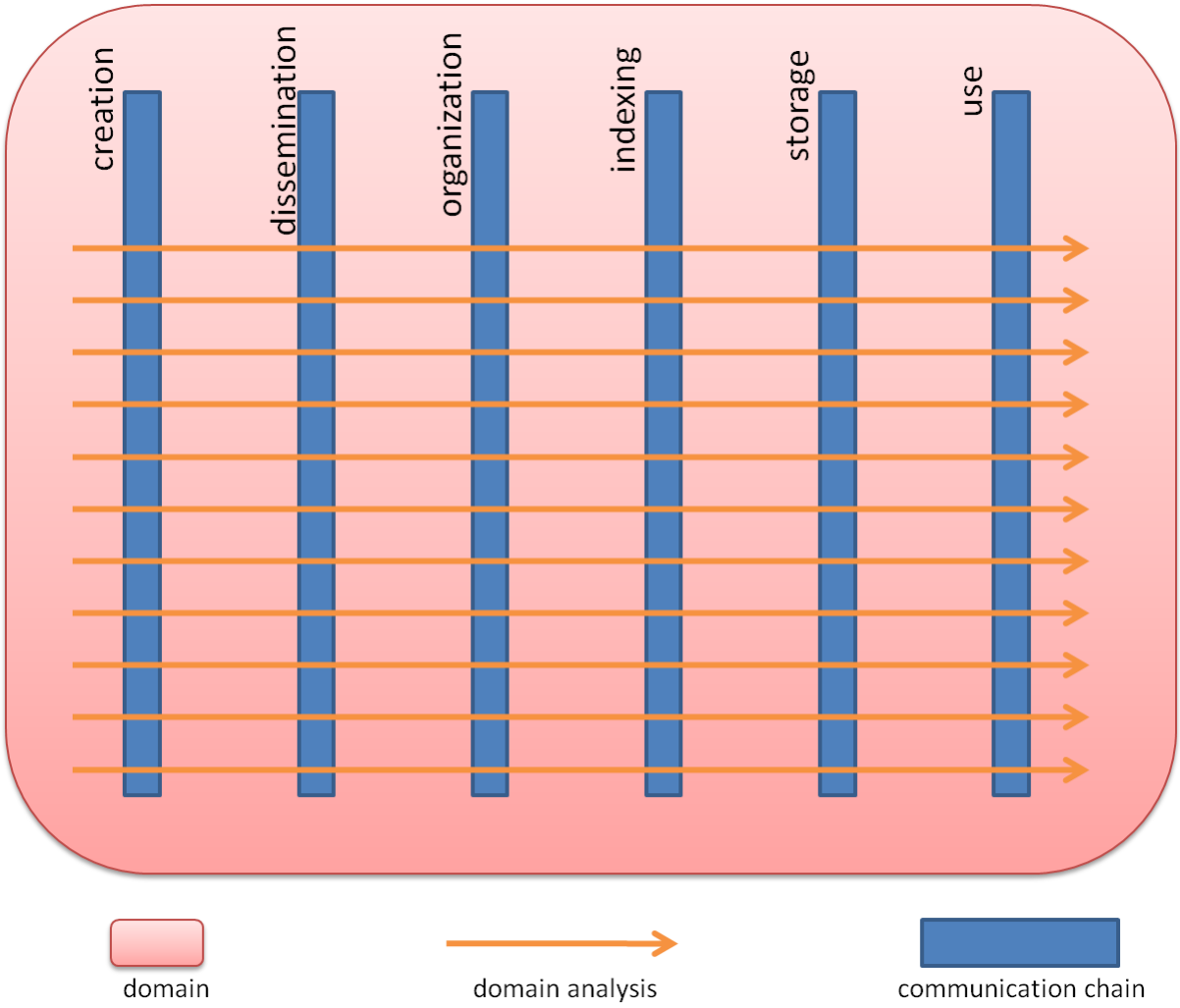
But what is meant by 'molecular biology'? Simply put, molecular biology is the study of the structure and function of biology below the cellular level. The sequence of genes, the synthesis of proteins, the kinetics of biochemical pathway in the cell, the interaction of biomolecules in the living cell, the function of proteins in the life processes of the cell: these are the objects of study for the molecular biologist.

Molecular biology is therefore the domain of interest for this thesis. Domain analysis though incorporates a plurality of research methodologies, any of which can be understood through the lens of domain analysis as tools to research a domain. Hjørland originally outlined eleven different aspects in the study of domains, and these are listed below:

- Production of literature guides and subject gateways
- Production of special classifications and thesauri
- Research on indexing and retrieval in specialist subjects
- Empirical users studies
- Bibliometric studies
- Historical studies
- Studies of documents and 'genres'
- Epistemological and critical studies
- Studies of terminology and special languages, discourse studies
- Studies of structures and organisations in the communication of information
- Studies in cognition, computing and artificial languages

This thesis will principally fall under the second aspect above, research into the production of special classifications and thesauri (of which the Gene Ontology is an example in the molecular biology domain). In the context of the domain and the communication chain, this approach can be summarised schematically as shown in Figure 2.

Figure 2: The domain, the communication chain, and domain analysis




### 2.1.3 Studying the production of special classifications and thesauri

Hjørland indicates that research into classifying special knowledge domains is "…limited in amount and methodology" [127]. In his opinion, information science has concentrated on universal classification schemes rather than specialist schemes, and where research does exist in this area, tends to focus on the creation and maintenance of thesauri using facet analysis or natural language processing techniques [127].

Since both special classifications and thesauri consist of concepts represented by terms and linked using semantic relations, rules and standards for the construction of both structures must share commonalities [128]. This is Hjørland's justification for grouping both entities under a single aspect to domain analysis, and as this thesis will show later, the Gene Ontology is itself a hybrid between a thesaurus of attributes for gene products, and a classification for grouping these attributes into meaningful classes.

As Hjørland emphasises, every database has a special classification and these classifications are frequently developed independently of information science experts. Despite the wealth of research in the LIS domain concentrating on methodologies for bibliographic classification, very little of this research has made itself applicable to specialist domains. The Gene Ontology is therefore an

instance of a special classification and, though it may differ from systems for organising documents, the principles are the same.

The Gene Ontology is used to classify the object of study of the molecular biology domain – gene products. Scientific classifications are governed by underlying theoretical commitments, and Hjørland maintains that to understand these commitments is to appreciate how, with changing theories, come changing systems for classification. Although the Gene Ontology describes a specialist domain and some domain knowledge in molecular biology is necessary to appreciate its intricacies and nuances, as a special classification it still shares many features and attributes common to universal schemes which are familiar to most LIS professionals. Domain analysis is an avenue for integrating research into these theoretical commitments with methodologies for creating classifications to describe this domain-specific knowledge.

Finally Hjørland appeals for a broadness of methodologies in creating special classifications, such as those for scientific domains. Facet analysis, although appealingly rationalist in philosophy, fails to account for the "...social and ideological embeddedness" [116] of classifications. One idea guiding my thesis is that this same limited purview is repeated in the Gene Ontology, and that to accept scientific knowledge as socially and ideologically embedded offers the opportunity to make better classifications.

### 2.1.4 Other approaches to domain analysis to supplement research into special classifications

Hjørland stresses that research into special classifications and thesauri "...can benefit from other approaches to domain analysis", and specifies the following complementary aspects of domain analysis:

- Research on indexing and retrieval in specialist subjects
- Bibliometric studies
- Historical studies
- Epistemological and critical studies
- Studies of terminology and special languages, discourse studies

The focus of my thesis is a study of a special classification in molecular biology domain which, as part of a domain analysis approach, will incorporate supplementary work on indexing, information retrieval, bibliometrics, information history, epistemology and discourse study all centred on better understanding the communication chain in the molecular biology domain.

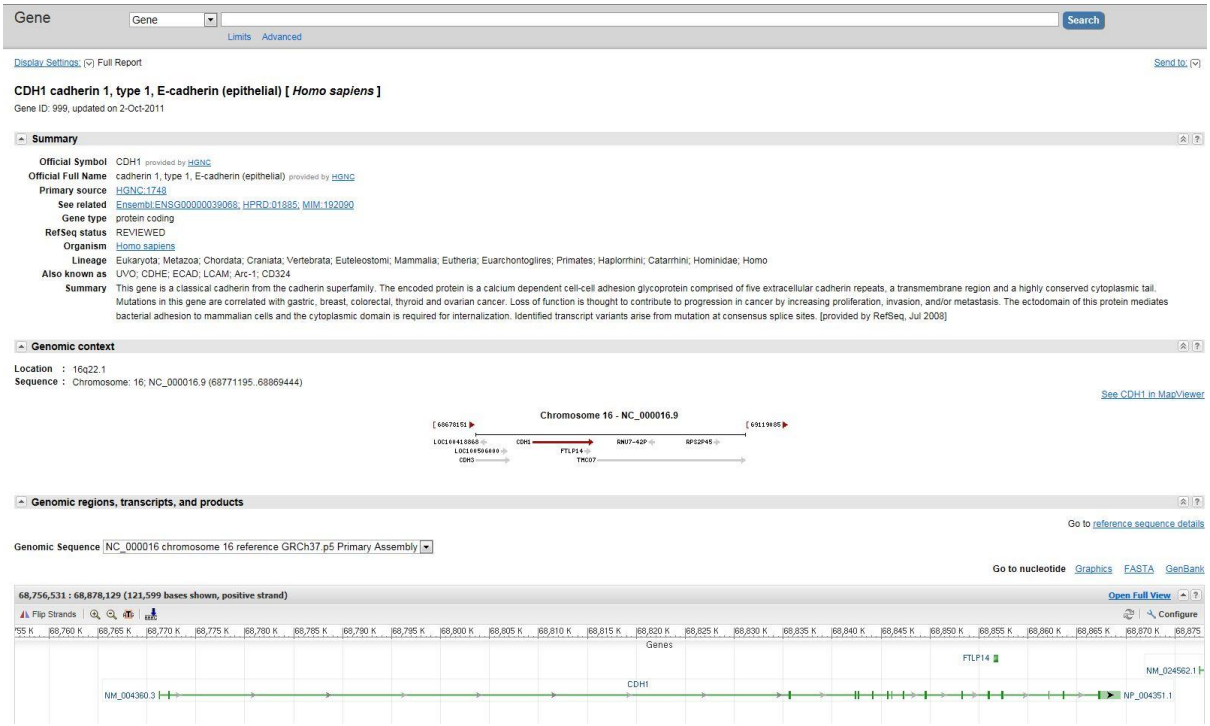
On 'Indexing and retrieving specialities', Hjørland briefly describes the limited nature of research in this area and argues that it is necessary to improve indexing practices, document representation and, consequently, retrieval. I maintain that the objects which Gene Ontology terms are used to index, namely gene product entries in species databases, share many features in common with documents and therefore research on indexing and retrieving bibliographic entities is highly relevant to ontology applications in biology, especially the question of annotation and annotation consistency ('annotation' being Gene Ontology parlance for indexing).

Gene products include entities such as messenger RNA molecules and proteins. Where information about the products of genes is lacking, regions of the genome are marked as putative protein-

encoding sequences. All these entities can be thought of as documents, identified by a source (a species or individual), a location (a genomic position on a chromosome) and content (usually sequence information, in the form of nucleotides or amino acids). Again, the LIS professional looking at the NCBI gene entry for CDH1 in Figure 3 may not be familiar with the language and content of the database entry. However, to a biologist such an entry is understood as a document, and the Gene Ontology is a way to index, represent and retrieve that document from a database. Domain analysis is therefore a LIS research approach which can help in understanding – and perhaps improving – how ontologies are used to index and retrieve gene product 'documents'.

Little work in the LIS domain exists treating biological entities like RNA molecules or proteins as documents. This thesis is a practical demonstration of how the treatment of gene products as documents can prove useful to classification problems in molecular biology.

**Figure 3: NCBI gene view for CHD1 (accessed 07 October 2011)**




In the information science field, the best example of research into retrieval tasks specific to the needs of biologists can be found in the long-standing TREC Genomics Track [129]. Though the focus is on processing texts to extract relevant information to meet various user scenarios, of special interest is the challenges known as 'GO Triage' whereby systems are tasked with identifying suitable papers for Gene Ontology annotation [130]. In comparison to other retrieval tasks, spotting useful papers to submit to GO annotators for review was found to be the most difficult task. No methods were found to improve results over an initial pass of the text corpus for the MeSH heading 'mice' [131], and this hints at the complexity of indexing gene products with GO terminology.

On bibliometrics Hjørland writes that it is a "...strong approach to domain analysis because it is empirical and based on detailed analysis of connections between individual documents". Relatively mature techniques exist for bibliometric studies, and the biosciences domain is rich in documents and metadata to aid the interested bibliometrician. Citation analysis is commonly used in the biology

domain to map the interaction between basic science and industrial applications [132], to indicate the relative impact of important papers in the domain [133], or to appraise the fairness of peer-review procedures [134]. However, despite the strengths of bibliometrics, I will use the techniques in only an elementary manner, to provide a broad overview of ontology-related publication in biology, that I may apply some of the more diverse methodologies proposed by Hjørland.

Hjørland advocates the use of historical studies to evaluate domain-specific classifications. Whilst it is overly ambitious to attempt to synthesise the entirety of the history of science into the domain analysis methodology, the interested information scientist can focus on historical interpretations of the "...development of terminology, categories, literatures, genres, communication systems". Historical studies therefore offer the potential to explore the origins and applications of the Gene Ontology, and to understand the role of ontologies in the biosciences communication chain [115].

On 'Epistemological and critical studies' Hjørland emphasises that in each knowledge-domain there are "...different 'paradigms', 'schools' and approaches" and that their study is fundamental to domain analysis because "...they represent the most general principles and theories that can explain information behaviour" [116]. Hjørland's thinking here is closely allied to the writings of Thomas Kuhn [135], who advocated the importance of sociological behaviours in the development of scientific theories.

The information scientist cannot understand a knowledge domain without understanding the epistemological assumptions underlying the research tradition within that domain. In the case of biology these assumptions would broadly speaking be the empiricist/positivist commitment (that the scientist as an impartial observer can gain objective knowledge about the world) and the rationalist doctrine (that there is indeed a logic and structure to the universe which Man can apprehend, independent of the vagaries of his personal subjectivities). Hjørland stresses that epistemological values determine the "...guidelines for selection, organisation and retrieval of information". My analysis of the Gene Ontology will attempt to understand the epistemological assumptions underlying current thinking in molecular biology, and appraise their role in the development of ontologies in the domain.

## 2.2 Methodology details

Briefly, the research presented in the following chapters applies several different techniques to explore the Gene Ontology and its design. Firstly, the idea of ontological realism, by which ontologies are taken to describe the relationships between entities in reality rather than concepts in the minds of users, is challenged by an analysis of a single concept in the Gene Ontology, and by attempting to show how this concept can be situated historically and pragmatically in the ongoing work of biologists.

Furthermore, a detailed investigation of the rules governing the structure and editing of the Gene Ontology in comparison to international standards in vocabulary construction is used to try and explain why the GO was specially created for the molecular biology domain. In the next chapter, a discourse analysis of several texts from the GO mailing lists further extends the findings on term obsoletion, in an attempt to explain how social roles and power affect decision-making in the Gene Ontology Consortium. This leads into a study of term obsoletion in the Gene Ontology which audits the reasons why terms are removed from the ontology.

Finally, a content analysis of a corpus of recent Gene Ontology publications explores how users in the molecular biology themselves report usage of the Gene Ontology, and whether this usage corresponds to the Consortium's own image of how ontologies ought to be applied.

Efforts were made to survey the personal perspectives of GO editors and annotators on their work via a questionnaire emailed to twelve such persons in the UK and USA. Participants were selected based on their current involvement in the Gene Ontology project, as determined by participation in recent (posting in 2009) ontology development discussions on Sourceforge. The questionnaire comprised eight general, open-ended questions designed to elucidate opinions and views about their ontology work. Only two recipients responded, and preferred to be interviewed rather than to write questionnaire responses. Semi-structured interviews were therefore carried out with these editors, using the same questions as in the original questionnaire (see Section 6.1 for further details). Despite the poor response, the interview results provided valuable context for the rest of the thesis analyses.

These studies will inform Chapter 4 in which the implications of Gene Ontology philosophy of biological functions are appraised, and suggestions as to how alternative classifications might create better vocabularies for indexing gene products in the domain are made.

### 2.2.1 Concept analysis

Different epistemological worldviews create different definitions for what a concept is. Hjørland outlines several of these theories of concepts [136], and argues that knowledge organization systems will differ depending on the concept theory a system designer can commit to. In Chapter 3.2 a single concept in the Gene Ontology is analysed according to four concept theories: empiricism, rationalism, historicism and pragmatism. Although ontological realism is the declared methodology, or doctrine, guiding the construction of most ontologies in biology, including the Gene Ontology, concept analysis can test the hypothesis that concepts in the biosciences domain are in fact not stable, context-free, nor value-free.

Various methodologies drawing on description logic, formal concept analysis (FCA) and formal ontological principles exist for analysis concepts in the biomedical domain [66, 113, 137-141]. Many of these methodologies rely on computer algorithms and natural language processing to automatically assess, according to predetermined rules, the semantic consistency of vocabulary terms. Jiang and Chute [142] describe a method for constructing concepts lattices according to FCA to represent concepts in the SNOMED CT, whilst at the same time assessing the completeness of the vocabulary. Similarly, Jiang *et al*. extend this approach to automated ontology construction using FCA to identify concepts from medical text corpuses [143].

Science mapping similarly relies on mathematical analyses of words and their semantics to classify domain concepts. Techniques such as cluster analysis can be used to group different words from different disciplines into common classes with shared meanings [144]. Yet these approaches do not account for how different models for what concepts actually are impact on conclusions from this type of research. For example, a science map which represents concepts and their inter-relations may on the one hand be considered as a visualization of the experimental approaches and their attendant theoretical language if one subscribes to the empiricist model of concepts. Or should one assume a rationalist approach to concepts, the same science map might be taken as a sketch of the structure of reality itself, drawn from science's surrogates for groups of instances in this reality,.

Concept analysis as used in this research is therefore not a mathematical analysis of the semantic structure of the language of molecular biology. It aims to show that different epistemological model for concepts, each being valid in different ways, can show the meaning of a single GO term to be plastic. The meaning of a concept, even a scientific concept, changes according to the concept theory one chooses to adopt, the scientific paradigm within which that term is used, and the intended research goals of a scientific community which uses that term. Concept plasticity for a Gene Ontology term undermines the doctrine of ontological realism as the principle philosophical method for guiding ontology construction in molecular biology.

Such a falsification of ontological realism using concept analysis is evidence that other philosophical bases for vocabulary construction in the sciences (such as pluralism or cognitivism) can be incorporated into ontology design to create more better classifications for biology.

In the first instance, a GO term, 'cardiac cell differentiation' has been selected. This selection is not unbiased, because of personal knowledge of cardiovascular biology. The term resides in the biological process ontology, and therefore other terms from the Molecular Function Ontology or Cell Component Ontology are not analysed.

There exist very few declared methodologies for concept analysis. The approach here is guided by Birger Hjørland's paper [136], and is inductive in the sense that various text were read iteratively to finalise the research methodology which is both transparent and reproducible. The broad approach is outlined in Figure 4 below.

**Figure 4: Concept analysis**

1. Using a selected GO term, search all other current vocabularies in the biosciences for equivalent or related terms; compare definitions
2. Analyse position of selected GO term within the GO hierarchy, how it was created, and what justifications exist for these relationships
3. Decompose selected GO term into facets, and repeat '1' using each facet
4. Extend search to the biosciences literature, looking for instances in which chosen concept and its constituent facets are used; read and compare these meanings and contexts to the stated GO definition
5. Search through biology textbooks for canonical uses of my selected GO term, including index and glossary terms
6. Using a historical approach, describe how selected GO term has been realised through the development of different theories in biology
7. Using this data, frame selected GO term in the context of the theories of concepts established on empiricism, rationalism, historicism and pragmatism
8. Test whether selected GO term can be insulated from the criticism of context-dependency
9. Suggest terms, synonyms or definitions related to selected GO term, but not excluded due to Gene Ontology design rules

### 2.2.2 Analysis of Gene Ontology vocabulary construction standards

Information history is a synthesis of historical study with the research questions central to the LIS discipline: how is information used different domains, how do people understand information, or how information changes society [145-148]. As Weller explains, information history is more than the history of the library for it seeks to understand "...the way in which information is or has been

thought of and applied in its own right" and the interplay of information with "...the social, political, economic and cultural climates of the time" [145]. Furthermore, the post-modernist take on information history can explore the role played by information in "...the everyday experience and attitudes of people" [145].

Information history takes the historical perspective, such as that adopted in economic history, social history, or cultural history, and applies it to the different elements in the communication chain. As Robinson argues, the communication chain (composed of the creation, dissemination, organization, storage, indexing and use of information) provides a conceptual framework formalizing the study of information in different domains [115]. Therefore this thesis takes a historical approach to understanding the origins and development of the Gene Ontology classification scheme as it stands as a tool for indexing gene products in the biosciences domain.

Eisenstein wrote of the role of the printing press in the development of modern society [149], and likewise one can consider the important role played by classifications in the development of modern biology. Every database requires metadata standards to organise the data it stores, and this data must be classified. The information history approach explores the recent history of special classifications created for the special purpose of indexing gene product records in computer databases. In looking at the historical precursors to the Gene Ontology and the development of the ontology standards in the biosciences, can it be better understood why ontologies have seen rapid adoption in the domain?

Arguably this strategy is more an exploration of the history of information in the molecular biology domain, and the shape of that recent history in the advent of the 'data deluge' in the sciences [99, 150, 151]. Few studies have made attempts at placing the information behaviour of biologists in a historical context. Some research exists mentioning Charles Darwin's usage of information in his researches [152], although advocates of information history argue that this example fails to place Darwin's information work in the wider historical context of Victorian England [145].

The approach in Chapter 3.3 compares the Gene Ontology editorial rules against the ANSI/NISO standard Z39.19, 'Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies'. Where the Gene Ontology diverges from this international standard for controlled vocabularies, evidence drawn from reports, papers, meeting minutes and message board discussions is used to explain how the GO Consortium developed its own set of ontology standards.

The major themes and conflicts in the design of the Gene Ontology are articulated, and placed in the historical context of molecular biology and the wider economic environment in which science, and especially this kind of e-science, is funded in the 21st century [108, 150, 151, 153-163]. These issues are considered in the light of the efforts towards 'open science' and high-throughput, collaborative research in [154, 160, 164-166] biology, married to the principles of the Opensource movement which guide much of the software development in the bioinformatics community.

### 2.2.3 Discourse analysis

A variety of methods and approaches to discourse analysis exist, but Chapter 3.4 uses what is commonly known as Critical Discourse Analysis (CDA) in a reading of Gene Ontology texts. Discourse is the manifestation, in the text, speech or action of everyday life, of signs. Fairclough describes CDA as the "...analysis of the dialectical relationships between semiosis (including language) and other

elements of social practices" [167]. Dialectics is the interplay or interrelationship between aspects or ideas, which in the case of CDA is the study of how signs used by humans communicate social roles, the exercise of power by authorities, and notions of truth.

This thesis looks at texts created by various authors during the course of development discussions on a Gene Ontology mailing list. The discourse therefore consists of the messages sent and received over time. What these words mean in the context of the Gene Ontology project, of problems in the domain of molecular biology, and in relation to the social status and roles of the speakers in the scientific community, are analysed by a variation of CDA described by Norman Fairclough. The critical approach of CDA is intended to dissect the role of power and authority in deciding the contents of the Gene Ontology – where disputes exist, how are they resolved and what are the consequences for the ontology?

Email mailing lists are examples of new media communication [168] and Gruber characterises email as communicative forms with the characteristics of being both conceptually written and spoken, primarily textual (with the possibility of hypertext), dialogical, both 1:1 or 1:*n* in terms of communication partners, of medium intended persistence, and asynchronous in delivery (versus instantaneous for an alternative form like a chat room).

The GO mailing list corresponds with Gruber's characterisation. Messages can be either conceptually written, such as short arguments justifying the inclusion of a new term, or conceptually spoken, in which correspondents are speaking to one another, in written words which read much like speech. The communication is text, with some hyperlinks to other Web documents, and most communication takes the form of a dialogue between the mailing list members. Some mailing messages do take the form of announcements, which may not illicit a response and could be described non-dialogical. So too do the number of communication partners vary, although since any list member may reply to any message, the majority of new messages are one-to-many. The GO mailing list persists today in an openly accessibly archive [169]. All threads are time-stamped and clearly show that communications are asynchronous, with a considerable delay between responses explicable by list members corresponding from different time-zones (primarily between the US and Europe).

Several authors have conducted discourse analyses using computer-mediated communication (CMC) texts such as emails, chat rooms and mailing lists. Chiluwa investigated the discourse of spam email messages which use various strategies to deceive readers and engender trust [170]. Park studied online conventions in chat rooms, and in particular strategies such as 'emoticons' which speakers use to compensate for non-verbal cues that people would normally use to communicate emotions in a polite conversation [171].

Several ethical considerations must be made when considering mailing list as text sources for critical analysis.

Firstly, should one anonymize the source of the data? The GO mailing list is a publically accessible list and, as Herring maintains, one may treat the messages as public broadcasts even though the discussions themselves may be between individuals [172]. In this discourse analysis the name of the group which is the source of the text is not anonymized, but after the manner of Herring, the anonymity of individual list members is preserved. Though this may seem contradictory (the names

of sources can be identified with relative ease), the anonymization of individual's names is an effort in linguistic research to maintain a distance between observer and observed. The advantages are two-fold: objectivity is fostered in the text analysis and the privacy of correspondents is, to a degree, respected.

Secondly, should the researcher seek informed consent to use mailing list communications as a data source? Electronic communications such as mailing list messages are copyright protected, with reproduction rights accorded to the authors. Herring argues that whilst messages ought not to be modified when used in academic research, seeking consent from authors to use what they have written does restrict the ability of the researcher to appraise the data critically. The practicalities of obtaining permission from every member of a mailing list and the dubious legitimacy of a list-owner granting permission on behalf of the other list members notwithstanding, GO mailing lists stand as public discourse. Public mailing lists are like public broadcasts, and one would not seek the permission of a broadcaster and interviewee in order to quote communications from a radio show.

Therefore no informed consent was sought from mailing list participants. The identity of speakers is anonymized, both for their own privacy and to facilitate research objectivity. The ability to appraise communications critically would have been severely impinged in seeking the consent of individual speakers to use the text of particular conversations. Although anonymizing messages does modify the original text, this is considered as fair usage within copyright law.

Fairclough distinguishes between three types of value that text features may have [173].

- Experiential values are to do with "...*contents* and knowledge and beliefs" and suggest how "...the text producer's experience of the natural or social world are represented."
- Relational values are to do with "...*relations* and social relationships" and indicate "...the social relationships which are enacted via the text in the discourse".
- Expressive values are to do with "...subjects and social identities" and suggest how the text producer may evaluate "...the bit of reality [the text] relates to".

Fairclough goes on to describe a simple framework for describing texts that accommodate these three values (see Table 5). 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. Fairclough's method provides a means to describe texts to facilitate the second phase of analysis, which is the interpretation of the context and role of language. Through this interpretation, the analyst can gain an insight into the social commitments and ideologies at play within the discourse.

Table 5: Fairclough's framework for discourse analysis

| Linguistic aspect | Values to consider |
|---|---|
| **Vocabulary** | 1. What experiential values do words have?<br>a. Classification schemes used by the text producer<br>b. Whether words are ideologically contested<br>c. Rewording or over wording suggests some contestation over reality<br>d. Synonymy, hyponymy and antonymy suggest ideology, either in discourse type or creatively established within a text<br>2. What relational values do words have?<br>3. What expressive values do words have?<br>4. What metaphors are used? |
| **Grammar** | 1. What experiential values do grammatical features have?<br>2. What relational values do grammatical features have?<br>3. What expressive values do grammatical features have?<br>4. How are (simple) sentences linked together? |
| **Textual structure** | 1. What interactional conventions are used?<br>2. What larger-scale structures do the text have? |

The micro-level discourse analysis in this research is based on the Fairclough framework above. Detailed analyses are complemented by macro-level reading through all GO mailing list messages between January 1999 and February 2002. In February 2002, the GO project initiate a Sourceforge ontology request tracker which took over the purpose of managing debates over detailed changes to the terms in the ontology, and therefore this date was used to truncate the analysis of the mailing list discourse.

In tandem with Fairclough's three-dimensional framework above, an adapted 'Linguistic checklist' suggested by Fowler [174] is employed. This checklist (see Table 6) was 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 Gene Ontology 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.

Mailing list message texts were processed to ensure a degree of anonymization, and to aid in reading and presentation. Further methodological details are given in Chapter 3.4 , including how texts for analysis were selected.

**Table 6: Fowler's linguistic checklist for analysis power in discourse**

| Language aspect | Approach |
|---|---|
| **Lexical processes** | "...vocabulary reflects and expresses the interests of the group", therefore texts are read for concepts which are used repeatedly, and which demonstrate either overlexicalization (where many words exist corresponding to the same concept, such as technical jargon or slang) or underlexicalization (where the lack of a word for an important concept forces a circumlocution); an example of a lexical process would be the many different words we use for 'wife' in Western society |
| **Transitivity** | After the work of Halliday and Fillmore, transitivity deals with processes designated by verbs and adjectives, and the entities, usually in the form of nouns, participating in these processes; in power terms, transitivity can show what relations are at play within a discourse, and the differences between power agencies; one reads the kinds of roles attributed to agencies by the predicate they are linked to, so for example an article about a new government policy to improve failing schools may be attributed to a Minster, whereas the agent responsible for the failure of the school is consistently omitted |
| **Syntax** | The choice of syntactic phrasing can obscure agency and act to insulate power players from critique; for example, the phrase 'The proposals going forward are not fixed in stone' tends to obfuscate on what the proposals are, who is making the proposals, what the object of 'going' is and what the changes implied by the choice of the phrase 'not fixed in stone' might be. We might construe these details from the context in which this statement is used, but one could imagine different syntactic paraphrases in which it was made abundantly clear what the potential changes in future might be. Syntactic choices include<br>1. Deletion: parts of a sentence construction are left out, through ellipsis (in which meaning can only be inferred from previous sentences), nominalization (in which a verb is transformed into a noun, such as 'You propose' is changed to 'The proposals') and passive tense (in which agency is deleted, such as 'Change was proposed')<br>2. Sequencing: passive voice permits the reordering of elements in a sentence to allow prominence of different parts. An example might be 'I proposed closing the library' altered to read 'Library closure was proposed by me' de-emphasises myself as the agent and re-configures the sentence to put 'library closure' to the for<br>3. Complexity: Described as the length of sentences, and the way in which parts of the sentence are coordinated or subordinated, greater complexity in linguistic structure correlates to the exercise of power. In scientific language, there can often be many clauses in a single sentence which modify the logical content of a statement. For example 'The biochemical characterization of lysosomes has so far depended on purification methods based on either density gradient centrifugations or magnetic purification of iron-loaded organelles' has several qualifiers and dependencies which marginalises the non-expert. |
| **Modality** | Simply put, modality refers to validity, predictability and desirability with forms such as may, must, need, probably and perhaps. An authority will exercise its power through these kinds of modalities to approve or resist ideas. Uncertainty and hesitancy in language suggests acts of deference, and these kinds of modal adjustments can indicate when a speaker is subservient. |
| **Speech acts** | Utterances, be they single words or groups of sentences, are intended to produce some effect or achieve a goal, depending on the context in which they are written or spoken. Speech acts therefore maintain social roles, and depend for their success on the speakers recognising the parts they play in a power relation. |
| **Implicature** | Introduced by Grice, implicature refers to non-stated meaning inferred by what is deliberately omitted from a conversation. Rights are presumed to exist as to who possesses status and authority enough to permit implicature, and these unspoken commitments can themselves add up to a tacit ideology which undergirds a discourse. For example a domestic argument between a husband and wife about cleaning a house could be interpreted in a different light if one took account of the fact that the husband had been unfaithful. Admissions of guilt or accusations of unacceptable behaviour become readable if one accounts for implicature and the refusal by both parties to talk about the silenced issue. |
| **Turn taking** | Power relations between the participants in a conversation can determine who may initiate or close a conversation, interrupt or keep talking. For example, in the classroom there are clear rules which govern turn taking between the pupils and a teacher leading the class. |
| **Addressing** | Dependent on the dimensions of power and solidarity, addressing refers to choices over pronouns and forms of address between parties. Persons of unequal standing, such as an elder and a young person, or colleagues with the same status in a job, will choose appropriate forms for addressing one another, and an analysis of address forms in a discourse can reveal much about the distribution of power. |

### 2.2.4 Content analysis

Content analysis is the qualitative coding and interpretation of data, usually in the form of text, to test a hypothesis. The process of coding reduces large quantities of text into a smaller number of categories, and instances in these categories are used as a measure of some variable. For example, the number of times a person says 'Um' during the course of a conversation may be taken as a measure of uncertainty in the mind of the individual.

Content analysis aims at a scientific approach to the study of content, in that the standards of the scientific method are adhered to. A scientific method is one which tests a hypothesis using an approach which is objective, reliable and reproducible, and yields results which can be generalized beyond the immediate confines of the study-setup [175]. A content analysis contaminated with the personal prejudices of the analyst, which fails to articulate a method that can ever yield reproducible results, or which has no clear hypothesis to test does not meet the research standards of a good content analysis.

Content analysis is a used in the LIS domain, and is applied in a variety of scenarios for hypothesis testing. In terms of research into classifications and their usage, content analysis can be a useful tool in interpreting the structure and usage of various kinds of controlled vocabularies. Jin provides an overview of attempts to translate Library of Congress subject headings into FAST, or Faceted Application of Subject Terminology, an alternative subject access system amenable to deployment in digital environments [176]. In this kind of work, subject headings are hypothesised to be decomposable into different facets, and content analysis takes the vocabulary itself as the object of analysis. Spiteri reviews the literature on efforts to use facetted classifications to organise socially-generated tag data [177] and these applications are all forms of content analysis. In this example, social tags are analysed to test the hypothesis that some tags are more informative than others, and these tags represents facets. The analysed data is lists of tags, the objects they describe, and their relationship with existing vocabularies.

Content analysis may be extended from studying controlled vocabularies to larger text corpuses. The expansion in volume of data for analysis can make computer-supported data mining techniques invaluable, and Leroy *et al.* [178] describe an interesting technique for automatically appraising the readability of texts. In this example, the results are validated against test subjects who themselves rated the readability of sample texts, and here is we can see the importance of the scientific method in supporting the evidence from a content analysis. Likewise Deokattey *et al*. [179] analyse texts, in this case abstracts, to identify plausible candidate concepts for a domain ontology representing nuclear physics technology. Computers are used to initially process the texts, whilst experts then analysed the output content to develop facets.

Content analysis can also be used to study existing datasets in new ways. Nisonger takes citation information, usually studied using bibliometric analysis, from two popular LIS database services and appraises their coverage compared to a checklist [180]. This method provides a measure for the quality of the two databases, based on the hypothesis that the better service will cover more items on the checklist. Nisonger's analysis demonstrates that even databases used to collate information can become the object of study, and his methodology is clearly reproducible and applicable to other database services.

This thesis applies content analysis to two different datasets. Chapter 3.5 is an investigation of the reasons behind why terms are removed or 'obsoleted' from the Gene ontology. This analysis uses GO database fields and free text comments from term-related Sourceforge discussion threads to ascertain why the GO Consortium decided specific terms are redundant. The null hypothesis is that terms are only ever obsolete from the ontology because GO editors and users agree that no instances exists in reality. This analysis of term removal from the ontology is in the context of the Gene Ontology's own rules and standards for vocabulary construction, which are themselves grounded in ontological realism, or the idea that only terms with instances in reality may be included in an ontology for biology.

The second application of content analysis in Chapter 3.6 uses a subset of full text Gene Ontology papers from a bibliometric dataset (see Appendix 6.7 ). Metadata from bibliographic records for each article are combined with categorisation of statements from the full text articles which describe results of a common type of Gene Ontology analysis called 'GO term enrichment'. The authors' usage of Gene Ontology files and GO terms is coded in order to test the hypothesis that user definitions for entities (like functions and processes) in the molecular biology domain do not entirely correspond with the representation of knowledge in the ontology. Divergence from GO definitions are measurable by author rewording of GO terms and restructuring of ontology classes and relationships as biologists try to compose meaningful categories that convince their target domain audience of the validity of their research.

Gene Ontology papers for the period 2000 through 2009 were retrieved from MEDLINE and Web of Science, identifying 2101 relevant papers. Papers without matching PubMed identification numbers from both searches were excluded because these hits had no associated MeSH headings, an important metadata field which was a dependency for other analyses. This reduced the paper set to 1935 hits (for further details, see Appendix 6.7 ).

All papers published in 2009 were retrieved, producing a set of 374 papers. Only papers which were cited at least once, regardless of whether this was self-citation, were retrieved from this set, leaving 163 papers for initial content analysis.

A first pass content analysis searched for Gene Ontology papers which did not deal with analytical papers. Non-analytical papers were considered to be reviews about the GO project, biological databases which incorporated GO data, mathematical models dependent on GO data, prediction algorithms using GO, and reports of software drawing on GO data sources. The final set of 113 papers for content analysis in Chapter 3.6 are therefore exclusively applications of biologists using GO to make inferences from empirical data, and published in 2009.

## 2.3 Data sources

This thesis relies on a diverse range of data sources as evidence for the design principles of the Gene Ontology and the working practices of the Gene Ontology Consortium. These data sources include the following:

- Gene Ontology data files (including archived versions) [181]
- Gene Ontology annotation files [96]
- Gene Ontology protocols, meeting minutes and published papers [182]
- Archived versions of the Gene Ontology website [183]
- Sourceforge Gene Ontology development trackers [184]
- Bibliometric dataset of Gene Ontology-related papers (see Appendix 6.7 )
- Alternative classification schemes for the biosciences (eg, MeSH)
- Open Biomedical Ontology files [85]
- Historical articles, textbooks and indexes referring to specific concepts in biology
- Interview data with Gene Ontology annotators (see Appendix 6.1 )

# 3 Results

## 3.1 Results summary

The following section presents the results from five different techniques applied in exploring the development of the Gene Ontology. In order to aid navigation through these results, a brief summary of the major findings is included below.

### 3.1.1 Concept analysis summary

A concept analysis shows that a single concept in the Gene Ontology, 'cardiac cell differentiation', can be presented in different ways according to different epistemologies for concepts. The Gene Ontology project is broadly committed to ontological realism which is closely allied with a rationalist theory for concepts, yet other epistemologies can be drawn upon to understand a single GO term, a process which potentially alters this term's relationships and definition within the vocabulary. Accepting alternative theories for concepts or scientific paradigms in understanding scientific concepts does not necessarily undermine their objective basis and application. This concept analysis is strong evidence that since there is more than one way to understand a single concept in the GO vocabulary, then potentially all the terms and their relationships in the ontology could be altered to reflect alternative, yet equally valid epistemologies for concepts. Consequently, these results act to undermine the philosophical basis of ontological realism as a core, irrefutable standard for creating ontologies in the biosciences.

### 3.1.2 Gene Ontology vocabulary standards summary

An analysis of the standards and rules used to construct the Gene Ontology compares GO internal guidelines to an international standard in controlled vocabulary development. Whilst the GO Consortium, without direct reference to existing standards, has in some aspects successfully devised appropriate rules for selecting terms and encoding complex relationships between these terms, major weaknesses are identified in this analysis. In particular, synonym control is very poor and with no identifiable warrant invoked regarding preferred terms and their relationships, the Gene Ontology is logically consistent according to the precepts of formal ontology, but models a very limited representation of knowledge in molecular biology. Results indicate 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 developers also risk exacerbating the linguistic problems they are trying to solve in the domain, by inventing non-canonical names for ontology nodes.

### 3.1.3 Discourse analysis

In an effort to better understand the decision-making processes behind the development of the Gene Ontology, discourse analysis is used to study developer and user conversations drawn from a popular GO mailing list. Texts indicate that social relations do play an important role in deciding ontology content. Despite aspirations of scientific objectivity, consensus on the inclusion of concepts in the vocabulary, and authoring of universal definitions for fundamental ideas such as sexual reproduction, is demonstrably difficult to achieve between different Gene Ontology contributors. Examples are found from GO mailing list conversations where senior ontology editors decide ontology content through the exercise of their power and authority, marginalising alternative conceptual views in the process.

### 3.1.4 Term obsoletion

GO terms are routinely removed from the vocabulary for through a procedure the GO Consortium calls 'term obsoletion'. In an analysis of the reasons why several hundred terms were obsoleted from the Molecular Function Ontology, results show that classes of justifications for the removal of certain kinds of terms recur over the history of the ontology. These terms are arguably valid concepts, and have common usage in the molecular biology domain. Reasons for their obsoletion are contrived by the GO Consortium in accordance with certain assumptions unique to ontological realism.

### 3.1.5 Content analysis

In the final section of the results, a content analysis of a set of Gene Ontology analysis papers published in 2009 gives an indication of how users in the biosciences community report usage of GO terminology. Results show poor compliance with GO data citation policy, weak transparency in the way GO analysis results are generated and frequent re-wording and erroneous quotation of GO terminology.

## 3.2 Concept analysis

In this section of the results, concept analysis is used to explore the origin and representation of a single term in the Gene Ontology. In so doing, alternative epistemologies to ontological realism are presented as a basis for classifications of gene products, and the potential advantages of these other 'theories for concepts' are presented.

### 3.2.1 Cognitive and social models for knowledge (and concepts)

#### *3.2.1.1 On the purpose of this concept analysis*

The Gene Ontology, as a subscriber to the OBO Foundry principles, adheres to ontological realism. GO represents mind-independent entities and not mental concepts. The ontology is a knowledge representation of molecular biology where every GO term is a placeholder for entities and processes in reality. This principle is sometimes discussed in ontology circles as 'instantiation' or 'instantiability' – every term in an ontology must have at least one instance in reality. The Gene Ontology resists the contention that GO terms are placeholders for mental concepts in the minds of biologists. Biological ontologies are not representations of scientific discourse or the collective belief states of biologists.

Later sections will discuss the implications of the Gene Ontology's commitment to ontological realism for the design and implementation of the vocabulary. Before exploring these standards for the construction of the GO vocabulary, an argument *for* concepts as the basis for biological ontologies is presented. The aim is to understand why ontological realism is core tenet in ontologies for the molecular biology domain. In this process, this chapter will demonstrate that in fact there are several different theoretical models for concepts which could equally serve as practical foundations for scientific classifications of gene products. The argument rests on the hypothesis that knowledge representations of the mental concepts and subjective understandings biologist's possess regarding processes, functions and entities in molecular biology are useful for indexing gene products, in terms of potentially improving information retrieval, supporting better information management strategies for biological data and using computers to develop new ideas and theories.

The chapter presents a concept analysis using several different theories for concepts, drawn from a review by Birger Hjørland. A single term from the Gene Ontology is presented in relation to its structure and relations according to the Gene Ontology philosophy, and according to ontological realism. This term is then de-constructed in order to understand it in alternative informative ways depending on the concept theory drawn upon. The chapter concludes by arguing that since a single Gene Ontology term can be framed in the context of a number of concept theories, then all the terms in the ontology can potentially be framed in this manner. The consequence of adopting a different philosophical doctrine for the meaning of Gene Ontology terms is that we can design alternative controlled vocabularies for indexing gene products which expressly aim to model concepts in biology, and ignore the principle of instantiability.

#### *3.2.1.2 Concepts for concepts*

What then is a concept? The word 'concept' is frequently used in the biological ontology literature, yet is rarely defined [36, 74, 137, 138, 142, 185-187]. Little systematic work exists attempting to understand the basis for concepts, or in describing potential concepts for concepts.

Birger Hjørland writes, in one of few review papers dealing with theories for concepts, the following description:

> "Concepts are dynamically constructed and collectively negotiated meanings that classify the world according to interests and theories. Concepts and their development cannot be understood in isolation from the interests and theories that motivated their construction, and, in general, we should expect competing conceptions and concepts to be at play in all domains at all times." [136]

Hjørland [136] considers, and I agree, that concepts are drawn from paradigms in the sense understood by Thomas Kuhn [135]. A paradigm is a common way of approaching theoretical problems in a scientific domain. The paradigm is constructed in the sense that the scientific problems, the tools available to tackle those problems, and the way scientists work all combine to limit the ways science can be done. If we say that concepts are drawn from paradigms, we are saying that concepts are socially and historically situated. Concepts are a synthesis of the knowledge and practices of scientific communities. The meaning of concepts is negotiated in the sense that scientific communities must come to a consensus if they are to productively tackle scientific problems. Drastic, sudden changes in scientific knowledge occur when the weight of evidence supporting one paradigm and one set of shared concepts is finally undermined by a new paradigm and by a new way of looking at old scientific problems. The shift to a heliocentric model for the Solar System is one example of a paradigm shift. The acceptance of cell theory or the theory of evolution is an example of paradigm change in the biological domain.

Some concept theories do not insist concepts are socially situated. The classical view of a concept considers a concept to be a set possessing necessary and sufficient properties to determine membership by an individual. On first reading, this understanding of a concept seems to be eminently sensible and logical. Based on the Platonic view of forms and ideals, the concept of a table is grounded in an ideal table, which specific properties such as legs and a supporting surface which combine to create in our minds the concept 'table'. The classical view of concepts though quickly founders on more complex concepts such as 'art' – what necessary and sufficient properties determine what art is? In biology, a very simple example of the failure of the classical view of concepts is in the concept 'alive' – what constitutes a living thing? Viral particles only seem to be alive when they are inside another living organism, and this discovery created and still stimulates lively debate in the biological sciences. A virus calls into question the necessary and sufficient properties governing membership of the set of things we call 'alive'. What has been called the 'spreadsheet approach' to concepts, where individuals are rows and columns represent properties, flounders when confronted by concepts like 'art' and 'life'.

Probabilistic theories of concepts consider some exemplars to be better concept representations than others, and draw on the prototype theories of ideas. We can point at an object and call it a 'tree', and then look at all the other objects growing in an area and decide how much like our prototypical tree these other entities are. Probabilistic theories do not rely on social theory to explain the provenance of concepts. An example in biology of an attempt at a probabilistic theory applied to a type of concepts in the phenetic system for classifying organisms. Phenetics attempts to quantify the observable traits in different species to create a numerical method for measuring similarity. Largely superseded by cladistic methods, phenetics was an attempt at making taxonomies more objective, an attempt which failed because the selection of characteristics is an inherently subjective procedure. As we shall see in the case of the Gene Ontology, it is difficult to eliminate the subjective even from seemingly scientific classifications.

Another non-social model for concepts is conceptual atomism, whereby a concept is not a definition, but has a causal relation to the structure of reality. The structure a tree has in reality is mirrored by semantic structures deployed in human language and there is a direct correspondence between structures in reality and language structures in our minds. Quite how conceptual atomism can be resolved with the process of science though, in which new and ideas and new concepts are invented and tested, is difficult to see. Hypothetical or false concepts which have no correspondents in reality regularly make a contribution to the discovery of new ideas in science. Conceptual atomism cannot provide a complete concept model for these types of circumstance.

#### *3.2.1.3 Theory theory of concepts*

After the manner of Hjørland, this thesis advocates what is dubbed the *theory theory of concepts*.

Scientific theories, theoretical frameworks, and the language of observation statements create concepts, and define their limits and applications. The rest of this chapter focuses on scientific concepts and concepts used in scientific explanations, although much of what follows is applicable to everyday or commonsense concepts.

According to the 'theory theory of concepts', scientists share common theories and common understandings about how the world works. Kuhnian paradigms require this commonality, since it annuls the problem of different scientific groups presenting logically incompatible, conflicting theories. Scientists form communities, share theories and it is these theories which shape and define conceptual language in the domain.

Within the 'theory theory of concepts' are four epistemologies which create four different models for concepts.

Firstly, there is the empiricist version of the 'theory theory of concepts'. Empiricism is the commitment to the derivation of new knowledge by our capacity to perceive reality, make consistent observations and to test theories against these observations. In science, empiricism is given form by the experimental method. The scientist has a problem to solve, and develops a hypothesis which will solve the problem. An experiment is designed under which conditions are controlled in such a way as the hypothesis will either be corroborated or nullified by the evidence. By empiricist doctrine, concepts are created and given meaning by experiment and observation. The physiologist can measure heart-rate, blood pressure and blood oxygenation levels in an individual and these concepts contribute to a common understanding of how the circulatory system operates in mammals.

Secondly, a rationalist approach to the 'theory theory of concepts' subordinates empiricism to the philosophical belief that knowledge exists independent of what we may observe. All observational data is filtered through our senses and given meaning within the context of the cognitive functions of the human mind. The rationalist asserts that beyond this sense data and the inherent logic these observations seem to obey, there are in fact deterministic laws and properties of matter independent of all we might measure and imagine. Scientific theories are coded representations of these objects and laws in reality. The concepts used by scientists are mental objects which directly correspond to reality. For the rationalist, concepts like 'apple', the number eighteen, or the laws of gravity all inhere in reality, and are only meaningful because they exist in reality.

Historicist epistemology creates a third philosophical doctrine for a 'theory of theory. Hjørland argues that historicist understandings of concepts cannot be ignored in understanding the source and meaning of concepts. The historicist accepts that theories and the conceptual language of theories are culturally and socially situated, even in the sciences. The history of science is the history of theoretical changes which are not independent of the technology, politics, economics and society in antecedent historical periods. Kuhn's description of theoretical change in the sciences is often described as a historical analysis. Paradigms are cultural and linguistic constructs of how scientists 'do science', products of the historical milieu in which scientists work and make new discoveries.

The early history of genetics is an interesting case in point for the biological sciences. Many European geneticists in the early twentieth century were devoted to the ideals of eugenics, or controlling the characteristics of human populations by breeding for favourable attributes such as intelligence or strength. The politics of the Nazi party were founded on the ideal of a Germanic master race with superior physical and mental traits; genetics provided a scientific basis to these notions of racial superiority. Early genetic theory was therefore influenced by politics and perspective of society which saw certain types of people as naturally more gifted than others. Genes and inheritance offered an explanation for why specific characteristics existed, and in turn concepts in genetics were themselves shaped by talk of eugenics and the hope that criminality or physical deformity might be 'bred out' of the human populations by the application of genetic theory.

Historicist epistemology therefore offers an interesting philosophical foundation for approaching the 'theory theory of concepts'. Historical contexts influence theoretical talk and consequently the meaning of concepts in scientific communities. Critics level accusations of relativism, scepticism and antirealism against historicism though, claiming that scientific theories are in no way contingent on history.

The fourth and final philosophy Hjørland describes in the context of theories of concepts is that of pragmatism. He writes: "Pragmatism understands concepts as a way to fixate parts of reality in thought, language, and other symbolic systems" [136].

The basis of a pragmatist understanding of concepts is that all knowledge is teleological. Scientists devise new theories, languages and practices in order to achieve certain research goals. Concepts are constructed to help achieve these aims, being flexible classes negotiated by working scientists to get a job done. In order to understand concepts from the perspective of pragmatism, one must account for research aims a community of scientists share, and endeavour to frame concepts in the light of these goals.

The four epistemologies outlined above – empiricism, rationalism, historicism and pragmatism – can be applied to explain the sources and meanings of scientific concepts. To take an example in biology, the concept of 'a gene', as understood through the lens of historicism, is situated in the historical context in which it was used. In the 1920s, before biologists knew the structure of DNA, a gene was an inheritable unit, somewhat abstract and subject to abstruse analysis by early geneticists. The modern concept of a gene is very much structural, with clearly delineated regions and physical attributes which govern its behaviour. From an empiricist perspective, the concept of gene has changed with the development of increasingly sophisticated methodologies, such as structural chemistry and genetic studies, which have permitted ever finer observable details to be gleaned about genes. The rationalist would argue that everything a gene is inheres in the structure of reality

regardless of how we as humans perceive that reality. The development of seemingly more complex understandings of the inner workings of genes is nothing more than an appearance, as we struggle to uncover what the logical essence of genes are in relation to the development of life on Earth. The pragmatic approach to the concept of 'a gene' is contingent on what purpose the concept serves in a theory or discourse. Genes as elements a biologist can control in order to manipulate the identity of a species can be contrasted with notions of our personal genetic information as a private, legal entity to be protected, or revealed, by the laws of society: these two pragmatic concepts of a gene classify reality into quite different aspects consistent with different goals.

#### *3.2.1.4 Concepts in biology*

The Gene Ontology claims the vocabulary contents and knowledge it represents correspond to entities in reality and not concepts in the minds of biologists. For the Gene Ontology, there are no concepts and Hjørland's talk of theory as being central to understanding what concepts are is largely redundant. There is only reality, regardless of the theories in biology.

In practical terms, although ontologists in the biosciences domain largely resist the idea of concepts as "...constructed and collectively negotiated meanings", I argue that working biologists and philosophers of biology would accept that the 'theory theory of concepts' is relevant to understanding knowledge and theoretical change in biology. Empiricism for example is largely consistent with the scientific method. Biologists make impartial observations of Nature and, in organising this data and inferring relationships between different observable entities, identify meaningful categories and classes in reality. By the empirical method, by experiment and the testing theories based on objective, value-free and context-independent observations, concepts are realised and validated.

For example, the classification of proteins based on functional classes inferred exclusively from their primary structure (the sequence of individual amino acids), is one empirical method by which biologists might try and discover sets of meaningful concepts to describe what proteins do [188]. As bioinformatic techniques increase in their power and application, and as structural detail about proteins increases in detail and reliability, so the concept of what a protein is and how it can behave in the context of the cell is transformed.

The biology of prion proteins is an interesting case of how empiricism changes concepts in biology. With the discovery that protein particles could transmit the disease CJD between organisms, the idea of the protein particle as an infective agent, capable of reproduction and serious pathology, was born. The experimental evidence suggested an entirely new role for single proteins in the life of organisms. Prions subverted even traditional categories for what could constitute an infective agent, being structurally much simpler than bacteria or even viruses.

Rationalism is another popular ideal in the biosciences domain, and can act as the source for new theoretical concepts. Bioinformatic models rely on reductionism to tackle complex biological problems. Biological processes and pathways are presumed to be decomposable into constituent processes and atomic functions. Computer models test supposed essential properties and entities in artificial environments and compare the results against real-world situations. The deduction of properties and likely explanations by logical models confers new biological knowledge *a priori* of any empiricist understanding of concepts. Computational biologists do believe they can identify new concepts by rationalist means, independent of any observation or experiment conducted in the spirit

of empiricism. The Gene Ontology is one tool amongst many in the bioinformatics field designed to support research according to this reasoning, and its products are conceptual changes increasingly free from empirical dependencies such as experiment and observation.

Species concepts in biology typify the interplay of empiricism and rationalism in the biosciences epistemology. Empiricist concepts for species include observable characteristics such as morphology or reproductive means which group organisms into species and taxa. Rationalist concepts for species take these characteristics to be secondary to the evolutionary history from a common ancestor which is independent of the subjective weaknesses of classification by appearance. These two types of concept theories therefore give rise to two different species concepts and two different taxonomic methods.

Do theoretical concepts in biology change over time? If theories and the concepts they delimit are a social product, and one accepts paradigm change as a fact of science, then clearly concepts change with the times. A hard realist, committed to rationalism, as many ontologists seem to be, would deny conceptual change. The laws of reality are constant. What appears to be conceptual change is only ever a better articulation in human terms of what those laws, final and consistent, actually are. The end road of this reductionist reasoning is therefore the 'final version' of biological knowledge, amenable to representation by ontology, whereby all the laws of Nature are fully explained and all observations in reality are explicable.

Section 3.2.3 test the four different epistemologies which provide a philosophical framework to the 'theory theory of concepts' against a single term in the Gene Ontology. Each GO term is the product of a set of theories about molecular biology. Do the theories and the concepts they inform support the proposition that GO terms can be viewed as concepts in the sense explored in the previous section?

Or can GO terms resist this framing as concepts and stand, as ontological realism contends, purely as entities in reality?

### 3.2.2 Anatomy of a GO term, 'cardiac cell differentiation'

Before proceeding to this test, the structure and origin of a single Gene Ontology term from the Biological Process Ontology, GO:0035051 or 'cardiac cell differentiation' is described.

#### *3.2.2.1 Why was this term added to the ontology?*

The GO term 'cardiac cell differentiation' (GO:0035051) was first added to the Biological Process Ontology in an edited ontology file dated November 2003. Terms are continually being added to the ontology since the GO project was first initiated. Since the ontology is the representation of all the biological processes occurring in any species, the work of term addition is considered by the developers to be an ongoing effort towards a complete description of all the entities present in reality relating to molecular biology. The intention is that eventually *all* the processes that can be said to occur in living cells will have a GO term.

Very few details are available regarding why 'cardiac cell differentiation' was added. This observation is quite normal in the context of the GO project. The term is not mentioned in the minutes of any GO meetings for the period prior to November 2003, and nor are there any discussions on the GO mailing lists dealing either with cardiac-related term or cell differentiation.

The GO file is not marked with which developer was specifically credited with the addition of this term, although external groups may be designated a set of GO IDs so that it is possible in some circumstances to trace additions back to particular interest groups, such as the Yeast Genome Database group or the Arabidopsis (Plant) Genome Database Group.

'Cardiac cell differentiation' was not added to the ontology by one of these groups, instead being created by a centrally located GO developer with editing privileges. A code in the ontology files indicates exactly who authored the definition for GO: 0035051, an individual originally affiliated with the Flybase database who, as of January 2012, is now a senior curator on the Gene Ontology and working with the European Bioinformatics Institute.

For years after its addition to the Biological Process Ontology, GO:0035051 remained in the ontology files. Its definition since addition is unchanged and reads:

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

Individuals at institutions linked to the Gene Ontology project annotate GO terms to gene product entries in species databases. These annotations are equivalent to subject headings in bibliographic records, or index terms in a book. Every annotation to a gene product is given an evidence code, indicating the empirical grounds for that annotation. Since November 2003, GO:0035051 has been directly annotated to a number of gene products, meaning that these gene products play a role in the biological process of cardiac cell differentiation (see Table 7).

**Table 7: Direct annotations of GO:0035051 to human gene products**

| Gene product symbol | Full name for gene product | Evidence-type for annotation |
|---|---|---|
| **BMP2** | Bone morphogenetic protein 2 | Inferred by Direct Assay / Inferred from Mutant Phenotype |
| **MYOCD** | Myocardin | Non-traceable Author Statement |
| **SOX6** | Transcription factor SOX-6 | Inferred from Electronic Annotation |
| **SOX6** | Uncharacterized protein | Inferred from Electronic Annotation |
| **SOX6** | Uncharacterized protein | Inferred from Electronic Annotation |
| **SOX6** | Uncharacterized protein | Inferred from Electronic Annotation |

BMP2 was annotated to 'cardiac cell differentiation' after a curator, having studied a primary paper, considered assay evidence to be strong enough to warrant an association between the GO:0035051 and the gene product. In addition, the annotation was validated by evidence from an experiment in which a mutated BMP2 gene disrupted the differentiation of cardiac cells.

MYOCD became annotated on the strength of a non-traceable author statement, which is a comment in a published paper or monograph. The four references to SOX6 in the table above are annotations created between the GO term 'cardiac cell differentiation' and four different, related

protein sequences. These sequences were generated from automated analyses of transcripts from the human SOX6 gene, variants of the original gene product which are likely to be synthesised in humans. This is a result which is yet to be validated, and is considered as a weaker type of evidence for annotation.

These then are the annotations to GO:0035051, the Gene Ontology in practice and yet over time, the ontology changes. A group with special responsibility for GO term annotation to heart-related genes was created in 2007 with the formation of the Cardiovascular Gene Ontology Annotation Initiative at UCL. This group was funded to employ several dedicated cardiovascular system experts who were charged with the task to create high quality, manual annotations to a list of cardiac-specific genes. These experts and the out-reach programmes they initiated regularly offer comments on existing GO terms and offer proposals for the inclusion of new, cardiac-specific ontology terms. They both users create annotations, and act as ontology editors by suggesting ontology changes via the Sourceforge tracker system.

Curators working for the Cardiovascular Gene Ontology Annotation Initiative hosted a workshop in September 2009, and experts in the field of cardiac biology edited and expanded the heart development process ontology. The GO term 'cardiac cell differentiation' survived this scrutiny by the expert panel. As of July 2010 the term is annotated to 463 unique gene products across several different species databases, including gene product databases for mouse and human genomes.

According to my dataset of GO papers identified in MEDLINE, 116 papers were published in the period 2000 – 2003 in which the Gene Ontology is either mentioned, or used in an analysis. GO papers were defined as those mentioning the Gene Ontology, either in full or abbreviated form, in the title, abstract or indexing. Papers for this period, or authors using out-dated versions of the Process ontology after the term 'cardiac cell differentiation' was added in November 2003, will not have indexed gene products or found any empirical associations to 'cardiac cell differentiation'. Until a term is added to the Gene Ontology, it effectively does not exist for the purposes of data analysis. This is a normal consequence of the GO Consortium strategy of regularly updating the ontology with new terms. Users are advised to download the latest versions of ontology files from the GO website for use in analyses and in association with software tools. GO developers are aware that applications may be using out-dated ontology files, and this is why the Consortium advises authors to report the ontology version they are using.

As Chapter 3.6 will reveal via a detailed content analysis of published papers using the Gene Ontology, authors rarely follow this requirement.

#### *3.2.2.2 Ontology relations*

The GO term 'cardiac cell differentiation' exists only in the Biological Process Ontology. Biological processes are considered in GO parlance to consist of multi-step reactions or multiple functions combining to achieve a biological goal. In this case, genes annotated to the term have been shown to be involved in a process which has the goal of turning a less specified cell type into a cell found in the heart. What is meant by a cardiac cell and the process of differentiation will be discussed in more detail below.

Broadly speaking, 'cardiac cell differentiation' is related to two major parent processes in the Process 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.

Figure 5 demonstrates the multiple pathways a term can take back to the root of the ontology. Each relationship between individual terms must obey what is described by the Gene Ontology as 'The True Path Rule' or 'Judy's Rule' (named after one of the instigator's of the GO project). The True Path Rule dictates that *all* paths through the ontology must be true. Relations between individual terms establish theoretical statements, and each statement must be true. Further to this, the path between all terms to the root must also be true. The consequence of the rule is that a user can infer that any gene product annotated to 'cardiac cell differentiation' must, by virtue of any 'is_a' relations outlined in the diagram below, share a true relation to higher level terms like 'anatomical structure development' and 'cell differentiation'.

The power of this approach lies in the application of the ontology and annotation files to gene expression analyses. Annotations to terms deep in the ontology can be collapsed into higher level terms, enabling functionally similar gene products to be grouped together. Gene products involved in cardiac cell differentiation in mice can therefore be functionally related to, for example, bovine gene products annotated to the GO term 'nose development' by virtue of common relations to the term 'system development' (GO:0048731).

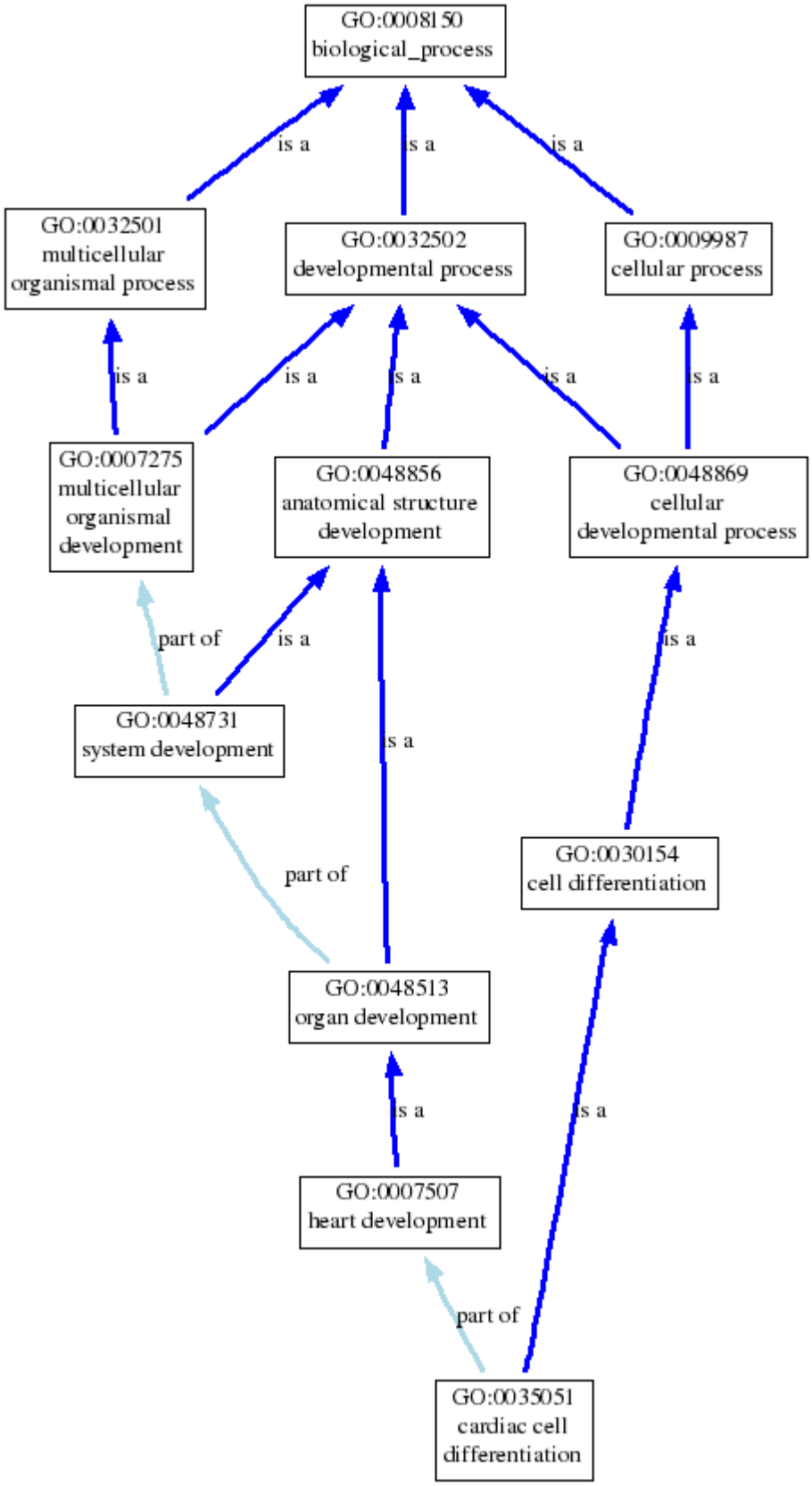
The GO term 'cardiac cell differentiation' is itself a parent term to a number of more specific children. As of July 2010, these terms and the dates they were added to the ontology, were as follows:

- GO:0003348, 'cardiac endothelial cell differentiation' (January 2010)
- GO: 0060935, 'cardiac fibroblast cell differentiation' (October 2009)
- GO:0060950, 'cardiac glial cell differentiation' (October 2009)
- GO:0055007, 'cardiac muscle cell differentiation' (November 2006)
- GO:0060945, 'cardiac neuron differentiation' (October 2009)
- GO:0003292, 'cardiac septum cell differentiation' (November 2009)
- GO:0060947, 'cardiac vascular smooth muscle cell differentiation' (October 2009)
- GO:0010002, 'cardioblast differentiation' (March 2003)
- GO:00039293, 'heart valve cell differentiation' (November 2009)
- GO:0007513, 'pericardial cell differentiation' (November 2003)

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'. These terms appear in ontology file revisions for October or November 2009 and judging from the meeting minutes, were added consequent to the Heart Development workshop held at UCL in September 2009 [189]. As of January 2012, no annotations had been created to these terms which, according to

the guidance of the Gene Ontology Consortium, refer to real-world biological processes. At some point in the future, gene products may be indexed with these terms.

Figure 5: 'Cardiac cell differentiation' relations



#### *3.2.2.3 Facet analysis*

The Gene Ontology is an enumerative system of classification: it lists the system of classes appropriate for indexing gene products from different species.

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 Gene Ontology yet it is the contention of my thesis 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.

The principles of facet analysis will next be applied to the GO term 'cardiac cell differentiation', in order to illustrate how facets structure the existing enumerative system of classification for gene products in the ontology.

Ranganathan originally suggested five facets for ordering knowledge based on the PMEST formula [190, 191]:

- Personality
- Matter
- Energy
- Space
- Time

According to the PMEST formula, what facets compose GO:0035051, 'cardiac cell differentiation'?

If personality is the character of a subject which distinguishes it from other subjects, then GO:0035051 is marked as dealing with the facet of the biological process *differentiation*. Differentiation is commonly defined in the biosciences domain as the maturation of less specialized cell type into another cell type with a more specialized form and function. A simple example is the differentiation of different cell types found in the blood from less specialized cell types originating in the bone marrow.

Matter is the physical substance of which the subject is composed. In this respect, GO:0035051 deals solely with the matter type 'cells', cells being unitary, membrane-bound subunits of multicellular organisms, or the independent unit of life found in unicellular organisms and noted for its capacity to self-replicate. Other matter types enumerated in the Gene Ontology include biochemical molecules, organelles within cells, and whole organisms.

Energy according to PMEST 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.

The space facet in PMEST deals with the positional component of a subject. The Gene Ontology articulates a large number of locations for cell activities, ranging from locations within an individual cell to places where events occur within whole organisms, such as limbs or organs. In the example of 'cardiac cell differentiation', the space facet is the heart or heart-like structure of multicellular organisms.

Finally, the time facet of the PMEST formula describes a temporal period associated with the subject. In biological terms, the biological time of an organism or cell within an organism is a point in the life cycle which runs from the very first moment of conception to the ultimate death of an organism. The development of organs within multicellular organisms is one such point in time of the life cycle, and 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 expanded Ranganathan's original PMEST formula to include further facets appropriate to the description of entities in science and technology [190]. 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 facetted classification formula one may choose to use. Although the Gene Ontology is an enumerative list of terms for indexing gene products, each term possessing a referent in reality, there are recurring themes for concepts which can be decomposed into facets. One may apply the process of differentiation to different types of cells beyond cardiac cells, and indeed this is partially accomplished in the Gene Ontology through sections of the graph dealing with differentiation of specialized cell types in the blood, nervous and respiratory systems.

Beyond processes, substances and locations, the facets embodied by a term like 'cardiac cell differentiation' are limited. The power and potential for a facetted classification to describe a very large number of properties for gene products in any number of biological situations will be explored in Chapter 11, 'Alternative classification standards for biology'.

#### *3.2.2.4 Synonyms in other vocabularies and indexes*

The NISO Z39.19 standard (see Chapter 3.3.2 ) advises that a vocabulary compensate for the problems created by synonymy by using a single preferred form for a concept. A vocabulary should, but is not required, to provide guidance on term variants commonly used in the domain.

The only synonym offered for GO:0035051, 'cardiac cell differentiation' as of July 2010 is a single related term 'heart cell differentiation'. Related terms in the Gene Ontology are synonyms which are neither narrower, broader or equivalent to the main GO term indicated, and their precise, somewhat unclear status in the ontology is explored further in Chapter 4.1 .

Is it valid to assume that the GO term, 'cardiac cell differentiation' has no other synonyms in the biosciences domain? One might extend a search to appraise the literary warrant behind the choice of this term by searching other controlled vocabularies, a methodology facilitated by the BioPortal ontology service provided by the NCBO [192] and the EBO Lookup Service at the European Bioinformatics Institute.

Results of detailed searches of these services for ontology terms related to 'cardiac cell differentiation' are provided in the appendix. In brief, the strategy taken was to retrieve terms relating to the two major facets of GO:0035051 'cardiac cell differentiation': the concept of cardiac cells and the concept of differentiation. The results were as follows.

A search for the term 'cardiac cell' or 'heart cell' yielded a total of eight related terms in non-Gene Ontology vocabularies, see Table 8.

**Table 8: 'cardiac cell' terms found via the search of the BioPortal facility, July 2010**

| Ontology source | Term name | Description |
|---|---|---|
| **Foundational Model of Anatomy (FMA)** | Heart cell | Homonym for 'endocardial cell' and 'cardiac myocyte' according to FMA notes |
| **NCI Thesaurus** | Adult Cardiac Cellular Rhabdomyoma | A neoplastic process occurring in cardiac cells |
| **Neuroscience Information Framework Standard Ontology (NIF Ontology)** | cardiac cell development | Deleted from the NIF Ontology without history notes (as of January 2012) |
| **Neuroscience Information Framework Standard Ontology (NIF Ontology)** | cardiac cell differentiation | Deleted from the NIF Ontology without history notes (as of January 2012) |
| **Cell Line ontology** | nodal cardiac cell | Obsoleted without notes, other than a reference to an obsolete class in the Cell Type Ontology (as of January 2012) |
| **Cell Type ontology** | nodal cardiac cell | Obsoleted with reference to use the Cell Line ontology term 'nodal myocyte' (as of January 2012) |
| **Mammalian phenotype** | abnormal cardiac cell glucose uptake | Defined in the Mammalian Phenotype Ontology as 'the anomalous ability of the cells of the heart to take in glucose' |
| **Cardiac Electrophysiology Ontology** | heart cell | General anatomical term without definition |

The Foundational Model of Anatomy ontology, an important and well-developed vocabulary in the anatomical domain, indicates that a common homonym for 'cardiac cell' is the term 'cardiac myocyte'. Extending the search to include this variant on the BioPortal service retrieves 37 different related terms from assorted ontologies, with variants of the term 'cardiac cell' found to include 'cardial cell', 'cardiac muscle cell', 'cardiac myocyte', 'Myocytes, Cardiac' and 'cardiomyocyte' (see Appendix).

A search of the EBO Lookup Service using the stem 'cardi%' retrieved over 800 terms, some unrelated to cell types in the heart, but a significant number being identifiable as specialised cardiac cells and tissues (see Table 9). Based on the findings of these searches, 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 Gene Ontology 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 such as the insect, *Drosophila*.

**Table 9: Examples of specialized cell types and tissues found in the heart, as determined by searches of the EBO Lookup Service and BioPortal**

| Examples of cardiac cells |
|---|
| **Cardiac Purkinje cell** |
| **Cardiac myocyte** |
| **Cardiomyoblast (precursor to the cardiac myocyte)** |
| **Nodal cardiac cell** |
| **Cardiac mesenchymal cell** |
| **Corpus cardiacum (a part of the heart which releases hormones)** |
| **Cardiac jelly (noncellular substance found in the embryo)** |
| **Cardioblast** |
| **Cardial cells (found in *Drosophila*)** |
| **Embryonic cardioblast** |
| **Superior cardiac nerve (and related)** |
| **Primitive cardiac myocyte** |
| **Small adult cardiomyocyte** |
| **Embryonic/larval heart anchoring cell** |
| **Embryonic heart anchoring cell primordium** |
| **Larval heart anchoring cell** |
| **Neoplastic cells in the heart tissues** |

A specific search for 'cardiac cell differentiation' in the BioPortal search facility yields a hit in the NCI Metathesaurus. This term is a duplicate of the same Gene Ontology term, and this duplication is a result of the BioPortal indexing both the metathesaurus and the source for the metathesaurus. Interestingly though, a search for 'cardiomyocyte differentiation' did retrieve terms from the Gene ontology for 'ventricular cardiac muscle cell differentiation' and 'atrial muscle cell differentiation'. This seems to be an effort on the part of the GO developers to distinguish between cardiomyocytes in different anatomical locations, a distinction not found in other vocabularies.

The BioPortal search was finally extended to the facet of general, rather than specifically cardiac, cell differentiation. A simple search for the string 'cell differentiation' resulted in 744 hits (as of 02 July 2010; a similar search in January 2012 indicates 757 hits). Cell differentiation terms in biological ontologies can be broadly categorised into four important groups:

1. Mammalian Phenotype Ontology terms listing abnormal cell-specific differentiation processes such as 'abnormal mast cell differentiation'
2. Large numbers of Gene Ontology biological process terms describing cell-type specific differentiation such as 'axial mesodermal cell differentiation'
3. Some broad NCI Thesaurus terms describing cell differentiation, together with some more specific terms for differentiation in specific cell types such as 'Clara cell differentiation'
4. Cell Cycle Ontology terms for the regulation of cell-type specific differentiations such as 'positive regulation of smooth muscle regulation' (these have specific cross-references to equivalent terms in the Gene Ontology)

Beyond the Gene Ontology and other vocabularies indexed by BioPortal and the EBO Lookup Service, the Medical Subject Headings (MeSH) offers an alternative system for indexing bioscience concepts. MeSH is well developed, having received significant investment and attention for the National Library of Medicine and does describe several concepts relevant to the GO term 'cardiac cell differentiation'.

There is no entry for 'cardiac cell' in MeSH, and the closest term that might be identified with this GO facet is 'Myocytes, cardiac' defined by MeSH as:

> "Striated muscle cells found in the heart. They are derived from cardiac myoblasts"

MeSH indicates three different synonyms for 'Myocytes, cardiac' including 'Cardiomyocyte', 'Muscle Cells, Cardiac' and 'Muscle Cells, Heart'. None of these synonyms are a match for this more general facet found in the Gene Ontology which is of a general category of cardiac cell, further substantiating the observation that this concept is relatively unique to the GO vocabulary.

Incidentally, cardiac myoblasts are the precursor cells to cardiac myocytes; they are not mentioned in any Gene Ontology biological process term names, definitions or synonyms.

With respect to the facet of cell differentiation in the context of GO: 0035051, 'cardiac cell differentiation', MeSH offers a scope note for cell differentiation as:

> "Progressive restriction of the developmental potential and increasing specialization of function that leads to the formation of specialized cells, tissues, and organs."

A further, related scope note offers guidance on disambiguating between a *cell lineage*, which is the developmental history of differentiated cells traced back to originating stem cell types in the embryo, and a *cell line* which is derived from cultured cells *in vitro*. The Gene Ontology makes no such distinction between cell differentiation processes that may occur *in vivo* or *in vitro*, and considers a cardiac cell differentiating in the developing heart of an individual to be undergoing an equivalent process to cultured cardiac cells dividing and differentiating under laboratory conditions.

Canonical textbooks in the domain can be an indicator of established knowledge and a source of literary warrant for controlled vocabularies. Berne and Levy's 'Cardiovascular physiology' subtends the cardiovascular system into various sub-systems such as hemodynamics and the control of cardiac output [193], as does Levick in 'An introduction to cardiovascular physiology' [194]. Cardiovascular cell types are dealt with only in passing in both titles which discuss the unique features of the myocardial cell and its distinction from skeletal cells. These unique features include the disproportionately large number of mitochondria in myocardial cells, and their unusually rich capillary supply. Ventricular cells, that is cells belonging to the walls of the ventricles in the heart, are specifically mentioned in both textbooks, and implies that the cardiovascular physiologist may distinguish between subtypes of cardiomyocytes, depending on their anatomical location.

Of note is the detail that Berne and Levy refer to, and index, ischemic cells. Ischemia is a pathological state, and in the heart refers to the starvation of heart tissue of oxygen and nutritional supply. Disease states are not dealt with by the Gene Ontology, although one may consider the creation of ischemic cell sub-populations as a form of cell differentiation enacted by gene products. The GO term GO: 0002931, 'response to ischemia' in the Biological Process Ontology is the closest concept

covering ischemic changes, yet is annotated to only a few gene products. The Gene Ontology represents an idealised form of the molecular biology of a cell. The very normal sense in which a cardiovascular physiologist will consider aberrations in cardiac cells, such as ischemia, are virtually silenced in the GO vocabulary, and serves to limit its application.

What do these kinds of searches tell us about the Gene ontology 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 Gene Ontology 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 Gene Ontology, probably 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 relatively new ontologies in the domain, 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 Gene Ontology 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 facetted 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. The consequences of this approach, particularly with reference to the proliferation of terms, will be discussed in later chapters.

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 Gene Ontology'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 Gene Ontology 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.

The long-term view of the Consortium is that eventually, processes and functions occurring in particular anatomical locations will be cross-referenced to a separate, species-neutral anatomy ontology, coupled with more species-specific anatomy ontologies. Much of the Gene Ontology, as it has existed since inception and still stands today, accepts the necessity of anatomy-centric terminology. Therefore 'cardiac cell differentiation' is a term which represents a larger conflict in the ontology, a denial of an explicit aim to be species-independent, yet agnostic to this requirement of ontological realism in the face of a need to create a workable classification to index gene products.

### 3.2.3 A GO term according to four different concept theories

This chapter has thus far outlined what is meant by concept theories, and Birger Hjørland's explanation of the 'theory theory of concepts'.

Further to this outline, one term from the Gene Ontology, GO:0035051 - otherwise known by its term name 'cardiac cell differentiation' was explored based on the reasons why it was added to the ontology in the first place, and the definition, synonyms, and relationships it has to other terms in the ontology. Furthermore the term, a compound, was decomposed into different facets, and each facet searched for via various controlled vocabularies and indexes to articulate the nature of these facets in the context of modern molecular biology.

The GO term 'cardiac cell differentiation' is one of many thousands of terms in the Gene Ontology. The suggestion here is that it is not a straightforward, unambiguous pointer to a commonly accepted biological process in reality. The term 'cardiac cell differentiation' exists in a nexus of other vocabularies, recorded literature, empirical data and knowledge in the subjective minds of communities of working scientists.

The concept analysis will now be extended as 'cardiac cell differentiation' is considered in the context of the four different concept theories mentioned at the start of this chapter. What does considering this GO term through the lens of these different theories of concepts tell us about the limitations of the Gene Ontology as a universal classification?

#### *3.2.3.1 Empiricism*

The philosophy of biology is a blend of different philosophical approaches such as rationalism, empiricism and reductionism which together seek to explain the justification for biological theories. In Chapter 4.1, I will look at the difficulty of adopting a single, comprehensive philosophical outlook as a basis for functional explanations in biology, and the consequences for the Gene Ontology's attempt to create a classification of biological functions. In this next section though, I will simplify the philosophical landscape of biology down to a single viewpoint, that of empiricism, and ask how an empirical attitude to the origin of concepts in biology changes the potential classification of GO: 0035051, 'cardiac cell differentiation'.

Empiricism is the belief that our knowledge of biological systems is derived exclusively from sense experience. There is no *a priori* knowledge as advocated by rationalists. Whilst there may be no difference in the substance of a biological theory justified by rationalism and the same theory articulated by empiricism, the empiricist claims that the justification for that theory rests entirely on sensory evidence collected by instruments and experimentation. All biological knowledge is *a posteriori* to sensory data, and our interpretation of that data. All concepts are therefore drawn from the empirical method, and we can gain no knowledge of universal concepts in biological theory, such as the concept of a 'gene' or the concept of 'sexual reproduction', other than by an empirical approach.

Concepts in biology, as understood according to the theory of empiricism, are therefore rendered comprehensible by the kinds of experiments, empirical tests and data types biologists handle in their daily work. With respect to understanding the concept 'cardiac cell differentiation', we can therefore describe the kinds of experiments performed on, and data generated relating to, cardiac cells and their differentiation.

In the next section, three different empirical data-types found in molecular biology, flow cytometry data, gene expression data and visual data derived from modern microscopy are described. It will show how these three empirical techniques, which generate these data-types, have contributed to differing understandings of cardiac cell differentiation, and suggest how empiricism changes the classification of a concept like 'cardiac cell differentiation'.

Flow cytometry is a common technique in molecular biology which uses monoclonal antibodies against specific receptors on the surface of cells to 'tag' these cells with fluorescent markers. The specificity of monoclonal antibodies is such that different populations of cells can be identified expressing different repertoires of markers on their surfaces. A flow cytometry machine can measure the scattering of different wavelengths (or colours) of light from antibodies artificially designed with different fluorescent tags. Molecular biologists can therefore count the proportion of cells in a sample with particular kinds of cell-surface receptors, and can thus sort a cell population into different cell sub-types, such as peripheral blood monocytes expressing the markers CD14 or CD16 on their surface [195].

Flow cytometry is commonly used to identify stem cell progenitors, including cardiac stem cells. Different complements of cell-surface markers are assumed to be indicative of stem cell status and can be used to distinguish cells which are likely progenitors of cardiac cells. Arsalan *et al*. [196] describe flow cytometry methods they use to identify cardiac cells expressing the marker CD117 or 'c-kit' which is taken as a measure of stem cell status. In addition, the same authors sort cardiac cells for those expressing the markers CD3, CD11b, CD19, and CD45 which are assumed to be indicators of cardiac cell progenitors. The expression of different receptors on the surfaces of cells is therefore taken to be an indicator of future potential to become a cardiac cell, and the changing repertoire of cell surface markers, indicated by flow cytometry, conveys how far along a path of differentiation to a specific cardiac cell type progenitors have progressed.

Gene expression data similarly can type cells as cardiac lineages, but whereas flow cytometry collates data on protein receptors expressed on cell surfaces, gene expression techniques look at mRNA transcripts extracted from cells and tissues. Two very common empirical methods for detecting mRNA transcripts are rtPCR and microarray analysis. In both techniques, short genetic probes composed of sequences unique to particular genes are used to bind mRNA transcripts extracted from samples, which may be derived from cultured cells or tissues. Molecular biologists can use several different methods for detecting this binding between the probe and a transcript, and detection is evidence that a gene is expressed in a system.

How do gene expression techniques relate to an empirical concept of cardiac cell differentiation? Authors will typically use quantitative rtPCR techniques to measure the levels of transcripts associated with cardiac cell differentiation. Van de Ameele *et al.* [197] explore the role of previously uncharacterised transcription factor using gene expression methods, and assert a role for eomesodermin in the developing heart of embryonic tissues. This paper is representative of a common trend in molecular biology to determine the factors involved in different biological processes, and several comprehensive reviews cover the extensive literature of gene expression studies pertaining to the differentiation of cell types in cardiac tissues [198-200]. What these kinds of papers tell us about the concept of cardiac cell differentiation is that there is a strong empirical basis for categorising cells according to the genes they express. The concept itself is defined

empirically as a set of idealised gene expression data which suggest that the process of cardiac cell differentiation is occurring. In much the same way, flow cytometry data can be idealised in a model sense to offer an image of cell-surface receptors which, detected with the appropriate monoclonal antibodies themselves define and are signifiers of the concept 'cardiac cell differentiation'.

A third source of empirical evidence used to form the concept of 'cardiac cell differentiation' is visual data derived from modern microscopy techniques. The microscope was one of the earliest instruments used for collecting information about the cellular structure of living tissue, and today sophisticated imaging techniques are routinely used for collecting data on the physical structure of individual cells and sub-cellular structures.

Empirical evidence for cardiac cell differentiation collated using microscopy is impelled by cultured cells exhibiting spontaneous contractility *in vitro*. As stem cell progenitors differentiate in laboratory conditions, progression towards cardiac cell types is often recognised by cells beginning to contract rhythmically by themselves, which is itself a defining feature of cardiomyocytes in heart tissue. Classic papers in the field such as Kehat *et al.* [201] present static pictures of stem cells at various stages of differentiation together with electron microscope images of the ultrastructure of cultured cells to support the argument that the process of cardiac cell differentiation is occurring.

Empirical notions for GO: 0035051, 'cardiac cell differentiation' are further moulded by this routine, visual evidence generated by microscopy. The sight of spontaneously contracting cells down a microscope is of itself not taken as proof that at a previous point in time, cardiac cell differentiation occurred. Couple this visual data with gene expression evidence and flow cytometry studies to characterise the genetic and molecular nature of these cells though, and the premise is validated to a common standard of empirical evidence, as required by the norms of science across the domain of molecular biology.

The GO concept 'cardiac cell differentiation' can therefore be framed and defined in three complementary ways using empiricism. The process can be described as but one step in the ongoing presentation of sets of cell surface marker proteins. Alternatively, the presence of cardiac cell differentiation can be defined by the expression of an idealised set of genes which act in concert to 'be' the process. Finally, molecular biologists can see cardiac cell differentiation as the progressive adoption of physical features which look like a cardiac cell down a microscope.

The Gene Ontology does not capture these empirical notions for the concept of 'cardiac cell differentiation'.

#### *3.2.3.2 Rationalism*

Contrary to empiricism is the rationalist perspective on concepts. Rationalism holds that all knowledge about molecular biology exist *a priori* of any experiment or experience of reality. The data the molecular biologist collects from the flow cytometer or microarray or through the lens of the microscope is a confirmation of the rules and laws undergirding natural phenomena. Knowledge is independent of how empirical data is collected and interpreted. Most importantly, knowledge is independent of how the individual molecular biologist thinks about and experiences that data.

By a rationalist theory of concepts, 'cardiac cell differentiation' is not a term that has developed from increasingly more sophisticated experiments exploring the nature of heart cells and their

development. Rather, the concept is logically entailed by reality, and exists as one of many processes occurring within instances of hearts. A rationalist theory of concepts is therefore closely aligned with the ontological realism and the requirement of 'instantiability' in ontologies.

The important consequence of adopting rationalism as a foundation for concepts in ontologies is the possibility that new categories and ontological relationships may be deduced according to axioms. This is certainly a stated aim of many ontology developers in the biosciences domain. The very complexity of modern biology in a sense requires the ability to reason across ontology graphs, in order to infer new annotation terms for uncharacterised entities or the deduction of new molecular function or processes that have yet to be studied in an empirical sense.

Annotations inferred from electronic algorithms, which make up a significant proportion of all GO annotations, are one example of rationalism in practice in the Gene Ontology. The attribution of functional categories to putative gene sequences which have never been studied under experimental conditions is performed by computer analyses looking at sequence similarity. The concept of 'cardiac cell differentiation', and the likelihood of instances of this process occurring in unexplored biological systems, is logically entailed by rationalist thinking and the sequence similarity alone.

Equally, the paths through the ontology graph logically entail that any gene product annotated to a more specific GO term must also adopt the less specific functions closer to the ontology root. All gene products associated with 'cardiac cell differentiation' must also participate in the higher order process 'anatomical structure development'. The weakness of this reasoning is apparent in artificial systems, in the experiments molecular biologists create in the course of an empirical study. The spontaneously contracting cardiac stem cell in a Petri dish may never form a complete heart. Rationalist concepts for GO terminology imply that in an idealised system, where all other variables were necessarily controlled, the cultured cell could progress in its differentiation to eventually form part of a true, anatomical structure of a heart.

Finally, rationalist thinking in ontologies guides the notion that cross-references between terms in different ontologies are *a priori* true. Categories of cells in a cell-type ontology can be cross-linked to cellular processes in the Gene Ontology to infer compound processes such as 'gastric cell differentiation', even if this node does not exist in either ontology as an independent term. The discovery of a new cell-type or even of a new species necessitates, by rationalist commitment, the existence of all attendant functions and processes.

To cite a flippant example, the discovery of a living unicorn would, based on best current biological knowledge and ontological realism, logically entail that 'cardiac cell differentiation' occurs in this once-mythical beast. The empiricist would resist this *a priori* reasoning until the empirical evidence had been collated. The rationalist though may accept logical entailment, and thus avenues of work exist where, for example, resident progenitor populations of cardiomyocytes are assumed to exist [202] because current best knowledge implies they are, *a priori*, inevitable, and are therefore discoverable.

Rationalism for concepts means that there must be necessary and sufficient conditions to unambiguously distinguish a cardiac cell from any other type of cell. This type of rationalism verges on a belief in natural kinds, much like the essentialism discussed in the Introduction. One may relax

this condition to admit probabilistic models for cell types. If variables for characterising a cell fall within certain parameters then an instance may behave more or less like cardiac cell. This probabilistic approach to rationalist concepts is ignored by the Gene Ontology. There is no difference between a cardiac cell in a human or a dog according the GO philosophy. Nor do cardiac cells in different groups of individuals – for example, cardiac cells in healthy people and cardiac cells in persons with atherosclerosis – merit distinction in the ontology.

All cardiac cells form a single logical class in the Gene Ontology. Similarly an all-or-nothing definition for cell differentiation holds in that cardiac cell differentiation is a binary condition: it either is occurring, or it is not. The suggestion that gene product expression is a stoichiometric process whereby sub-optimal levels of gene products do not permit a biological process to complete, is resisted. By this I mean that what may ostensibly look like the process of cardiac cell differentiation would, by dint of insufficient levels of even just one gene product, not be cardiac cell differentiation.

This is a simple biochemical principle, yet the Gene Ontology's simple classification for process concepts does not admit this level of sophistication.

#### *3.2.3.3 Historicism*

Historicism in relation to concept theory for molecular biology is the idea that knowledge in the domain, and the knowledge one may represent in an ontology, changes with culture and with society. What molecular biologists believe, even though it may be presented as entirely scientific according to empiricism, rationalism, or some combination of the two, is in fact shaped by much larger social structures and norms.

Can this historicist approach be used to argue that the concept 'cardiac cell differentiation' is socially and culturally situated?

Before genes were discovered, biologists knew that cardiac cells existed. They looked and behaved differently to other cell types, and were quite distinct from their closest counterpart, skeletal muscle cells. Biologists were aware that this unique cell population, possessing special properties such as auto-contraction, and only found in the heart organ, originated from some simpler cells in the developing embryo which did not share these properties. Cardiac cells came from somewhere via the process of differentiation.

This process could not be understood in term of genes and gene expression because genetics had yet to be discovered. Instead, the process was understood in terms of the best conceptual tools biologists had at the time, which was biochemistry. Cardiac cell differentiation was therefore the release and effect of various, as yet uncharacterised chemical factors which, in concert with the cell machinery and cell division, acted to form what appeared down the microscope and was found to be located pulsing in the hearts of living organisms.

'Cardiac cell differentiation' was a type of biochemical process, a chemical transformation of a primitive, prototypical cell kind into a specialised cardiac cell.

The discovery of genes and their mode of action transformed what experts in biology understood to mean by the concept 'cardiac cell differentiation'.

Rather than a primarily biochemical process, the concept differentiation was changed by an understanding of genes and the mechanisms of gene action. Every cell in even complex, multi-cellular organisms like Man was found to contain a complete copy of the genome. This genome itself contained all the information necessary to create and sustain a living organism from conception to death. The discovery of genetics and the genetic code fostered the notion that within each living cell is a complete copy of the programme necessary to create life. The idea of the genetic 'programme' was itself tied to developments in other areas of science such as physics and computer science, with organisms being likened to complex machines, coded for by this complex genetic information source.

The different cell types which compose an organism such as Man, rather than being categorised according to their unique physical and chemical properties, became defined by their genetic properties. Different cell types were found to express different sets of genes under specific conditions. Coded into the genetics were developmental and differentiation programmes, the controlled expression of gene products aimed at orchestrating the complex cell divisions and timed cellular changes necessary to lay down tissues, form organs and sustain physiological processes in the mature adult.

'Cardiac cell differentiation' in this post-genomic era took on the characteristics of a computer programme, the articulation of a complex set of genetic events and triggers which ultimately created the cardiac cells in a heart. Knowledge about cardiac cell differentiation was prioritised as knowledge of the underlying genetics controlling this process, rather than the physical or chemical knowledge about how cells change.

Concepts for the cell and for terms in the molecular biology domain like 'cardiac cell differentiation' are now shaped by the most important cultural and social phenomenon of our present era which is the proliferation of technology into everyday life.

Hi-tech solutions such as e-science are now viewed as the best tool to solve intractable biological problems, and the problems themselves have been translated into a new paradigm in biology, the paradigm of systems biology.

As molecular biologists have discovered more and more, a major difficulty has become that of separating the effects of the genetic code from the ever-changing environmental background against which genes are expressed. Whereas research once concentrated on single genes and their effects, modern molecular biology needs to handle the effects of many thousands of genes at many different systematic levels. This more holistic approach to understanding biology, known as the systems biology approach, has driven the development bioinformatic and mathematical modelling to support knowledge discovery in the domain.

Biological processes have become events modelled in a computer, a set of parameters and classes designed to imitate reality. The technologisation of knowledge discovery in molecular biology is accompanied by a commensurate shift in the understanding of concepts like 'cardiac cell differentiation'. This shift is a transformation of processes and functions into *logical classes*, to be manipulated by computer programmes, to enable reasoning by machines.

Ontologies representing this kind of machine-tractable knowledge are technological thinking extended, the reification of biological knowledge into computerised forms to facilitate the observation, interpretation and control of a stubbornly complex biological landscape. A historicist approach to concepts can accommodate changes in paradigm thinking in the domain. In the case of molecular biology, three paradigmatic ways of thinking about biological problems – the chemical, the genetic and the system.

#### *3.2.3.4 Pragmatism*

Pragmatism as the basis of a theory for scientific concepts advises that concepts are what scientists need them to be, in order to achieve certain goals. As Hjørland puts it:

> "Pragmatism is the ideal of basing knowledge on the analysis of goals, purposes, values, and consequences." [136]

If the work of classifying molecular functions and biological processes into an ontology is a pragmatic task, and we are to approach the concepts therein pragmatically, what kinds of goals might molecular biologists be intending to achieve? I will cover three potential purposes that molecular biology might aim at according to a pragmatic epistemology:

1. Defining a teleological explanation for cardiac cells
2. Illustrating a common transcriptional programme between species
3. Outlining potential therapeutic targets and valuable research objects

Broadly speaking, the aim of the biological sciences is to offer explanations for natural phenomena, for life [203, 204]. Whilst in later chapters I will be dealing with functional explanations, one important explanatory strategy I will touch upon now is described as *teleological*.

Teleology is the study of purpose or design in natural phenomena, and in molecular biology would refer to the purpose of specific molecular features in the cell. Teleological explanations are controversial [203-209] yet arguably the intended aim of the Gene Ontology is to represent knowledge about the purpose or design of biomolecular systems.

Thus the concept 'cardiac cell differentiation' is used to annotate gene products with purposes or reasons for their expression. A broad category of purpose for gene products or, in other words, a feature of the *design* of cellular processes is to facilitate the differentiation of unspecified cells into more specialised cell-types. The purpose of some gene products is therefore to create specialised cells like cardiac cells. A pragmatic approach to concepts in molecular biology, intended as a means for creating better classifications for gene products, therefore claims teleological explanation as the guiding principle for ontology development. Molecular biologists may choose what they perceive to be the most interesting or most important goals that cellular systems appear to present, such as differentiation, or cell division, or cell-cell communication, and base concepts and their relationships around the representation of these cellular goals.

Arguably, the Gene Ontology has already adopted a teleological classification of goals for cells, and classified gene products according to a teleological reasoning.

A second pragmatic goal of molecular biology is the validation of the theory that all species share fundamental cellular systems or features, and that one classification can be created to represent this

common transcriptional programme. This reasoning is the basis for all model organism biology [210], for the functional predictions of new genes [211], and comparative genomics [212, 213]. In fact, the Gene Ontology has always stated that its intended aim was to harmonise different domains in biology, to provide a common language for describing molecular functions in different species [214].

The Gene Ontology therefore explicitly follows this second pragmatic approach to concepts by assuming that it is possible to create one classification for functions and processes in all species. Following this approach yields numerous practical advantages for molecular biology, by facilitating cross-database searches and making it easier to describe and manipulate large quantities of biological data. The practical consequence of this pragmatism for a concept like 'cardiac cell differentiation' is that there can be no distinguishing between cardiac cells from different species, or different biological contexts. Cardiac cells form a single class with common properties, and all cardiac cells are formed from an essential process – 'cardiac cell differentiation'.

Only the rules of ontological realism dictate that there should be mutually exclusive ontologies describing different aspects of reality. This is a design choice with benefits which restricts the kind of concept admissible to the ontology, and limiting the potentially diverse ways molecular biologists may wish to describe different kinds of cardiac cells or differentiation in different biological contexts. Pluralism as a way of representing biological knowledge is avoided, and the pragmatic goal of a single, coherent ontological schema for instantiable entities is thus realised.

The advantages of choosing a different pragmatic goal in reference to creating a single classification for all molecular systems in any species will be discussed in later chapters. A different pragmatic aim here could be to reflect the diversity of conceptual understandings across the molecular biology, and this alternative goal could yield different kinds of benefits.

Thirdly, pragmatism as a theory for concepts in molecular biology is exemplified by the measuring of the value of research. Most biologists might resist the claim that some research is more important than its counterparts, yet national science policies and the distribution of research funding does prioritise some research goals over others. Recent trends in biology such as 'translational research', which seeks to drive practical benefits for society from basic research, is based on the notion that research must have a demonstrable impact on the lives of ordinary people.

The concept of 'cardiac cell differentiation' is as much a fundamental biological process as it is a potential therapeutic target in diseased organisms. Stem cell biology and notions that we can repair and rejuvenate our failing physical forms shapes such concepts, and a concept like 'cardiac cell differentiation' attests to the fact the heart and processes which occur in the heart are perceived as valuable objects of research.

### 3.2.4 An argument for the concept 'cardiac cell differentiation' as a boundary object

The representation of the GO term 'cardiac cell differentiation' can be modulated depending on the epistemology for concepts one may choose to adopt.

As an organisation, the GO Consortium has selected the theory for concepts I have described above as 'rationalism'. Ontological realism is a philosophical interpretation of rationalism. Concepts represented in the Gene Ontology, and in all partner ontologies in the OBO Foundry project,

therefore correspond to entities in reality. Every node in the GO graph is a representation of a class of instances in reality, be it a molecular function, a biological process or a cell component. Hypothetical and imaginary entities are not permissible in the ontology under this brand of rationalism. Nor is the Gene Ontology intended to capture the different subjective, cultural or historical representation of theoretical concepts in the molecular biology domain, as this would run counter to ontological realism.

If we consider the GO node 'cardiac cell differentiation' according to empiricism, historicism or pragmatism, we change the relationships it has to other concepts in the ontology, and accept the possibility of different definitions for the same term. Probably the most obvious manifestation of this approach would be the plurality of different representations for the notion of a cardiac or heart cell. The vast range of anatomical variation between individuals of the same species and, of course, between blood-pumping organs in different species raises the question - what is a heart? In primitive or diseased hearts, are all the pre-formed or mal-formed cells composing these organs manifestly heart cells?

The analysis presented in this chapter is evidence that there exists a strong argument for accepting that every single GO term in the Gene Ontology can potentially be considered as a *concept* rather than an abstraction of pure instances in reality. Paradigms in the molecular biology domain, grounded in the ways scientists generate data, the historical and cultural milieu in which they work, and the intended goals their research aims to achieve, govern variations in conceptual understandings for GO nodes like 'cardiac cell differentiation'. I assert that a result of my concept analysis is to show that even though a GO node may satisfy the condition of instantiability, it cannot exhibit context-independence and value-independence.

In order to make good controlled vocabularies for indexing gene products across the domain, it is therefore imperative to try and capture the differences in meaning evoked by adopting different epistemologies for concepts.

Star and Griesemer [215] developed the idea of boundary objects in a 1989 paper drawn from their work in the science and technology studies domain . A boundary object serves to bridge disparate communities of practice, acting in information terms to facilitate communication and representation, even when different groups of people are thinking about different things. The example Star often cites is of a road map, which to some people may show the way to a pleasant camping ground, whereas to the geologist it shows the living history of rock formation, or to the ecologist indicates potential habitats and ecosystems for different species [216].

My conclusion to this concept analysis is that we can think of the Gene Ontology as acting like a boundary object in the molecular biology domain. It offers a map of the knowledge about molecular biological systems and how they combine and interact to form a complete cell. As a boundary object, the vocabulary serves a purpose in the larger information structure of molecular biology, indexing gene products and facilitating data integration across different species databases. This concept analysis offers evidence that, depending on whatever theory for concepts one may choose, individual nodes in the GO graph can be flexible; they bridge alternate conceptualisations for ideas in the domain.

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. Yet the Gene Ontology serves a larger purpose, as a boundary object bringing together pluralities of meaning for scientific concepts across the domain and enabling information retrieval and data integration from resources about gene products, regardless of their source.

Whether you be a yeast expert, a human geneticist, a mammalian physiology researcher, a proteomic bioinformatician, a cardiovascular physician or biotechnology expert in artificial hearts, the Gene Ontology acts as a boundary object drawing these different viewpoints and understandings into a single, concrete, reified vision. The Gene Ontology is an alliance between a system of users, a standard for molecular biology. Yet it is only one form of a standard, constructed on ontological realism and necessarily ignoring the plastic and creative ways individuals from different communities of practice, thinking according to different scientific paradigms, can take that standard and use it as needed.

The Gene Ontology fails to account for the fact that it is a boundary object, and that a commitment to ontological realism is a necessary compromise in order to become a scientific sort of boundary object. The problem with the ontology is derived from its power as a standard in molecular biology. In creating an information infrastructure that can index gene products in any species database, all the other interesting, dynamic and most importantly *innovative* ways in which scientists normally think about scientific concepts, are demoted and marginalised. At its simplest level, the Gene Ontology does not even permit the incorporation of hypothetical concepts to describe gene products into the vocabulary, and this is a serious sticking point if the ontology is to make *new* discoveries and grow *new* knowledge.

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 thereby limited.

## 3.3 Analysis of Gene Ontology vocabulary construction standards

In the next results section, an information history approach explores the current Gene Ontology vocabulary construction guidelines. These guidelines articulate a perspective on gene product information in the molecular biology domain, and how that perspective is indelibly tied to the technical work and research goals of the modern biologist. Gene Ontology guidelines are compared to an existing international standard for the construction of controlled vocabularies, the NISO Z39.19. Deviations from this standard are explained in terms of the specific information needs of the molecular biology domain, and these explanations are cast in the context of the specific technical, social and economic climate of the wider e-science infrastructure.

### 3.3.1 Current GO rules

The Gene Ontology, despite those 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. Of interest though is the fact that despite the existence of international standards for vocabulary construction, the Gene Ontology has never made any reference to these standards and, since the project's inception, has always favoured the development of internal rules.

The next chapter provides an overview of the major rules guiding expansions of, and changes to, the Gene Ontology as it stands in 2011, and how these rules relate to the philosophical principle acknowledged by GO developers to be paramount to the creation of effective ontologies for biology, the principle of *ontological realism*.

The main rules of an important and long-standing international standard for controlled vocabulary construction are outlined, and an overview of how these key rules contribute to the creation of useful vocabularies is given. The standards governing the Gene Ontology will then be compared against these key rules, to ascertain whether GO has indeed met these standards through its programme of organically creating its own rules, and to assess why, when GO deviates from the international standard, differences exist.

In concluding this chapter, it is questioned whether there are requirements for special classifications and controlled vocabularies in the domain of biology which merit the modification or elimination of long-standing standards in vocabulary construction. To some extent, efforts by the Gene Ontology developers to be objective in their development of the vocabulary, and aspirations towards universality across the molecular biology domain, are judged to be reasons for certain ontology rules.

#### *3.3.1.1 Overview of rules covered*

The current, stated rules for additions and edits to any of the three ontologies which comprise the Gene Ontology are published electronically to the Gene Ontology Documentation webpage [217]. This page provides links to several important sub-sections including an FAQ and introduction to the Gene Ontology, guidance on the ontology itself, methods and rules for annotation (indexing gene products with Gene Ontology terms), standards for the GO databases and associated file formats, plus other sources of accessory information.

For the purposes of this discussion, I will concentrate on the information in the 'Ontology' section, which explains the structure, relations and conventions used to construct the Gene Ontology. Since

the Gene Ontology project is constantly being changed and updated, it is important to note that GO rules are a live document and subject to revision. Part of the purpose of this discussion is to explain how and why revisions are made, and I will explore the origins of these changes in the recent history of the GO Consortium. I will not be covering standards for GO annotation which cover both best practice for manual annotation by real people and automated annotation by computer methods. Nor will I be dealing with the technical aspects of the various GO file formats and database structures, many of which are unique to the GO project and would merit research attention in their own right.

The key points covered in this overview of the Gene Ontology are as follows:

- GO is a graph rather than a hierarchy
- GO is three separate ontologies
- Each term has defined fields
- Cross products link to different ontology terms
- GO has created several different types of relations
- Specific rules exist governing the contents of each of the three GO ontologies
- GO provides specific guidance on special topics

#### *3.3.1.2 GO is a graph rather than a hierarchy*

GO describes itself as a controlled vocabulary, and in common with all normal controlled vocabularies, terms are used to index and retrieve information. The reader should bear this in mind; although the structure of GO and the technical nature of its content may be unfamiliar to the non-biologist, it is in fact much like any other controlled vocabulary in any other domain and, as such, is amenable to investigation using principles and theory the LIS professional will understand.

Structurally, GO adopts what it describes as a directed acylic graph (DAG) for terms and relations [214]. This term is mathematical in origin, and means that each vertex in the graph is connected via edges to other vertices in the graph. A rule of directed acyclic graphs is that it is not possible to start at a vertex and find a path of edges back to this original vertex. There are no cycles. This is a logical constraint on the ontology structure which facilitates reasoning across the graph, and precludes the possibility of tautological arguments embedded in the ontology.

DAGs are useful to describe the flow of information through a series of processes in a network directed towards some outcome. Biological systems take this kind of graph structure – a cell is an information processing unit passing biological data through various cellular modules to achieve a goal, like cell division or movement. Various file formats exist to specify the syntax and semantics for describing the Gene Ontology graphs, and the most important is the OBO flat file format [218]. In simple terms, the OBO file format is a way to represent classes, properties and individuals, which correspond in GO to terms, relations and instances. OBO also captures synonyms, cardinality constraints and axioms relating these different entities.

The main application for the Gene Ontology was always intended to be information retrieval across multiple species databases. Historically, these databases stored gene product information in a myriad of formats and structures which made it difficult to integrate gene data from different sources. The graph structure of GO is designed to facilitate this main application. Classes can be

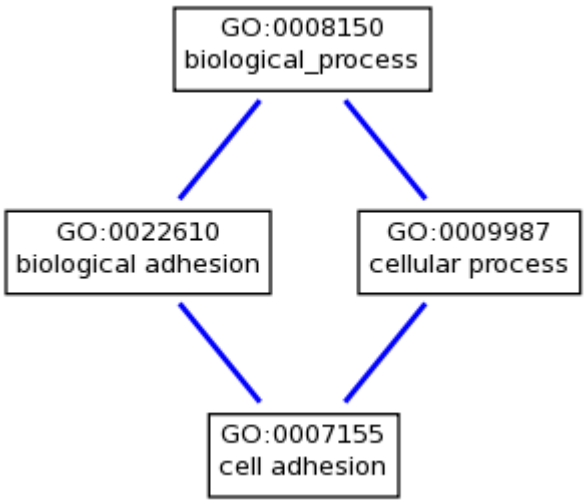
grouped and matching annotations found in any GO-compliant database, regardless of how the gene product information is represented in external databases.

In this sense, the graph structure of the Gene Ontology is like Medical Subject Headings (MeSH) which offer a means to collapse or explode searches as needed. 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. An example is given below.

The important feature to note is that any term can have multiple locations in the ontology. One biological process for example may have numerous direct parent terms offering different paths back to the root of the Biological Process Ontology. Traditional controlled vocabularies restrict terms to only one parent, much like the MeSH vocabulary, yet a synonym in GO may have more than one arc to represent its relationship with the root node.

In the Figure 6 below, the GO term 'cell adhesion' has two parents, 'biological adhesion' and 'cellular process'. The same term in MeSH has only a single parent 'Cell Physiological Processes'.

**Figure 6: GO graph view for GO:0007155, 'cell adhesion'**




#### *3.3.1.3 GO is three separate ontologies*

The Gene Ontology is in fact three distinct and unconnected vocabularies. Each has its own root node, each is described as covering its own 'domain', and there are no formal linkages created between these three 'sub-ontologies'. The three sub-ontologies are:

- The Cellular Component Ontology which "...describes locations, at the levels of subcellular structures and macromolecular complexes"
- The Biological Process Ontology which is "...a recognized series of events or molecular functions"
- The Molecular Function Ontology which are to a gene product "...the jobs that it does or the "abilities" that it has"

GO documentation is relatively silent on why three separate sub-ontologies are a necessity, and the assumption is that users will clearly identify cell components, processes and functions as three

distinct areas of knowledge. The Gene Ontology has designed separate rules guiding the inclusion of terms in each respective sub-ontology, and authors ought to make it clear which source ontology they are using in a data analysis like microarray data interpretation. However, by GO's own admission, it is difficult to distinguish a biological process from a molecular function, and cellular locations cannot be entirely kept from the other sub-ontologies. A biological process is defined as a series of molecular functions, although the Gene Ontology has never made any tools or sources available defining exactly *which* functions in the one ontology combine to manifest the series of molecular events recognisable as a biological function.

#### *3.3.1.4 Each term has defined fields*

Every term in the Gene Ontology, regardless of which sub-ontology it may occupy, shares a number of common features. Every term has a unique identifier of the form 'GO:xxxxxxx' where 'x' is an integer. Terms have names, and are labelled with a 'namespace' specifying which sub-ontology root it is associated with. Terms must have a text definition, and all terms have relationships to other terms. It is these characteristics which determine XML and database fields in several different ontology formats and files used by biologists in bioinformatics applications.

Terms may not share identifiers, and most terms have unique names. Historically, there are examples of terms in the different sub-ontologies sharing the same name; this was not viewed as problematic because identifiers were unique. Some definitions for terms are drawn from canonical dictionaries and books in the biosciences domain (and are referenced appropriately). However many definitions are authored by GO curators or domain experts, as indicated by short letter codes in the ontology files. Ontology editors with the appropriate privileges can change term names and re-write definitions as required. How these changes are implemented, and the role of the biological community in effecting these changes, is explored in more detail in the chapter on discourse analysis, where results explore the social dynamics behind changes to the ontology.

GO terms may have synonyms, and these include exact, broader, narrower and related. Synonym coverage is by no means comprehensive, and many synonyms are created automatically using software guided by simple semantic rules. Different GO terms deemed to be conceptually identical may be merged together, and the identifier for the replaced term becomes known as a 'secondary ID' to the surviving term.

GO terms can also be obsoleted. Obsoletion is a process reserved for a term which is "...outside the scope of GO, is misleadingly named or defined, or describes a concept that would be better represented in another way". Obsoleted terms are not deleted; rather they are labelled as 'obsolete'. This is important, as databases which annotate to terms which are later obsoleted must be able to track these changes.

How terms are obsoleted and why they are removed from the ontology is the subject of Chapter 3.5 .

#### *3.3.1.5 Links to other ontologies and vocabularies*

When the Gene Ontology was released, it was one of the first ontologies in the molecular biology domain. Now, there are many hundreds of different ontologies describing different entities in biology, such as anatomical structures, cell types and elements in experimental setups, and it is possible to link from the Gene Ontology to these other ontologies.

Progress on linking the Gene Ontology to this other vocabularies has been relatively slow. In 2001, Mungall *et al.* published a proposal for creating formal cross-products between the Gene Ontology and related ontologies. A small group of GO developers are now working on implementing these cross-products, and for example have established preliminary linkages between the Gene Ontology and the Cell Type Ontology, to permit relationships of the form such as:

*[Cell Type Ontology term] + [Biological Process Ontology term]*

*[epithelial cell] + [cell proliferation]*

The move towards formalising cross-products is a move towards a more facetted approach for constructing ontologies, but represents a major challenge for restructuring the Gene Ontology. As it stands in 2012, there are as yet still no cross-products even between the respective sub-ontologies within GO.

On the other hand, the Gene Ontology has long provided mapping files to other controlled vocabularies. These include to Enzyme Commission numbers, Interpro protein families, and the Uniprot knowledgebase (formerly Swissprot keywords). Mapping files list Gene Ontology keywords and their correspondents in these partner vocabularies, and mappings are either created manually, or via semi-automated methods.

It is interesting to consider precisely what the relationship between the Gene Ontology and these other vocabularies actually is. Many mappings are between GO terms and keywords which would be inadmissible according to GO standards, such as Molecular Function Ontology terms mapping to classes of gene products. However the GO Consortium still validates these connections by providing mapping files. If these mappings were considered to be synonyms, it would undermine many of the rules determining the content of terms in the Gene Ontology.

The GO Consortium provides no definition for a mapping relationship, and simply requires that should results be published using mapping files, both the Gene Ontology files and mapping must be cited appropriately.

#### *3.3.1.6 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 [36, 73, 78, 80, 186, 219]. 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. A relationship is valid if it describes a connection between two terms which is logically true.

The Gene Ontology uses three broad types of relationships:

- *is_a*
- *part_of*
- *has_relation*

For example, the 'is_a' relationship designates that A is a sub-type of B. The child class is subsumed by the parent class, and if A *is_a* B, then all type Bs are necessarily also of the type A. GO treats these

classes as Aristotelian kinds, with essential properties creating nested hierarchies of classes. For example, if we were to create an ontology of the utensils found in a kitchen and stated that the class 'spoon' *is_a* 'cutlery', we are stating that 'cutlery' forms a super-class which subsumes all child classes like 'spoon'. All spoons are therefore a kind of cutlery, and any object in a kitchen we deem to be cutlery must share the essential properties of the 'cutlery' class. Cutlery might be defined as any utensil we lay out on the table for an individual to use during a meal, such that teaspoons and dessertspoons, being members of the class 'spoon' are kinds of cutlery. However wooden stirring spoons could not be children of the class 'spoon' if the parent class were 'cutlery': this would the break the truth of the *is_a* relationship.

Another important relationship is the 'part_of' relationship. 'part_of' creates a subtle distinction from 'is_a' in that to state B *part_of* A, is to imply that the B is necessarily a component of A, whereas the presence of A does not necessitate the presence of B. For example we can state in ontology terms:

'flour sifting' *part_of* 'cake baking'

Here we are saying that the class 'flour sifting' has a 'part_of' relationship to the super-class 'cake baking'. All flour sifting events imply that a cake is being baked. However the *part_of* relationship means that the converse is not necessarily true: I don't need to sift flour in order to bake a cake.

GO also encodes the 'has_relation' type of connection between terms. This 'has_relation' can be positive or negative in application, and broadly speaking suggests how a term such as a biological process may regulate another process. Therefore we can say in GO terms that:

'pedalling' *has_relation-positive* 'bicycle velocity'

This ontology relationship suggests that increases in pedalling rate will act positively on bicycle velocity.

The inclusion of the three main types of relationships outlined above allows the Gene Ontology to encode a wealth of theoretical statements and knowledge from the field of molecular biology – decomposed into the ontological domains of cell components, biological processes and molecular functions. Ontological relationships aid in computer-driven reasoning across the three parts of the Gene Ontology: computer applications can be created which make inferences based on these logical relationships. This is enormously useful for information retrieval purposes, as one can ask questions across the GO graph like 'Find me all gene products involved in limb development' and users can search for annotations at all levels of the GO tree on the understanding that all paths in the ontology are true.

A trivial example might be a process 'ringing a bell' which positively regulates the process of cake baking. If 'flour sifting' has the *part_of* relationship outlined above, a computer might infer that 'ringing a bell' will also positively regulate the sifting of flour in a kitchen. The converse would not be true though: sifting flour would not make the bell ring more, and a computer programme would understand this to be the case.

#### *3.3.1.7 Specific rules exist governing the contents of each of the three GO ontologies*

I will now deal with the more specific rules governing the structure of the three sub-ontologies, and the intended scope of these vocabularies.

##### 3.3.1.7.1 Cellular component

The Cellular Component Ontology is a vocabulary describing the parts of cells and their relationships to one another.

The Cellular Component Ontology does not describe cellular structures at the level of individual proteins. However, it does represent protein complexes, where a complex is a structure composed of more than one protein molecule joined non-covalently to another. This ontology is also restricted to unicellular structures: it does not describe multi-cellular anatomical structures such as bones or kidneys.

All terms in the Cellular Component Ontology are required to meet several simple conditions, to ensure logical consistency in the ontology structure. Every cellular component must have an 'is_a' relationship path back to the root of the ontology because every entity represented by this ontology is a child of the 'cell' node beneath the ontology root. The ontology is therefore a model of a cell – all component terms contribute, via 'is_a' relations, to this idealised cell. In addition to this requirement, all Cellular Ontology terms must have a 'part_of' relation to another Cellular Component Ontology term, because the presence of a cell component means that it must be part of some larger structure which contributes to the entirety of a cell. Cellular Component Ontology terms without 'part_of' relations can become orphaned without links to any other cellular structure – they are cell components without being a part of a cell, which is logically inconsistent.

The Cellular Component Ontology is used to represent cell parts from any species. Therefore it has cell wall-specific terms to represent structures in plants and bacteria, even though eukaryotic cells (such as human cells or mice cells) do not have cell walls. It also contains terms designed to represent gene products that normally lie outside a cell, in extracellular spaces. Signalling molecules like hormones or factors which facilitate interactions between host cells and infective particles are captured by these kinds of Cellular Component Ontology terms.

##### 3.3.1.7.2 Molecular function

The Gene Ontology considers a molecular function to be the 'ability' a gene product has to perform a job in the cell. Functions are distinguished from biological processes (see below) by virtue of the fact that a process consists of more than one molecular ability or activity. Molecular functions are, in a sense, atomic – they are indivisible functions.

As the GO guidance notes state "...it would be wrong to create a function term that represents multiple functions". GO indicates that information about gene products with several functions should be captured at the annotation stage. Gene product information mistakenly coded into to the Molecular Function Ontology is often identifiable by terms with multiple parent functions. These parent functions are not obvious from the term name or definition, and demand supplementary domain knowledge. GO rigorously removes these kinds of gene product terms from the Molecular Function Ontology because they cause serious problems with reasoning across the graph, and introduce other sources of error pertaining to species-specific functions.

However it is worth noting at this point that the notion of an indivisible function in biology is not immune to criticism because arguably almost any function can be decomposed into further 'sub-functions' (this will be covered in more detail in later chapters).

GO divides molecular functions into four broad categories for the purposes of editing the Molecular Function Ontology. Function terms may describe binding events (where a gene product interacts with another molecule), enzyme activities (where a gene product catalyzes a biochemical reaction), receptor activities (where a gene product interacts with a target to initiate a change in a cellular process) and transporter activities (where a gene product engages in the movement of a target molecule).

Molecular functions are further defined by various other criteria. GO deals with molecules, not atoms, and nor does it describe spontaneous events occurring without a gene product (such as the spontaneous release of an ion). It attempts to capture knowledge about biochemical reactions, but only if intermediaries are released or catalytic subunits can clearly be identified with a specific activity. These rules help ontology editors in distinguishing between processes and functions. These rules do however exclude many enzyme entries found in the Enzyme Commission database, because these entries describe enzymes with more than one function by the GO philosophy.

The Molecular Function Ontology avoids cell component information, activities which occur simultaneously but are not dependent on one another, removes gene product entries (or appends 'activity' to gene product names to create a surrogate function for that gene product), controls for regulatory sub-units in enzyme complexes, does not consider catalysis to involve binding (for practical reasons), and does not group functions because they are involved in common processes.

The Biological Process Ontology serves this final function.

#### 3.3.1.7.3 Biological process

Given the extent to which the Gene Ontology guidance extends for molecular functions, the notes on terminology in the Biological Process Ontology are surprisingly brief.

Biological processes are defined as "...a recognized series of events or molecular functions [...] with a defined beginning and end". Definitions must include where the process begins and ends. For example, the GO biological process term 'heart development' is defined as the "...process whose specific outcome is the progression of the heart over time, from its formation to the mature structure".

The Biological Process Ontology permits both examples of complete processes and collections of processes. The former will have children related to the parent by 'part_of' of relationships, such that 'heart development' is decomposed into 15 different sub-processes each with a 'part_of' arc to the parent term. An example of a collection of processes might be 'systems development', whereby terms such as 'digestive system development' are related to the parent by an 'is_a' relationship.

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. As of August 2011, the Biological Process Ontology consisted of 21338 terms, whereas the Cellular Component Ontology and Molecular Function Ontology described 2892 and 9067 terms respectively.

#### *3.3.1.8 GO provides specific guidance on special topics*

The Gene Ontology also provides guidance on special biological topics such as the cell cycle and metabolic processes (see Table 10 below). These guidance notes 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.

Table 10: GO guidance on special topics

| Topic | Comments |
|---|---|
| **Cell Cycle** | Split into physical processes and temporal phases (these are not linked in the ontology); standard definitions are used to describe the phases; special note on 'cytokinesis' explains why it is not a child of 'cell cycle' (as most biologists might expect) |
| **Development** | Extensive notes on standard terms for describing the development of cells, tissues and organs. Ontology is not considered to be complete, and missing terms are 'implied'. Notes list standard form for definitions, and explain the GO treatment of closely related terms like 'development', 'differentiation', 'formation', 'morphogenesis' and 'maturation'. Also, cell types in different organs are given separate terms, giving terms for epidermal cells in hearts and epidermal cells in kidneys |
| **Metabolic processes** | GO distinguishes between metabolism at the level of the multicellular organism, and metabolism at the level of the individual cell. Many standard synonyms for these metabolic processes are suggested, together with qualifiers ('in the presence of oxygen') and complex terms to simplify the construction of additional specific processes |
| **Other organisms and viruses** | Guidance created in response to a need to create terms describing symbiotic or parasitic interactions between species; these term were also resolved with pathogenic interaction terms. GO creates terms relevant to both the host and the symbiote/parasite/virus which is performing the interaction. Guidance exists on how to annotate locations for the respective species |
| **Regulation** | The meaning of 'regulation' is defined. In GO terms, any regulation can be positive or negative, and regulations only occur in processes and functions, rather than with objects such as specific proteins – only activities are regulated. Other stages for regulation terms are recommended, including 'activation', 'inhibition' or 'termination' |
| **Response to stimulus** | Describes standard structures for terms describing responses to stimuli, such as molecular signals. GO describes detection phases, and various responses such as behavioural or cellular changes |
| **Sensory perception** | Restricted to the neurophysiological responses normally understood as 'The Five Senses'. Describes how to construct ontology relations for the transduction of sensory signals into cell processes (such as pain from a hot iron converted into electrical nerve impulses) |
| **Signalling** | Several standardised definitions for signalling terms |
| **Transport and transporters** | Covers terms used for "...the processes involved in positioning a substance or cellular entity". Split into the establishment of a location and the maintenance of that location. A standard ontology structure and definitions is given for transport terms that can used for any substance or cell component. |

### 3.3.2 LIS standards for vocabulary construction

Despite this commitment to ontological realism, the Gene Ontology still identifies itself as a controlled vocabulary, representing a biological knowledge in controlled language after the manner

of a thesaurus. 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 manner of* a controlled vocabulary, to index gene product entries in various species databases. Therefore there is a strong justification for comparing those standards used to construct the Gene Ontology against existing, internationally recognised standards for vocabulary construction.

Various authors in the biosciences domain have suggested requirements and improvements for ontologies representing biological knowledge [63, 67, 79, 82, 220-223]. The working group of the cancer Biomedical Informatics Grid (caBIG) proposed criteria for evaluating terminologies in biology and medicine [113] and found that GO largely performed well when compared to other vocabularies. However like many authors auditing terminologies in the biosciences domain, no reference is made to existing standards for controlled vocabulary construction created by professional LIS bodies.

The ANSI/NISO standard Z39.19, 'Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies' "...presents guidelines and conventions for the contents, display, construction, testing, maintenance, and management of monolingual controlled vocabularies" [224]. It acts as the US equivalent to the 1986 edition of the International Standard ISO 2788, 'Guidelines for the establishment and development of monolingual thesauri' and the British version, BS 5723. For more details on the relationships between these standards, see [128].

NISO Z39.19 has been developed by LIS domain experts, approved by the NISO members, peer reviewed, and verified by the American National Standards Institute. As a standard, it is a high quality series of recommendations for principles and rules that, if adhered to, ought to permit the creation of high quality vocabularies that will consistently index knowledge objects and facilitate information retrieval.

My assumption therefore is that the NISO Z39.19 standard is a relevant standard to hold the Gene Ontology against. Although the Gene Ontology was not created using any such standard, it is still a form of controlled vocabulary and, as a controlled vocabulary, the internal rules created by the GO Consortium ought to correspond to the recommendations of Z39.19. The principle of ontological realism does establish an important difference between how Z39.19 and the Gene Ontology approach terms. NISO Z39.19 states that "...a term is defined to be one or more words used to represent a concept" meaning that it assumes terms to stand for mental concepts in the minds of thinking users. The Gene Ontology understands terms as "...representing gene product properties" or objects in reality, not in the mind. Part of my argument in this thesis is that the two cannot be separated: the structure of reality and the different ways we can discover to describe it in our minds.

NISO Z39.19 distinguishes between primary content objects and secondary content objects. Primary content objects are that which is being described, and may be a document or work. In the case of the Gene Ontology, primary content objects are gene products (for example, proteins). These exist in physical forms within living individuals, and in electronic forms as sequences in genomic databases. Secondary content objects are the metadata associated with primary content objects, and examples in the context of the Gene Ontology are the names of genes, the species they are expressed in, and genomic locations. In biology, databases of electronic sequences create hybrids of primary content objects with the secondary content objects, or metadata, describing them.

I intend to use Appendix A of the NISO Z39.19 to investigate whether GO standards address several core principles in vocabulary construction. Most of these principles are recommendations rather than requirements, and raise the following questions about the Gene Ontology's own standards for maintaining its controlled vocabulary:

- Ambiguity (section 5.3.1): How does GO ensure that terms have one and only one meaning?
- Synonymy (section 5.3.2): Does GO guarantee that each concept has one and only one preferred term?
- Using warrant to select terms (section 5.3.5; section 6.6.1): What warrant does GO use to select terms?
- Choice of terms (section 6.1): How does GO choose terms for inclusion in the vocabulary?
- Scope notes (section 6.2.2): How does GO state the chosen meaning of terms?
- Compound terms (section 7): How does GO handle compound terms?
    - Factors to be considered when establishing compound terms (section 7.3)
    - Criteria for splitting compound terms (section 7.6)
- Relationships (section 8): How does GO control relationships between terms in the vocabulary?
    - Equivalence relationships (section 8.2)
    - Hierarchical relationships (section 8.3)
    - Associative relationships (section 8.4, especially 8.4.2)

### 3.3.3 How GO rules were created

#### *3.3.3.1 Ambiguity*

"Ambiguity occurs in natural language when a word or phrase (a homograph or polyseme) has more than one meaning [...] A controlled vocabulary *must* compensate for the problems caused by ambiguity by ensuring that each term has one and only one meaning." p.13, [224]

##### 3.3.3.1.1 Homographs

Homographs are words which share a written form and potentially have the same pronunciation. They may also possess a common origin yet homographs are distinguished by their conveyance of different meanings. For example the word 'mouse' can either mean a small rodent organism, or a control device connected to a computer. The biologist with a professional interest in murine biology needs to be able to determine the difference between these two meanings for the word 'mouse' when it is used in either sense to index documents in an information retrieval system, such as an Internet search engine.

Information systems which use controlled vocabularies to index content must control for the ambiguity problem caused by homographs, in order to distinguish between the different meanings associated with the same word. The Gene Ontology, despite identifying itself as a knowledge representation of entities in part of reality dealing with molecular biology, cannot ignore the problems created by homographs. The ontology uses words to identify each node in the graph, and these words are subject to the same vagaries of interpretation associated with homographs in any other controlled vocabulary.

To a large extent, the Gene Ontology circumvents the problem of homographs by a 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 Gene Ontology and refers 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. GO scope is well defined and limited to the description of entities in molecular biology. In indexing, retrieval, and information management tasks, the Gene Ontology does not crossover into the domain of physics, and *as yet* does not need to compensate for the ambiguity caused by homographs for a word like ‘nucleus’. However, that is not to say that the Gene Ontology will never encounter this problem in the future. Reductionism in the sciences implies that even molecular biology might be reducible to the principles of atomic or sub-atomic physics, meaning that eventually domain-specific vocabularies such as the Gene Ontology will be forced to control for ambiguities in natural language.

Potentially problematic homographs litter the Gene Ontology, and many have more than one meaning in the sciences. For example, the word ‘compound’ occurs frequently in the GO vocabulary, mostly in the context of biochemical structures and their movements in the cell, but also in the sense of anatomical structures such as the compound eyes found in insects. Both senses are applications of ‘compound’ as a noun, and because the ontology never uses ‘compound’ as a verb (meaning ‘to combine things together’), GO never qualifies individual instances of this string.

Many GO terms contain homographs which themselves 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 phrase which is never decomposed into its constituents in the context of the GO vocabulary. If the Gene Ontology were extended to the domain of biological psychiatry, events at the molecular level in the cell may have the downstream effect of modifying the happiness and psychiatric well-being of a whole organism, and it may prove necessary to disambiguate synaptic and psychiatric depressions.

Equally the exact synonym to GO:0035810, ‘increase in urine flow’ does not qualify that ‘increase’ is assumed to be a noun meaning ‘to make urine flow become larger’ since the parent term ‘positive regulation of urine volume’ cannot be interpreted as a verb. In fact, it is stipulated in the GO documentation on biological regulation that for biological process terms, the noun form ‘positive regulation of’ is preferred in situations where one process increases “...the frequency, rate or extent of [another] process”. All processes are abstract nouns describing states of being in biological systems rather than actions being performed on objects.

However the ontology does struggle to compensate for the problems caused by certain homographs found in exact synonyms. Synonyms in GO are normally created either automatically, on an *ad hoc* basis, or when obsoleting terms (an obsolete term string will often become a synonym for a new term with a new definition). Exact synonyms inherit definitions from their parent term, yet without this definition, many GO terms are potential sources for confusion.

An interesting case in point is the word ‘effect’, a common homograph found in the English language. As a noun it can mean the result or outcome of some process, whereas as a verb it means to bring about or make (‘to effect a change’).

In prokaryotic biology there exists a process known commonly as the glucose effect, whereby glucose inhibits the activity of specific metabolic pathways in bacteria. Since 'glucose effect' is not an especially descriptive term for non-prokaryotic specialist and was inconsistent with the structure of comparable terms, GO:0045014, 'negative regulation of transcription by glucose' was added to the Biological Process Ontology in September 2002 with the synonym 'glucose effect'. However to the naive user it may not be clear that 'effect' is used in the sense of a noun ('the outcome produced by the action of glucose'), or in the sense of a verb ('to bring about a change by the action of glucose').

Likewise, the word 'import' is a homograph, and can be either a noun ('a thing brought in from somewhere else') or a verb ('the process of bringing a thing in from an outside source'). Thus there are various compound terms in the Biological Process Ontology which are types of imports such as 'lysine import' or 'glucose import'. From the scope of the ontology and interpretation of the term definitions it is possible to infer that GO treats 'import' as a verb.

For example, GO:0017038, 'protein import' is defined as 'the directed movement of proteins into a cell or organelle' whilst GO:0032975 'amino acid import into vacuole' is defined as 'the directed movement of amino acids into the vacuole'. These definitions are verbal treatments of the word 'import', yet one can see how a naive user could easily interpret a gene product indexed with the term 'protein import' to be 'a type of protein which has been imported from somewhere' rather than 'a gene product responsible for moving proteins into a location'.

This example also displays a major problem of definitions in the Gene Ontology in that many are circular, using words from the term string to define the term. Users must understand what is meant by 'amino acid' and 'vacuole' if they are to comprehend the definition for 'amino acid import into a vacuole'. From the guidance provided by the NISO Z39.19 standard, poor definitions with circular forms only serve to increase ambiguity in the vocabulary.

#### 3.3.3.1.2 Polysemes

Polysemes differ from homographs in that the different meanings of a word or phrase are in some sense *related*. Although there is no formal test of polysemy, a word's etymology may indicate a common root. However this is by no means an absolute criterion for polysemy.

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 Gene Ontology 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".

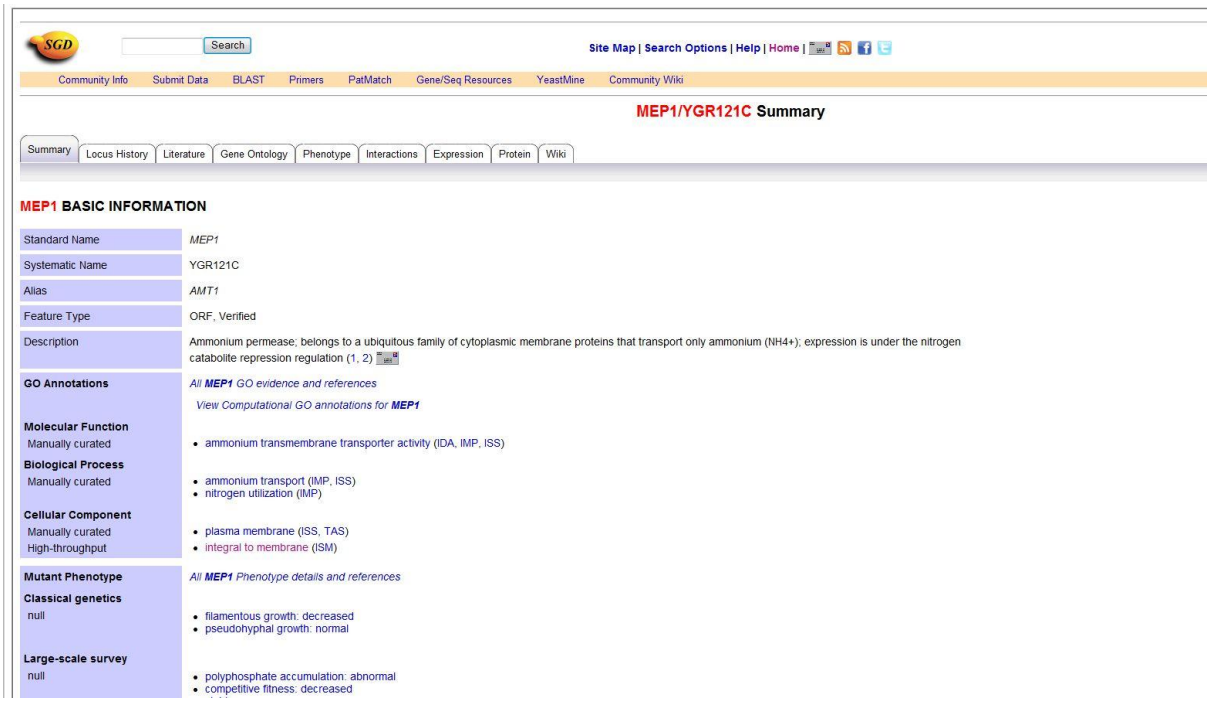
However the word 'membrane' has many more connotations in the biosciences domain. There are the mucosal membranes which line the cavities and canals of anatomical structures in the body, or the foetal membranes which surround and enclose the developing foetus. These types of membranes are not controlled for in the ontology since they fall out with the scope of the Gene Ontology. 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 GO is that in the Cell Component Ontology alone, as of November 2011, 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.

GO terms are frequently presented in information systems without definitions, for example when displaying the index terms for gene products in a species database. In Figure 7 below, taken from the SGD website page describing the yeast gene *MEP1*[1], a gene product is described with GO Cell Component Ontology terms 'plasma membrane' and 'integral to membrane'. From the context of the gene product and these two terms together, one can infer that 'integral to membrane' is likely to mean integral to a phospholipid bilayer membrane, rather than any other type of biological membrane. This disambiguation is only possible by searching the GO terminology for an explanation of what sense the word 'membrane' is being used. Even though the Gene Ontology does provide various alternative qualifiers for products integral to membranes, such as 'integral to plasma membrane' or 'integral to Golgi membrane', gene products annotated to general, higher level terms for reasons of insufficient evidence are inevitably ambiguous in their function because polysemes are poorly controlled.

**Figure 7: SGD database online record for the gene product MEP1**




[1] http://www.yeastgenome.org/cgi-bin/locus.fpl?dbid=S000003353

#### 3.3.3.1.3 The species-specific problem

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' [225]. 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, had different meanings when mammals were compared to insects. It was felt that the processes, although sharing the same name, were so different in their action when observed in reality 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. The GO term GO:0048477 'oogenesis' therefore had two related synonyms:

Oogenesis (sensu Mammalia)

Oogenesis (sense Insecta)

For some members of the GO Consortium, many GO terms were essentially 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 (as of November 2011) 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. The Mycobacterium tuberculosis for example invades cells and produces kinases which enter the host cell nucleus and hijack nuclear processes. The GO term 'host cell nucleus' is a species-centric qualified cell component, added after extensive discussions around the importance of describing host-pathogen interaction (GO Mailing List, June 2001).

#### *3.3.3.2 Synonymy*

> "A controlled vocabulary *must* compensate for the problems caused by synonymy by ensuring that each concept is represented by a single preferred term. The vocabulary *should* list the other synonyms and variants as non-preferred terms with USE references to the preferred term." (p.13, [224])

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 traditional 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:

> "...the term "synonym" is used loosely for any kind of alternative label for a class (the "name" tag is used for the community preferred label)." [218]

NISO Z39.19 states that true synonyms are rare in natural language (p.56, [224]), with synonym usage being highly dependent on context, such as between professional and lay circumstances. 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 OBO file format approaches exact synonyms quite differently to NISO Z39.19, and considers true synonyms for class names to be appropriate "...if there exists some user or user community (existing or historic) for which this label unambiguously denotes the class" [218].

Preferred terms are indicated in the Gene Ontology 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 Gene Ontology 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 exact synonym 'outer membrane of cell' is a lexical variant, and the ontology strives for consistency in these variants both in structure and semantics (for example, preferring the American spelling over the English). What the ontology fails to account for in this example is that a cell outer membrane is commonly understood in biology as a part of a *bacterial* cell membrane. GO usage differs from this common usage in that the string 'cell outer membrane' is used to index both eukaryotic and prokaryotic gene products, a meaning for the concept some bacterial biologists may contest.

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? Major flaws do exist in the GO approach to synonymy, and special problems are created by the Gene Ontology's decision to maintain the three ontologies as separate entities. The Consortium tends to re-use term strings in the three different ontologies, yet users searching the ontologies have previously become confused when presented with near-identical term strings for different nodes in the ontologies. An example queried on the GO mailing list from July 2001 asked why 'mRNA cap binding' and 'binding to mRNA cap' existed in the Molecular Function and Biological Process Ontologies respectively. This and other confusing uses of the same word strings to represent different concepts were resolved by re-wording and re-defining the GO terms. The term 'binding to mRNA cap', which was never defined, was replaced with the term 'mRNA capping', and any reference to the conceptual relationship between the two is written from the ontology's history.

- Molecular Function Ontology term GO:0000339, 'RNA cap binding' NT 'binding to mRNA cap' OR NT 'mRNA cap binding'
- Biological Process Ontology term GO:0006370, 'mRNA capping'

Largely, GO does satisfy this requirement of the NISO Z39.19 standard, although this statement should be qualified by the problem created by the Gene Ontology's status as an ongoing 'work-in progress'. The GO Consortium accepts that the structure of the ontology largely dictates 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.

Additionally, the incomplete nature of the ontology is reflected in terms which might legitimately be added now, but which have not been added for practical reasons of time, work involved, 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 is a sense in the GO Consortium that at some point in the future, development terms may be cross-referenced to an existing cell-type ontology [226].

GO mailing list discussions back in March 2008 mention this plan ("We also still intend to turn attention to cross-products with the Cell ontology after the regulation work is secure") and funding for work on the Cell Ontology and its interoperability was received in 2009 [227, 228]. However until the necessary cell development terms are added or cross-references from the Gene Ontology to the appropriate cell-type vocabulary are created, users are again forced to use more general higher-level terms or composites of several other terms to try and capture the conceptual sense for gene product activities which ought to have their place in the vocabulary.

The advantages of a fully facetted classification schemes for gene products as a solution to this problem of synonymy and the representation of concepts by single terms will be discussed in later chapters.

#### *3.3.3.3 Using warrant to select terms*

The NISO Z39.19 standard suggests three different types of warrant which may be used to inform the choice of preferred terms in a controlled vocabulary:

1. Literary warrant
2. Organizational warrant
3. User warrant

Literary warrant is indicated by the natural language used in a domain to communicate concepts. In the molecular biology domain, this would be the words and phrases found in scientific texts. Organizational warrant is determined by the institutions and groups which will be using the vocabulary; the language of organizations may therefore differ from the terminology identified according to literary warrant. User warrant is governed by the search phrases used by users to discover content in information systems. Search terms in species databases in the molecular biology domain may therefore give some indication of preferred terms for concepts according to user warrant.

Literary warrant is shaped by existing dictionaries and word lists in a domain, together with natural language evidence drawn from primary and secondary information sources, such as published articles and reviews. The NISO Z39.19 standard suggests that "...word or phrases chosen *should* match as closely as possible the prevailing usage in the domain's literature" (p.28, [224]). The Gene Ontology Consortium selectively ascribes to this methodology for term selection. At times it will follow canonical dictionary definitions for terms drawn from titles such as 'The Oxford Dictionary of Biochemistry and Molecular Biology' [229]. Yet since the organisation considers the Gene Ontology to be a solution for the problematic way semantics in biology has proven an obstacle to 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. A good example of this practice is found in the GO mailing lists for March 2000, where curators, dissatisfied with the term 'ubiquitin' decide to invent a new term called 'protein degradation tagging' to describe a concept.

This process of negotiation and devising of a new term is explored in more detail in Chapter 3.4, but is mentioned now to show how literary warrant is often ignored by the Consortium.

Terms which have been selected according to literary warrant are readily identifiable in the Gene Ontology according to ISBN references to textbooks and dictionaries, or PMID identifiers to source articles. Large numbers of terms though will be attributed via a source code to groups or individuals within the Consortium. For example GO:0043228, 'non-membrane-bounded organelle' carries the source code 'GOC:go_curators' meaning that GO curators devised the term string and definition for this term. A brief search for this exact phrase, 'non-membrane-bounded organelle' reveals no exact hits via PubMed (searched 25 November 2011) and even an expanded search provides only 10 hits in the entire publication database. This is an example of a failure to adopt some measure of literary warrant in designing ontology terms, since GO is indexing gene products with a term which has little correspondence to the literature.

Are preferred terms like this example therefore chosen according to organizational warrant? In a sense, GO terms are created for the purposes of the Consortium members; Gene Ontology terms are tools to cross-reference gene products in different databases. Since the project is a collaborative work between several different sub-domains in molecular biology (Yeast researchers, mice researchers, bacterial experts and so forth), 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 Gene Ontology 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. Yet as the NISO Z39.19 standard advises "...[t]erms should reflect the usage of people familiar with the domain" (p.30, [224]).

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. The front-facing role of the Gene Ontology as an organisation is viewed as important because engaging with users, and getting users to apply the ontology in their work, are crucial to the long-term survival of the Gene Ontology project. The Gene Ontology therefore tries to reflect the structure and usage of terms in the domain, and through various meetings, annotation 'jamborees' and online systems for term requests, such as Sourceforge, user warrant does play a role in the selection of terms and their preferred forms.

The Gene ontology really only applies user warrant 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, for the ontology is constructed according to strict philosophical rules and failure to adhere to these rules will often lead to rejection of new terms. A relatively simple example of resistance to user warrant which has had a massive 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 the molecular biology domain is the usage 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 Gene Ontology. 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 senior member of the Gene Ontology team even went so far as to suggest a fourth ontology to develop alongside the Process, Cell Component and Molecular Function ontologies, a vocabulary which would 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 huge subset of terms that may have been meaningful and useful to users was excluded from the Gene Ontology.

#### *3.3.3.4 Choice of terms*

> "Selecting terms for inclusion in a controlled vocabulary is one of the most important factors in creating a product that has broad user acceptance." (p.20, [224])

The NISO Z39.19 standard highlights the importance of choosing appropriate terms for inclusion in any controlled vocabulary, since term choice affects both the work of indexers and users.

The standard recommends that existing vocabularies covering the same information domain should be consulted to ensure there is no overlap or duplication of the time and effort invested in creating a controlled vocabulary. In the case of the Gene Ontology, it was determined early in the project that no universal classification for gene products spanning all potential species existed in the molecular biology domain. Developers were aware of a number of different classifications for protein functions and structures, together with various vocabularies covering species-specific content, such as classifications of plant genes. However there was no single vocabulary suitable for classifying gene products in all aspects of the molecular biology domain. In fact, much of the content in the Gene Ontology does duplicate existing vocabularies, and formal efforts have been made to map GO content to these other classifications, such as the EC classification. Developers of alternative classifications in the same domain have also been invited to contribute ideas and content for the Gene Ontology, and this too has led to duplication and eventually subsuming of smaller projects into the GO Consortium. Largely, the Consortium has used the argument that because the Gene Ontology is predicated on ontological realism, it distinguishes itself from all pre-existing classifications in the molecular biology domain which, in their attempts to offer representations of the knowledge and concepts in biology, are inherently weaker products, essentially proving to be unfit for the purpose of data integration and e-research in the life sciences.

As discussed above, the GO Consortium has no clear policy on the warrant used to select terms and the preferred form for terms, even going so far as to routinely invent new terms for functions and processes. As far as this approach goes towards "...creating a product that has broad user acceptance", the GO Consortium enjoys the luxury of having no major competitors for indexing gene products. The idea of an alternative vocabulary constructed according to user warrant is in fact nonsensical according to ontological realism as the Gene Ontology is a representation of entities in reality; there can be no 'better' vocabulary which is more like reality. The idea of an appropriate warrant for selecting terms is relevant to the purpose of the Gene Ontology, since users who apply the ontology in an effort to understand the meaning of a dataset must recognise and understand the meaning of ontology terms if they are to make inferences about that data. Users appraise the relevance of documents retrieved by an information system supported by a controlled vocabulary, and in much the same way, biologists appraise the relevance of Gene Ontology term lists produced

by a term enrichment analysis. The GO Consortium, in adopting what is essentially an organizational warrant, determined by the members of its own organisation, risks alienating users from its product.

Furthermore, the NISO Z39.19 standard offers guidance on the specificity or granularity of terms. The proliferation of highly specific terms in peripheral areas of the vocabulary risks creating a structure which is difficult to manage, and difficult to understand for users; highly granular terms should be reserved for the core domain covered. The problem the Gene Ontology finds in terms of specificity is that it aims, ultimately, to cover all of molecular biology with the same specificity. Certain parts of the ontology, worked over by domain experts at ontology workshops or by special groups tasked with expanding important sections of the vocabulary, are therefore much more detailed than other segments. Cardiovascular biology is a clear case in point, whereby grant funding for a cardiovascular-specific ontology team has up-scaled the detail and coverage of terms relating to the heart and vasculature. Terms relating to relatively more obscure areas of biology, such as the gastrovascular cavity of jellyfish, are entirely omitted from the Gene Ontology, and this is not because functions and processes do not exist in reality, but simply for pragmatic reasons. The sheer number of researchers interested in mammalian cardiovascular biology and its immediate relevance to human disease supersedes the interests of jellyfish researchers. Although it is always the intention of covering all molecular biology eventually, in practical terms certain sections of the Gene Ontology are highly specific, and the ontology risks becoming overwhelmingly complex if it were to expand proportionately. With over 30000 terms, the Consortium only ever plans for expansion.

Related to the choice of terms is the appropriate management of mapping files between the Gene Ontology and other controlled vocabularies. As the NISO Z39.19 standard suggests:

> "Once the various relationships among terms across multiple controlled vocabularies are identified, some provision *must* be made for retaining and maintaining these relationships for future use." (p.86, [224]).

Ontologies in the biosciences domain are organised centrally via the work of the OBO Foundry [85], and are designed to have no equivalence relationships between different ontologies. Each is entirely independent, and OBO Foundry principles advise ontology editors to avoid creating synonyms for terms in pre-existing vocabularies. Relationship and concept *types* are well controlled, grounded as they are in the tenets of Basic Formal Ontology, and thus the Gene Ontology shares a common structure with all the partner ontologies listed by the OBO Foundry.

With respect to mapping terms to other biological vocabularies, the Gene Ontology Consortium offers what it terms 'mapping files' which cross-reference GO terms to keywords in classifications such as Enzyme Commission enzyme numbers [4] and the UniProt Knowledgebase [230]. The Gene Ontology is the master vocabulary and is mapped to various subsidiaries. For example, as of November 2011, 699 mappings between UniProtKnowledgebase keywords and GO terms were listed on the GO website [231]. The Gene Ontology has also been part of the Unified Medical Language System since 2004, and GO concept types such as 'biological function' map to specific nodes, forming a part of the wider UMLS semantic map.

#### *3.3.3.5 Scope notes*

"The scope of terms is restricted to selected meanings within the domain of the controlled vocabulary. Each term *should* be formulated in such a way that it conveys the intended scope to any user of the controlled vocabulary." (p.20, [224])

The NISO Z39.19 standard advises against the use of parenthetical qualifiers or gloss to disambiguate homographs. Current versions of the Gene Ontology largely follow this guidance, although this was not always the case. Early ontology versions relied on the use of qualifiers to explicate the application of terms. Examples include:

1. GO:0015416, 'phosphonate transmembrane-transporting ATPase activity' which in 2001 was worded as 'phosphonate/organo-phosphate ester porter (broad specificity)' in order to clarify the enzyme specificity
2. GO:0015411, 'taurine-transporting ATPase activity' which in 2001 read as 'taurine (2-aminoethane sulfonate) porter' to clarify the taurine chemical species
3. GO:0016810, 'hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds' which is still present in GO versions as of November 2011, and qualifies the scope of the term in terms of which reactions *are not* catalysed by gene products annotated to this term

In addition, the 'sensu' qualifier is another example of GO application of parentheses to disambiguate homographs, although this was later dropped from the ontology.

As mentioned before, the current versions of the Gene Ontology do follow the standard, and so compound terms are preferred to qualified single words, qualifiers are omitted where alternate meanings for homographs fall beyond the remit of the molecular biology domain and qualifiers are standardised throughout the vocabulary. Large-scale efforts have also been made to disambiguate terms throughout the Function Ontology, as indicated by the decision made at the St.Croix GO meeting in 2003 to add the word 'activity' to all function terms where "...it reduces the ambiguity of the term name, therefore helping when GO is included in other systems (specifically UMLS), and [...] reduce user confusion".

The March 2003 Function Ontology edit file included 10 instances of the word 'activity' in term strings.

This next month's release of the ontology expanded instances of terms with 'activity' in the name to a total of 12,921 terms.

Whilst the Gene Ontology does not use scope notes in the traditional sense familiar to controlled vocabularies, 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 global policy embedded in the OBO format requiring the addition of comments.

One major omission in the Gene Ontology, given that the project is updated on an almost daily basis and terms are added or removed as required by ontology editors, is the absence of a comprehensive system for tracking changes and deletions from the ontology files. Terms which are removed are classed as 'obsolete' in GO parlance, and are moved to an 'obsolete' node where they remain for future reference. However 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 will be discussed in the chapter on term obsoletion.

#### *3.3.3.6 Compound terms*

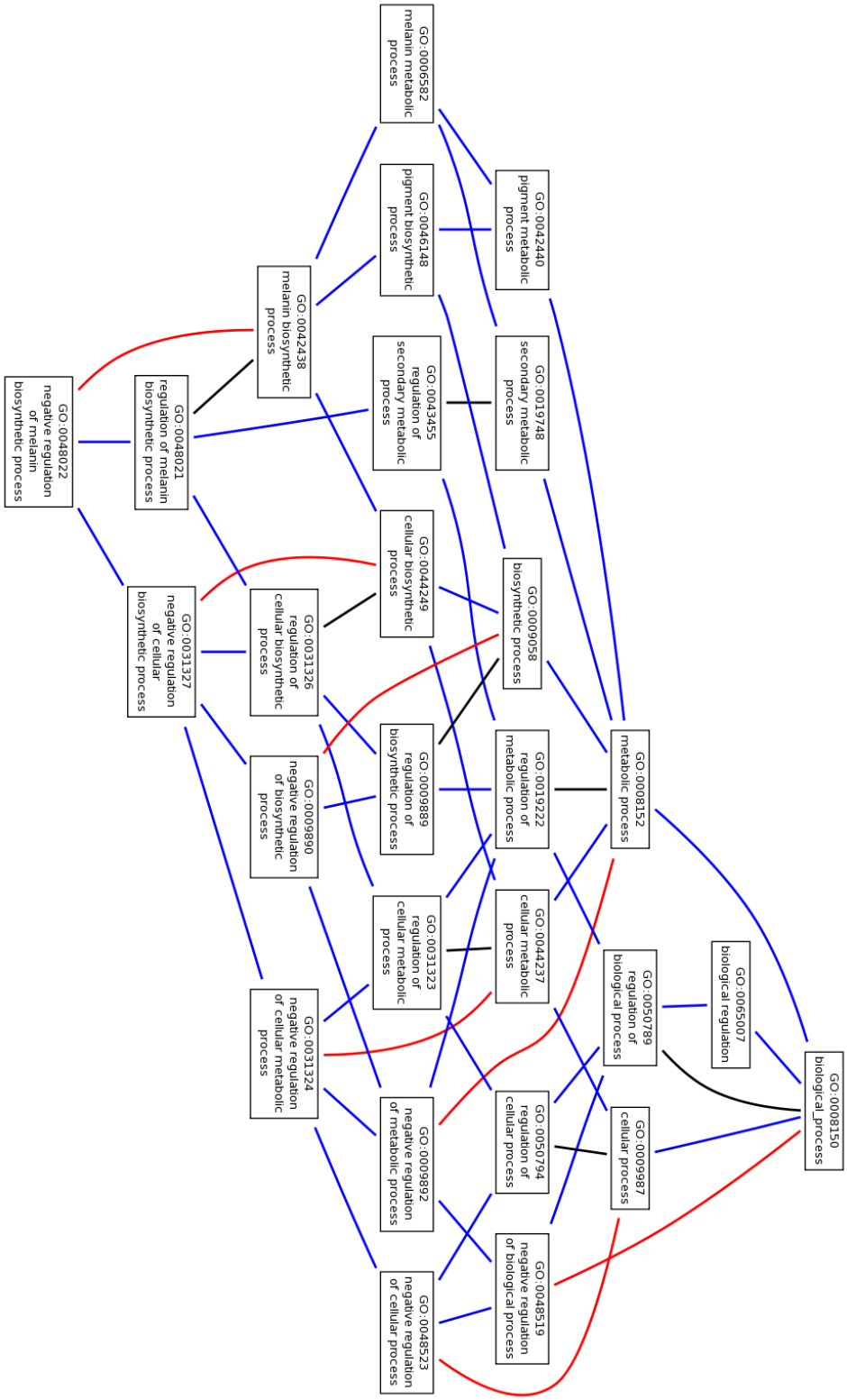
> "Because of the difficulty in defining "single concept," objective criteria for dealing with compound terms are provided..." (p.36, [224])

The NISO Z39.19 standard offers detailed recommendations on handling compound terms in controlled vocabularies, and advises on different conditions under which it may be advisable to split compound terms. The Gene Ontology, like all ontologies under the auspices of the OBO Foundry, do not ascribe to the concept view mentioned in the quote above, since ontological entities refer to classes in reality. However, the vast majority of terms in the Gene Ontology are structured like compound terms in a traditional controlled vocabulary, and some explanation for GO treatment of these kinds of terms is necessary.

The Gene Ontology is notable for its extensive usage of 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 that naturally fall into common types of classes. All molecular functions for gene products for example are deemed by the Gene Ontology to be kinds of activities which may occur in time and space in the cellular environment, and therefore all molecular functions fall into the class of a molecular 'activity'. Similarly, all cellular components intimately associated with the nucleus is some way are members of the class of cellular location identified as 'nuclear', and so throughout the Cellular Component Ontology one finds compound terms with the modifier 'nuclear'.

An example may better illustrate the Gene Ontology approach to compound terms. GO:0008152, 'metabolic process' is 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. The graph in Figure 8 below depicts parent terms for GO:0048022, 'negative regulation of melanin biosynthetic process', and the repetition in compound term wording is quite clear. As of November 2011 for example, there are over 1600 terms and term synonyms in the ontology which contain the wording 'biosynthetic process' and over 1400 terms with the wording 'metabolic process'.

**Figure 8: Graph structure for 'negative regulation of melanin biosynthetic process'**



An ongoing consideration in the design of a controlled vocabulary is when to split compound terms. NISO Z39.19 suggests literary warrant may support keeping compound terms, yet in the 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 arbitrary decision, based on what is comprehensible to the user, and what distinguishes particular classes of terms within the Gene Ontology itself (for example, appending the word 'activity' to most Molecular Function Ontology terms to distinguish them from similar Biological Process Ontology terms).

Compound terms should not be split if it leads to ambiguity. In the case of terms in the Molecular Function Ontology, the Gene Ontology is essentially forced to keep all compound terms with the a focus of 'activity' and such modifiers as 'regulator', 'transporter', or 'binding' because without these parts of the term, most function terms would simply be gene product names. For example:

- GO:0035033, 'histone deacetylase regulator activity' would split to 'histone deacetylase'
- GO:0005319, 'lipid transporter activity' would split to 'lipid'
- GO:0016597, 'amino acid binding' would split to 'amino acid'

Compounds terms should be retained if one part of the compound term is not relevant to the domain described by the controlled vocabulary. This is not normally a problem in the Gene Ontology because most terms are technical and have only one meaning within the biosciences.

If the compound term as a whole is more than the sum of its parts – that is the compound means something different to the parts of which it is composed – then the compound ought to be retained. The large number of chemical species in the Gene Ontology such as 'nitric oxide' means that splitting these compounds would indeed be nonsensical.

Proper names with two parts should be preserved as compound terms; this is not an issue in the Gene Ontology, which has few, if any, proper names in the vocabulary.

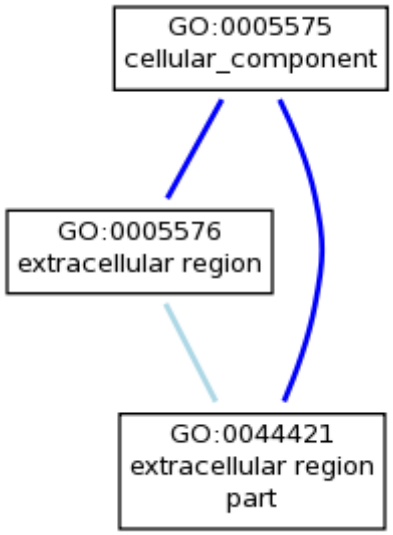
Where compound terms are used, the Gene Ontology often demonstrates poor compliance with the principle that modifiers should not be parts of the focus. The example given in the NISO Z39.19 standard is that of the term 'aircraft engines', whereby the engine (focus) is a part of the aircraft (modifier). The standard recommends that compound terms like these be split into two terms, to preserve logical consistency and to aid in creating a hierarchical structure of terms. In the Gene Ontology there are numerous instances in the Cell Component Ontology where the focus of compound terms are 'parts' such as 'cell pole' and 'cell envelope', 'endoplasmic reticulum lumen' or 'actin filament'. In each example, the focus of the compound term such as poles, envelopes, lumens and filaments are parts of a larger super-structure like the cell, endoplasmic reticulum or actin fibre.

The NISO Z39.19 standards also suggests splitting compound terms where it "...consists of a term representing a transitive action modified by a term for the object on which the action is performed" (p.40, [224]). This example given is of 'office management' where 'management' is a transitive action and is modified by the word 'office'. The Biological Process Ontology does contradict this suggestion, with several terms at detailed levels deliberately constructed as compounds of transitive verbs such as 'vesicle targeting to, from or within Golgi'. These examples are quite rare though; there are very few transitive uses of verbs in any parts of the Gene Ontology.

Again with respect to intransitive forms of verbs, the Gene Ontology generally avoids any verbal forms. Terms like 'germination', 'growth' or 'migration' have modifiers as the object on which this action is performed – 'pollen germination', 'lung growth' or 'leukocyte migration'. Yet as grammatical forms, these kinds of foci in compound terms are regular nouns or nouns derived from verbs by the addition of common suffixes (as in 'germination' is formed from the verb 'to germinate'). This approach circumvents any problems associated with compound intransitive verb forms.

The NISO Z39.19 mentions the following guidance on the purpose of node labels, which may be compound phrases: "A compound term *should not* be created solely for the reason that it forms a logical level in a hierarchy and would serve to group a set of narrower terms." The Gene Ontology strongly diverges from this recommendation, many levels in the hierarchy and their relationships being added specifically *because* they form what the developers consider to be a logical level in the hierarchy of entities. A simple example is given in Figure 9 below.

**Figure 9: GO:0044421, 'extracellular region part' and relations**




GO:0044421, 'extracellular region part' has two relationships, an 'is_a' relation to the root the Cellular Component Ontology and a 'part_of' relation to the 'extracellular region' node. Despite it having over thirty child terms, beneath the 'extracellular region' node it has only a single sibling, and is itself never used as an annotation term. The majority of the 8929 gene products annotated to this term inherit the relationship from annotations to the children, with only 18 annotations made directly to the 'extracellular region part' (data checked 01 December 2011), making the node a strong candidate as a superfluous logical division in the ontology. The NISO Z39.19 standard suggests it is preferable to replace these terms with node labels, of the form outlined below.

extracellular region
    *[extracellular region part]*
        activin complex
        angiogenin-PR1 complex
        [...]

#### *3.3.3.7 Relationships*

The NISO Z39.19 standard describes three different types of semantic linking to indicate relationships between terms in a controlled vocabulary:

1) Equivalency: exact, related and other types of synonyms
2) Hierarchy: terms related as instances or parts
3) Associative: such as cause and effect or process and agent

Relationships between terms must be reciprocal, whether they are equivalent, hierarchical or associative. Much importance has been placed by the Gene Ontology 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 Gene Ontology itself.

With respect to equivalency 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 an identifiable warrant, are subject to revision without history notes, and possess synonyms as added at the GO developers' discretion or user suggestion. Broader and narrower synonyms are added according to:

> "...an informal notion that encompasses both subsumption and possibly mereological and temporal containment" [218]

In practice, there is no rigorous or systematic attempt by the Gene Ontology to provide exhaustive equivalency relationships. The reference to an 'informal notion' above is a clue to the fact that there are opaque areas of ontology design. The hesitancy regarding the precise status of broader and narrower synonyms voices an acceptance that in biological language and the minds of users, the application of term names to concepts is an indeterminate area.

Lexical variants are handled as EXACT synonyms in the Gene Ontology, with American spellings used consistently across the ontologies and alternate, Anglicised spellings available ('signaling' rather than 'signalling', 'fiber' rather than 'fibre'). Definitions may use either US or British variations for words, and are usually corrected to US form for consistency.

Related synonyms are of particular interest in the Gene Ontology. They are described as synonyms which are not exact, narrow or broad and further extend the indeterminacy of synonym handling.

When GO terms are reported in published papers for example, an important convention is the citation of classes using the main term string and never a related synonym. In the context of ontological realism, the status of related synonyms in the Gene Ontology is not clear because there can be no conceptual 'fuzziness' relating one concept to another – terms either represent instances in reality or they do not. For example, GO:0070852, 'cell body fiber' is defined as a "...neuron projection that is found in unipolar neurons and corresponds to the region between the cell body and the point at which the single projection branches" and possesses the related synonym 'primary neurite'. The relatedness of 'cell body fiber' and 'primary neurite' is not clear, and for information retrieval purposes, related synonyms behave as exact synonyms with respect to ontology search

services. The question is, as the NISO Z39.19 standard describes, whether GO terms with related synonyms can "...be distinguished in the controlled vocabulary domain with sufficient precision to justify their representation as separate terms".

If this is the case, the standard suggests creating individual terms with different definitions. However, the Gene Ontology has no criteria on which to distinguish related synonyms as separate entities, other than an appeal to reality, in which case a 'cell body fiber' is the same as a 'primary neurite', or it is not. The Gene Ontology is not a probabilistic classification system. It cannot articulate degrees of relatedness, but it is significant that it is still necessary to include related synonyms in the ontology structure.

Hierarchical relationships are covered in detail in the Gene Ontology rules, and extend some of the suggestions made by the standard for controlled vocabularies, especially with respect to whole-part relationships. Since terms can take more than pathway back to the root node of the ontology, the Gene Ontology is a polyhierarchy, and this was perhaps the principal innovation of the ontology as compared to the simpler vocabularies which preceded it in the molecular biology domain. Associative relationships between terms in different hierarchies are not used in the Gene Ontology, and since the ontology consists of three distinct and, as yet, unconnected vocabularies (Cell Component, Biological Process, Molecular Function) it is perhaps in this area that GO might exploit the recommendations of the standard. Associative relationships can cover a range of kinds of links between different terms. Some examples of how the Gene Ontology might use associative relationships to relate terms from the different sub-ontologies are described in Table 11 below.

**Table 11: Suggestions for associative relationships in the Gene Ontology**

| Associative relationship type | NISO Z39.19 standard example | Example of potential relationship taken from the Gene Ontology |
|---|---|---|
| **Process/Agent** | hunting RT hunters | cell-matrix adhesion (BP) RT focal adhesion (CC) |
| **Process/Counteragent** | fire RT flame retardants | [Covered by 'negative regulation of' relationship type] |
| **Action/Property** | pollution RT environmental cleanup | photoreceptor activity (MF) RT detection of light stimulus (BP) |
| **Action/Product** | weaving RT cloth | cell division (BP) RT cell division site (CC) |
| **Action/Target** | binding RT books | cell surface binding (MF) RT cell surface (CC) |
| **Cause/Effect** | death RT bereavement | viral reproduction (BP) RT pathogenesis (BP) |
| **Concept or Object/Property** | poisons RT toxicity | axon (CC) RT transmission of nerve impulse (BP) |
| **Concept or Object/Origins** | Beluga caviar RT Caspian Sea | mitochondrion (CC) RT aerobic respiration (BP) |
| **Concept or Object/Units** | electric current RT amperes | No equivalents in the Gene Ontology |
| **Raw material/Product** | wheat RT flour | lipid binding (MF) RT (lipid transport (BP) |
| **Discipline or Field/Object or Practitioner** | neurology RT nervous system | No equivalents in the Gene Ontology |

### 3.3.4 Assessment of GO vocabulary standards

The NISO Z39.19 standard for controlled vocabulary construction was published back in 2005 and may be revised in the near future. It is unlikely to be changed drastically [169]. The standard was originally intended for thesauri, but was extended in the 2005 revision to encompass lists and taxonomies under the catch-all 'controlled vocabularies' [181]. Newer standards for thesauri construction like ISO 25964 owe much to NISO Z39.19 [96]. As the vocabulary landscape has changed with developments in electronic communication, so these newer standards do nod towards the perhaps unfamiliar domain of ontology. The rationale I maintain in this chapter is regardless of the name we give to something like the Gene Ontology – ontology, thesaurus, controlled vocabulary – the basic elements of a standard like NISO Z39.19 are still sound and relevant.

In short, when constructing a vocabulary of any kind for indexing information – be clear.

The Gene Ontology is therefore a sophisticated, pre-coordination system of terms for indexing gene products across the entire domain of molecular biology in a species-neutral manner. It encodes various types of relationships of assorted complexity for facilitating machine-driven reasoning across GO graphs, and this is important in biology for making predictions about new gene products, and about what kinds of functions and processes are occurring in experimental systems.

But is the Gene Ontology clear?

The fact that the GO Consortium has 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 the desired outcome, which is clarity. The GO project's internal rules, oftentimes labyrinthine in their own way, do achieve, in part, many of the goals the NISO Z39.19 standard aims 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 Gene Ontology covers a very specific domain. As the example of the word 'membrane' illustrates though, there are major risks presented by the synthesis of the Gene Ontology with other ontologies across the biosciences, where polysemous words can, when removed from vital contextual clues in text bodies, pose significant confusion.

Synonym control is very poor in the Gene Ontology, 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.

My argument here is 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. The GO Consortium sees this as a solution to an inherent flaw in the way biologists talk about functions and processes, such as using gene product class names as surrogates for what from the GO perspective are 'true' processes or molecular functions.

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 Gene Ontology 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 Gene Ontology, these kinds of notes are brief, omitted, or copied from existing term notes, and are rarely informative. The GO project relies on definitions to limit scope, coupled with the understanding that expert users can apply their existing knowledge to grasp how particular GO terms relate to the products they index. Persistent problems in GO annotation consistency, not explored in this thesis but alluded to in interviews with GO developers (see Appendix 6.1), suggest that expanded scope notes may be necessary.

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. As later chapters will explore, 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.

On the subject of compound terms, there is little more to add from the previous section other than to repeat that the Gene Ontology uses compound terms extensively. This is a design choice, and the NISO Z39.19 standard merely advises that a controlled vocabulary may consider splitting terms where it can reduce ambiguity. Of particular interest though 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 Gene Ontology has lent the impression that hierarchical levels do matter.

Much as semantic relationships between GO term names and their synonyms are dealt with in a casual way by the ontology, and arguably do not lend clarity to how the language of the ontology ought to be applied, so too has the proliferation of compound terms designed to cut the ontology

into what look like more logical slices, an approach more aesthetic than objective, only created mis-impressions and confusion in ontology applications.

My conclusion is that were the Gene Ontology to follow some of the more traditional standards for vocabulary construction, as suggested by the NISO Z39.19 standards, 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.

## 3.4 Discourse analysis

In the next results section, I apply Critical Discourse Analysis to a body of Gene Ontology mailing list texts, in order to understand the relative importance of social roles and the exercise of power by authorities in representing scientific truth within the ontology.

### 3.4.1 Introduction to discourse analysis in LIS

In the Chapter 3.3, I explored the vocabulary standards behind the development of the Gene Ontology. In a Chapter 3.5 on term obsoletion, I will be investigating the practical consequences of these standards, and the kinds of reasons the GO Consortium uses for excluding classes of terms from the ontology.

In this chapter though, I want to first establish a link between the Gene Ontology vocabulary standards and the processes, like term obsoletion, by which the standards are implemented. The Gene Ontology project markets itself as an open system, reflecting the needs of the molecular biology research community. It seeks contributions from users through various mechanisms such as ontology request forms, annotation events and promotional 'road-shows'.

At the same time, the Gene Ontology Consortium details "...[r]ules governing content and stylistic aspects of GO terms, standard definitions and term relationships" [232]. The guidelines are ordered, logical and scientific, and appear to offer objective criteria for the admission of terms to the Gene Ontology. Yet they were published long after the Gene Ontology had reached a mature form, and in the absence of compliance with any international standards of vocabulary construction, are liable to change according to the needs of the GO Consortium.

How does the GO Consortium resolve the needs of users in the molecular biology domain with the standards it has established for the ontology? This is a question about the social relations between the Gene Ontology organisation and the wider scientific community, about how the organisation listens to and acts upon proposals for additions and changes to the ontology structure.

Evidence from my previous concept analysis suggested that GO terms can be grounded socially and historically within scientific paradigms, and according to different epistemologies for concepts. This argument was developed in opposition to the overarching philosophical basis to ontologies in biology, which is ontological realism. If one accepts that concepts in a scientific classification are in part, or are entirely, constructed according to social and historical parameters, what evidence is there from the work of the Gene Ontology to support this contention?

One source of evidence for the social relationships between the Gene Ontology Consortium and the wider scientific community is the GO mailing lists.

The Gene Ontology 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 are not the whole design process and, since the ontology structure changes over time, they lie in the history of the Gene Ontology. Yet the within the mailing lists, we can witness how a scientific classification is socially constructed by studying the discussions between developers, curators and users as the ontology developed.

As documents of the design process, the mailing lists offer an opportunity to articulate how the GO Consortium has dealt with problems including:

- How to choose scientific concepts to represent within the ontology
- How to agree on term names for these concepts
- How to define terms
- How to decide when to change the ontology

How the Gene Ontology Consortium resolved these problems is of particular interest to my thesis because, as explained previously, no traditional standard for controlled vocabulary construction was used in the design of the ontology. Since standards like the ANSI/NISO Z39.19 [224] would normally have offered guidance on how to handle the particular problems listed above, the processes by which the Consortium has fashioned its own solutions will reveal the sociological aspect to the construction of this scientific classification.

This chapter contributes to answering the research question of how the Gene Ontology developers have tried to make the vocabulary objective, and aims to:

1. Describe the macro-level linguistic features of the GO mailing list
2. Articulate what the GO developers use the mailing list for
3. Describe the micro-level features of the language used by authors on the GO mailing list
4. Analyse how disagreements over the structure of the ontology are resolved by the mailing list participants
5. Further to '4', explore whether the mailing list discourse suggests that pre-existing ideology plays any part in resolving disputes

My investigations of the Gene Ontology are conducted according to the research principles of domain analysis [116, 127]. Hjørland advises that linguistics is an important tool to understand the meaning of texts in a domain, and in particular discourse studies offer ways to study the domain-specific meanings of special scientific languages [127]. I share Budd's broad distinction between two different types of discourse analysis [233]. The former is the analysis of 'transactional language' or, more simply, conversations and the transfer of information between speakers. The latter sense is discourse which "...embraces the social, cultural, political" [233], where texts and speech are a social acts, consequential to the social milieu in which speech occurs, and subject to the powers, ideologies and beliefs of speakers.

Fairclough defines this second sense of 'discourse' to mean "...any spoken or written language use conceived as social practice" [234]. In this chapter, I approach the Gene Ontology developer mailing lists as examples of the language used in the social practice of building a special scientific classification. My aim is to discover whether these mailing list exchanges embody "...epistemological, rhetorical, communicative, obfuscatory, political, cultural and other intentions" [233] unique to the Gene Ontology, and suggest what these intentions may mean for the construction of a classification.

As Wetherell writes, discourse is social action [235]; the GO mailing lists represent more than a record of discussions between individuals. The mailing list texts are fragments of a wider scientific discourse in which science as a social institution situates itself in society and exercises power over

the ‘scientific’. Electronic communication is a small part of this discourse (see Table 12 below). Discourse analysis is a tool to explore social actions in a small part of the scientific discourse represented by the GO developers as they built the Gene Ontology.

**Table 12: Typology of scientific discourse, by style and mode, and including CMC [236]**

| | Style | | | | |
|---|---|---|---|---|---|
| **Mode** | Frozen | Formal | Consultative | Casual | Intimate |
| **Written** | Theory<br>Laboratory notes | Scientific article<br>Conference proceedings<br>Website | Email<br>Mailing list | Internet chat<br>Twitter<br>Blog | Text message |
| **Spoken** | Witness statements | Lecture<br>Presentation<br>Podcast<br>Video | Teaching<br>Workshops<br>Meetings<br>Conference call | Laboratory chat<br>Telephone | Friendship |

Discourse analysis is a way to test my thesis that ontological realism fails to account for the social processes guiding how biologists have constructed the Gene Ontology, and for how these social relations and the exercise of power have determined ontology content.

The context of the words used is important: the people speaking, their position within the GO Consortium, when the messages were sent, the subjects discussed. What do the words used, the responses, the arc of the debate, say about the social relations, if any, at play in the decision-making process behind the ontology? Is deciding the contents of the ontology a trivial task, guided by clear logic and methods as the GO Consortium ontology guidelines now seem to promote? Or is the business of deciding terms, relations and ontology contents a messier business?

I will apply Critical Discourse Analysis (CDA) to a selected GO mailing list corpus. As van Dijk describes, CDA aims to “...describe and explain how power abuse is enacted, reproduced and legitimised by the text and talk of dominant groups or organisations” [237]. In the context of the Gene Ontology, I will use CDA to explore whether the Gene Ontology Consortium, as a developing organisation created to manage the GO project, applies objective and value-free scientific principles in deciding problems associated with ontology’s design.

This approach is consistent with several other discourse analysis studies in the LIS domain. For example, Haider outlines the development discourse which has served to construct the institutional notion of ‘information poverty’ in LIS [238].Developed from a Foucauldian perspective, Haider defines discourse as “...as a socially constructed regime of knowledge and truth that forms the social reality about which it speaks” and it is this idea of a ‘regime’ which is an important element in Foucault’s take on understanding language.

Haider elegantly states the aim of her approach to information poverty in LIS being “...to assess critically the strategies leading to the construction of the concept and thus to challenge some of the underlying assumptions.” Foucault’s work therefore provides a theoretical framework for understanding the construction of concepts, and the role played by institutions, authorities and power in manufacturing these concepts [239]. A tension exists, between power and knowledge.

Foucault resisted the idea that power could be attributed to a definable agency or *politique*. Rather, power was distributed and net-like and realised by the interplay between the opposing forces wrestling within institutions. From power is derived types of knowledge, and Foucault wrote in some detail on the idea of expert knowledge: knowledge created according to a regime of truth recognised by a specific group of experts. From the expert, truth is produced which, in Haider's analysis, is the professional librarian producing the truth of what 'information poverty' means [238, 239].

Talja outlines discourse analysis methodology for approaching qualitative research data like interview responses [240]. Of special note is what Foucault refers to as 'statements' which, in Talja's paper, are "...implicit starting points behind a particular way of speaking about a topic". When interpreting an interviewee's responses, one must account for these statements which may underlie that which the speaker utters; these statements are assumptions, tacit and unscrutinized.

#### *3.4.1.1 Methodology details*

The Gene Ontology project was conceived in the summer 1998, and by January 1999 a mailing list was established for communications between the partners and users of the nascent project. This mailing list was eventually to become part of a number of different mailing lists dealing with different aspects of the project, such as end-user issues, annotations, or special interests focused on specific domains of biology. Other mailing lists on the Stanford servers[2] include GO Discuss (for general discussions about the ontology) and GO Friends mailing lists (for users interested in the GO project).

Since December 2000 the GO Mailing list has been moderated by key Consortium members for spam and inappropriate messages. As far as possible though, the list has been kept open, and has invited discussion both between Consortium members, and between the Consortium and the GO users. Archived GO Mailing List posts are freely accessible to the public for download from a server in Stanford[3].

GO developers started using the GO mailing list in January 1999. The Gene Ontology was conceived by three originator species databases, Flybase (for fruit fly researchers), SGD (for yeast researchers) and JAX (for mice researchers), and it is principally researchers and computer specialists at these three institutions who began to use the mailing list as a means for debating how the ontology should be constructed and change. The mailing list is still active today, and messages are archived as of May 2012.

Mailing list posts were downloaded in their entirety from the Stanford servers and saved as plain text files. These text files are the raw data sources used for a discourse analysis of communications between Gene Ontology developers.

These discussions are a small sample of many documents created in the creation of what is the Gene Ontology today. There exist other mailing lists, the minutes of meetings, private emails, published papers and the like which document how the ontology was created. Some of these other documents will be considered elsewhere in this thesis, although many sources are and will remain unavailable, such as private correspondences between GO developers.

[2] http://fafner.stanford.edu/mailman/listinfo
[3] http://fafner.stanford.edu/pipermail/go/

The main challenge of applying CDA to these mailing list texts is the sheer volume of material that can potentially be analysed. With many hundreds of pages of conversations, broad themes and power relations in the discourse can be identified. However, this kind of reading may ignore the interesting linguistic details that are present at the level of syntax or grammar. My approach is therefore to conduct a broad reading of the entire mailing lists for the period studied, and then to choose several, individual threads to analyse the language used, and its social implications, more closely. The weakness to this research approach though is the potential bias inherent in selecting texts for detailed discourse analysis.

Data from my discourse analysis readings of the mailing list are presented in two forms for this thesis. In Appendix 6.2 are notes of the macro-level reading of all mailing list posts for the period 1999 – 2002. This period was chosen because it was a time when all shared, electronic discussions about the contents of the Gene Ontology were conducted using the GO mailing list. After February 2002, many development communications were migrated to Sourceforge threads, which are not included in this discourse analysis.

Further to this macro-level reading of the main features of electronic discussions on the GO mailing list, I have also included several detailed analyses of discrete message threads. These threads were selected as they deal with contentious issues in the GO development process, and contain evidence social actions and interplay of powers between the GO Consortium and GO users.

### 3.4.2 Results

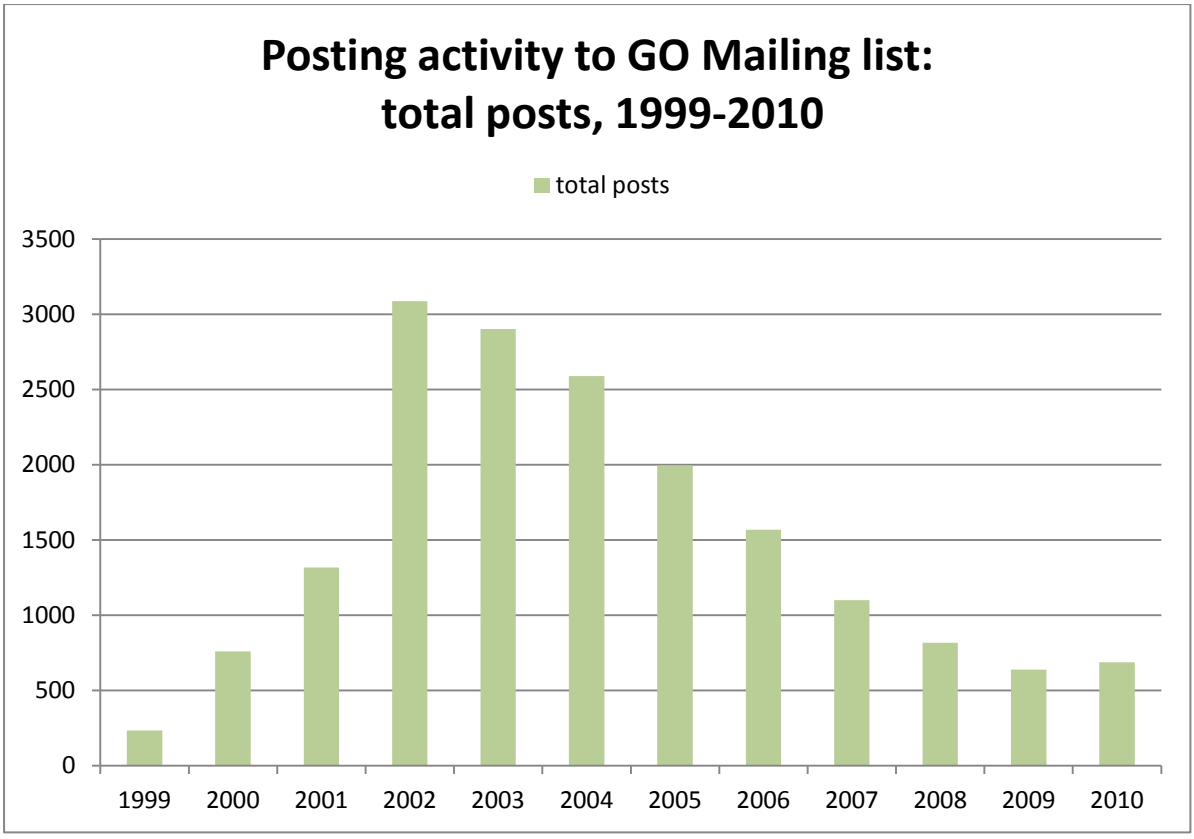
#### *3.4.2.1 Overview of posting activity on the GO mailing list*

As a broad introduction to just how active the GO mailing list has been, Table 13 below shows data and post counts for the list from 1999-2000. As can be seen from Figure 10 charting annual data and post counts to the list, posting to the list peaked in 2002 and has slowly declined in the intervening years. This is principally explained by the introduction of alternative communication methods for discussing ontology issues, such as Sourceforge trackers for coordinating additions and deletions from the ontology files.

**Table 13: Posting activity to the GO mailing list**

| Year | Total data posted | Total number of posts |
|---|---|---|
| **1999** | 175 | 233 |
| **2000** | 550 | 760 |
| **2001** | 654 | 1316 |
| **2002** | 2350 | 3087 |
| **2003** | 1472 | 2903 |
| **2004** | 1293 | 2590 |
| **2005** | 1024 | 1996 |
| **2006** | 810 | 1569 |
| **2007** | 612 | 1100 |
| **2008** | 387 | 816 |
| **2009** | 159 | 637 |
| **2010** | 380 | 688 |

Figure 10: Posting activity to GO Mailing list: total posts, 1999-2010




#### *3.4.2.2 Macro-level reading of the mailing lists*

Early discussions on the mailing list are reserved for a small number of the first Gene Ontology 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.

The principal attributes of sources featured in the macro-level reading of mailing list posts are shown in Table 14.

Table 14: Attributes considered in the macro-level reading of the GO mailing list discourse

| Attribute | Explanation |
|---|---|
| **Author** | Who is speaking? Posts are sent from identifiable email addresses and are usually signed by the author |
| **Theme** | What are they speaking about? Posts cover ontology topics, problems and design questions |
| **Sequence** | What is the order of speech? Message threads are sequential discussions, with posts and responses delayed in time |
| **Quotes** | What was said? Pertinent quotes are recorded verbatim and are included without editing or typographical corrections |
| **Observations** | What did the authors mean? My interpretation of discussion content based on the CDA approach: how is the power and authority of senior GO Consortium members used to legitimise certain truths in the text of the GO mailing list |

In analysing the arc of discussions on the mailing list, for the sake of brevity and sense I have not recorded every single exchange, although I have as far as possible attempted to note every single major announcement, comment or shift in topic.

The analysis, my comments and my decisions regarding what to include in my notes is therefore subject to my own bias and interpretations. However in being explicit as to what I have read and why I have interpreted it in such a way, I hope to develop an argument which is both open to criticism and defensible. This is in keeping with the CDA approach.

##### 3.4.2.2.1 Attribute: Author

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.

Of the most important, the people I will refer to as the *Senior Developers*. These Senior Developers retain rights as editors over the ontology, and maintain top-level control over how the ontology changes, although other users can be privileged with ontology-editing rights. The Senior Developers initiated the project, determined the overall direction of the project, resolved disagreements between other editors and users, and took responsibility for securing and distributing financial backing for the project.

Related to the Senior Developers are the *Curators*. Curators are partners from other biological databases who may or may not be given ontology-editing rights, but are responsible for implementing GO annotations within the context of their own system environment. Editing privileges are a measure of an individual's seniority and trust within the GO project, and they have represented an important contributory voice in deciding the look and feel of both the ontology and many of the software applications developed specifically for the purposes of the project, such as the AmiGO, a web-based GO browser or OBO-Edit, an ontology browsing platform. Curators are usually biological experts, and often will be manually creating links between GO terms and gene product entities in these databases. They may alternatively be bioinformatic specialists who curate gene products automatically using computer algorithms.

From the very start of the Gene Ontology project, a number of *Commercial Partners* have been allied with the GO Consortium. The most important of these is the pharmaceutical company AstraZeneca, which provided early financial backing to get the project off the ground, and offered direct technical assistance to the developers as they were beginning to construct the ontology structure and standards. As the project developed, other commercial biotech companies reached out to the GO project offering expertise or asking for technical assistance in applying what was a new technology for analysing high-throughput biological datasets. As such, these Commercial Partners form a distinct GO user group communicating on the GO mailing list.

The Gene Ontology was created to facilitate cross-database searches for gene product functions, and representatives of these partner species databases have always been invited to GO Meetings and offered their opinions of the direction the project should take, often based on the specific requirements of their database environments and user base. As the Gene Ontology project expanded, other partner databases were invited to join and aid in both the manual annotation effort

and in expanding the ontology structure to include terminology from species-specific domains, such as plant science or virology. *Partner Database* speakers on the Gene Ontology speak for the database they represent.

Beyond the Gene Ontology project itself, the GO mailing list has always been open to what I call *Expert Users*. Expert Users are proficient in applying the ontology and annotations to empirical problems, and used the GO mailing list to establish direct contact with the GO developers, reporting errors in the ontology files or requesting support in applying the Gene Ontology and its annotation files to specific biological problems.

Finally, the GO mailing list is open to *Naive Users*. These users, even though they may be biological experts, may know very little about the ontologies and technical aspects of the Gene ontology project, yet use the GO mailing list to communicate their ideas on the project, and to get support from GO experts with their work.

#### 3.4.2.2.2 Attribute: Theme

Several broad categories of discussion are identifiable through reading the GO mailing lists between 1999 and 2002. The GO Consortium was later to develop specific electronic communication channels to handle these themes, and each type of discussion represents an element in the routine work of building and maintaining the ontology.

*Announcements* form part of the house-keeping activity associated with any IT project. GO Mailing List announcements include changes to the ontology, software releases, publication of minutes from GO meetings, reports of errors in the ontology files, calls for papers, or progress on annotation. Announcements on the GO mailing list are rarely discussed in any great detail. Rather, they form a series of statements in the lifecycle of the Gene Ontology, the discussions behind the announcements being largely concluded. Examples of announcement threads include:

- 'New version of GO.' (April 1999): a Senior Developer announces the release of GO file version 0.2a
- 'Notes from Feb 2000 GO meeting' (March 2000): a Curator at the Saccharomyces Genome Database shares the minutes of the last GO meeting held at the AstraZeneca offices in Massachusetts
- 'New GO-EDIT version available' (March 2001): a software specialist Curator at the Berkeley Drosophila Genome Project announces a new version of an ontology editing package

An additional house-keeping activity conducted through the medium of electronic communication on the GO mailing list is what I term *Administrative threads*. Much like Announcements, Administrative threads are not subject to extensive discussion, although they may extend to several additional thread posts in which the speakers will negotiate the details of organizational work. Conversations include the details of meetings such as locations and travel arrangements, agendas for meetings, financial questions, together with threads about organizational and personnel changes within the GO Consortium.

As the Gene Ontology project developed, the GO Consortium became increasingly aware that it had a role to play in liaising with users. The GO mailing list is therefore used as a place to communicate

problems. In *Problem Communications*, errors are flagged for discussion, requests for assistance made, and solutions offered by GO users with the necessary expertise.

Problem communications are now dealt with via Sourceforge trackers, which replicate the mailing list and offer additional features for handling queries, or via a dedicated 'GO Helpdesk', where users can directly message a technical support team. Sourceforge trackers cover many of the original problem communications I identified in reading through the GO mailing lists, with topics covering GO annotation, software development and ontology file format issues. These trackers indicate how the GO Consortium views and prioritises its communication role with the molecular biology community, and one of the most heavily used tracker handles what is a recurrent thread topic on the GO mailing lists, that of *Ontology Requests*.

Ontology requests are suggestions for new terms, changes to ontology relationships, additions to synonyms and definitions for existing terms, and the obsoletion of redundant. The Gene Ontology is continuously changing, and has long been viewed by its developers as a constant 'work-in-progress'. An important work process for the GO Consortium is therefore the management of ontology requests, to ensure that best, current biological knowledge is included in the ontology, and that changes are made in accordance with commonly agreed GO rules and the broader commitment to ontological realism.

#### 3.4.2.2.3 Attribute: Sequence

A feature of electronic communication like a mailing list is the 'one-to-many' nature of each message – one message will be read by many people on the list. The other feature which is unusual in electronic communication is the time-delay associated with a conversation, whereby speakers may not respond immediately, and have time to consider and edit responses before posting a reply.

The macro-level reading of the mailing list discourse 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.

#### 3.4.2.2.4 Attribute: Quotes

Messages are quoted in full, and are partially anonymized in these results. Message headers, which include email addresses and data on posting times, have been removed.

Where authors have replied and included large sections of previous messages in that reply, that quoted section has been removed to aid readability. Spelling errors and typographical errors have been preserved as they were in the original message.

It is worth considering that these messages and the replies speakers offered would normally have been read and composed in some form of email client. It is assumed that the kind of client used, and the way the client would have presented messages and provided reply options, had no bearing on the way speakers addressed the GO mailing list.

#### 3.4.2.2.5 Attribute: Observations

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 in the resolution of disputes between speakers.

Discussed next is a short example of the type of observations made in an Administrative thread about a statement:

> "Please send any additions, suggestions, etc. In particular, curators please send ideas for things to cover in the Saturday session."

Taken from a thread titled, 'March 2001 meeting: call for agenda items' (February 2001) here a Senior Developer makes the speech act of a request, with the intended illocutionary force of getting other GO mailing list participants to send in agenda items. Here we see several common features of mailing list discussions such as deletion, whereby the object of a sentence is omitted, requiring the reader to infer the meaning of statements from the context of the message, and a clear relational value whereby the author has selected a polite phrasing with the use of 'please' to stimulate response from other Curators on the list.

In this example, a Senior Developer responds shortly afterwards, with information that a Curator will need time at the meeting to report on a proposed change to the ontology.

An example of observations made on a Problem Communication involving a user from the Jackson Laboratory (the main GO partner curating mouse gene products) follows:

> "Your home page has a typo."

In response to this message thread titled 'your home page' from September 2010 a Senior Developer at Stanford edits the GO home page, and reports the changes to the GO mailing list. This is a simple exchange, and makes no reference to the actual contents of the ontology. The choice of the slang word 'typo' implies a certain level of informality and as an expressive value captures the sense that the error is really quite minor.

A significant number of Problem Communications relate to ontology file formats and typographical errors in ontology files. As ontology file formats in the sciences have developed and common standards agreed, these problems have slowly disappeared from GO discussion lists, although it is worth noting that for the period covered by these texts, they did make up a significant number of Problem Communications.

Another Naive User mails the list with a more complex Problem Communication:

> "Please let me know if there is any computer program that can classify genes according to classification system propounded by Gene ontology."

This thread titled 'Software for classifying genes according to Gene Ontology system' and created in December 2001 stimulates a single response, which both offers a solution, and indicates to the user that this is an issue the GO Consortium intends to address for all users in the future. The experiential value of the request above is such that the user holds a belief that there should be a way to classify genes automatically using the Gene Ontology.

> "InterPro - you can use their web interface or download interproscan to do it in batch.
>
> There are other ways of doing this - I should have a page up about this next week"

The response is written by a Senior Developer at the *Drosophila* laboratory in Berkeley, and illustrates several important features of the mailing list discourse. 'InterPro' is a proper name and refers to a protein information database. It is one of many names for objects or services specific to the molecular biology domain, and is an example of the kind of domain-specific terminology common to the GO mailing list. Of interest here is the brevity of the response. The teaching of methodologies for classifying genes can occupy weeks of teaching on bioinformatic courses, yet the response to this user is a simple re-direct to the appropriate tool. The response shows a common feature of the GO mailing list discourse which is deletion – the sentence "...you can use their web interface..." only makes sense by inferring what is meant in the context of the previous message. In addition, alternative methodologies for achieving the user's request are alluded to.

Yet it is the choice of modality in the last sentence – "I <u>should</u> have a page up..." – which communicates both an uncertainty about whether the requested information will be made available in the near future, and deference on the part of the Senior Developer. Yet this deference does not quite match the politeness of the original user request, where the request prefixed by 'please' suggests a formality in addressing common to a speaker conversing with a more senior authority figure. This exchange illustrates the power differential between Senior Developers and Naive Users.

An FAQ for how to use Interproscan was added to the EBI web pages in April 2002, several months after this exchange on the GO mailing list. The first record I can find of the tool being described on the GO Consortium web pages is February 2007.

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.

Of the kinds of discussion found on the GO mailing lists, Ontology Requests represent a major source of discourse data for understanding both the social relations at play in the construction of the ontology, and the role played by scientific logic and objectivity in deciding ontology contents.

A simple example of an ontology request by a Naive User is illustrated in a thread from March 2001, titled 'GO Compartments and nuclear bodies'. After explaining the context to a suggestion, and supporting the argument with references to peer-reviewed articles, the Naive User requests the addition of 'nuclear body' as a child of the parent term 'nucleoplasm'. The addition is duly implemented and the user comments:

> "This is fun!"

Here we have a simple example of the informality common to the GO mailing list. The use of slang, emoticons and personal expressions of feelings on a topic indicate that at times, speakers are relaxed about ontology discussions. What is significant in this context is *who* is speaking in this discourse, and in the example above, it is a Naive User seeking acceptance from the group, the larger group of more senior ontology developers and expert curators. When a Naive User declares 'This is fun!', this statement is a speech act with an intended effect and the effect in this context is to demonstrate to more senior persons on the GO mailing list that this new user is happy to comply with the rules, procedures or standards required for adding new terms to the ontology – this user is seeking social acceptance.

A Senior Editor contacts the GO mailing list in January 2000 with a message titled 'a question about spindles':

> "I'm not sure that %establishment of cell polarity ; GO:0007163 should be a child of %cell adhesion ; GO:0007155 -- certainly it wouldn't work for yeast (though yeast cell polarity has its own node)."

Here is an ontology request to change a relationship – the position of a particular term and its children in the hierarchy. The Senior Editor is querying the relationship of 'establishment of cell polarity' to the parent 'cell adhesion' in the Biological Process Ontology and the tone is quite different to the previous example.

In multicellular organisms, the binding of cells to one another is conceptually related to the establishment of polarity in cells, or the sense that cells have an axis. An example would be epithelial cells lining the gut: by virtue of their adhesion to adjacent and underlying cells, these epithelial cells have polarity, with one side of the cell facing into the gut lumen absorbing nutrients from food, and the other side of the cell passing these nutrients through into the bloodstream.

The same requirement for cell adhesion is not true of unicellular organisms like yeast, a species in which this Senior Editor has expertise. In yeast, the establishment of cell polarity is more commonly associated with cell division, and changes to the yeast cell machinery. This Senior Editor is therefore trying to convince the rest of the GO mailing list of the validity of the yeast perspective.

The syntax of the ontology request is interesting. The Senior Editor starts with a modal adjustment "I'm not sure..." suggesting hesitancy and deference to the other readers on the mailing list. There is then a switch to the passive voice with "...it wouldn't work for yeast". This function de-emphasises the speaker, and instead refers to the authority of 'yeast', presumably meaning the collective expert knowledge of yeast biologists. This technique is certainly an attempt to exercise power to marginalise one concept for 'establishment of cell polarity' in favour of a more generalised concept. Technical language common to the Gene Ontology project is used in the request, such as GO term identification numbers, references to 'nodes' and 'children', and the percentage symbol which in a GO file format means 'this is a GO term'. The request is formed in appropriate, scientific language, and communicates the exercise of a certain authority in the social context of an ontology request.

The request is successful, with GO:0007163 'establishment of cell polarity' moved to be a child of 'cytoskeleton organization and biogenesis'. The move is not entirely uncontroversial though - the Senior Editor in charge of the term re-location writes in response:

"I am not convinced this is the best place, but [MH] is quite right cell adhesion was wrong !"

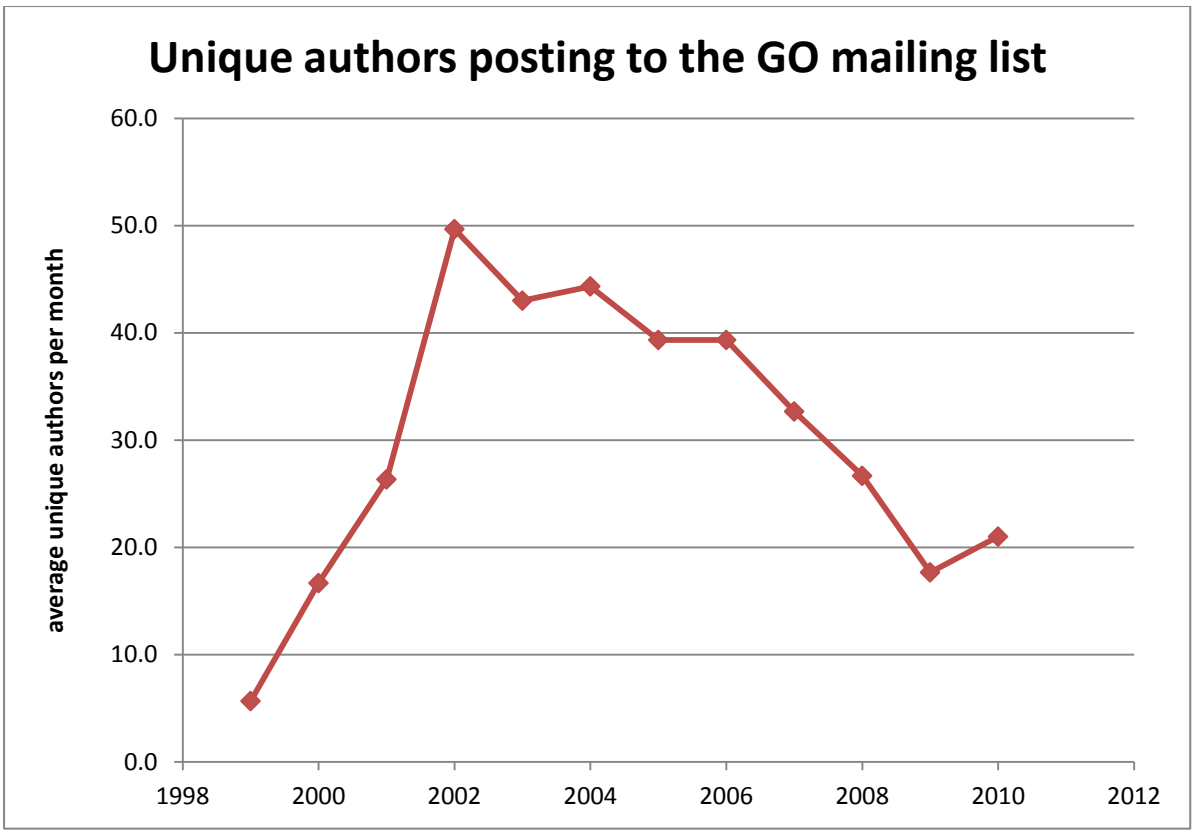
There is no such deference or hesitancy with regard to removing the relationship to 'cell adhesion' in this message. The speaker is one of the key instigators of the Gene Ontology project, and the statement that "...cell adhesion was wrong" is a good illustration of ontological realism at work. Knowledge represented by the Gene Ontology is either true or false. The relocation of this term is the correction of an error. Transitivity runs through this short piece of text. The thought processes of the individual that chose this location for 'establishment of cell polarity', or the potential conceptualisations of individual users for whom the previous ontology relationship may well have made sense, are marginalised. The ontology is a set of locations and ontology terms are rendered as abstract entities to be moved from 'wrong' to 'right' places.

In fact, the term in question has had its location and ontology relationships changed several times over the course of the Gene Ontology history. It is now a high-level term in the Biological Process Ontology, as a child of 'cellular process', with references to cell adhesion and yeast division entirely erased from the vocabulary. Its name was modified to 'establishment and maintenance of cell polarity', and these repeated edits do support my thesis that the ontology developers can struggle to agree on a shared meanings for terms.

Ontology request discussions are typified on the GO mailing list by several responses to the original posting, and often with several speakers contributing to the discussion. Decisions on ontology requests were either made quickly (within weeks) or deferred to later date, often when Consortium meetings or expert panels were due to convene.

What is perhaps surprising is the small number of speakers involved in ontology request discourses (see Figure 11 below). 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 Gene Ontology 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 Gene Ontology covers a huge, complex domain, GO terminology has not always been authored by those with expertise in the relevant area of molecular biology.

Figure 11: Unique authors involved in GO mailing list discussions




Individuals with proposals for changes frequently cite their inexperience on a topic, or confess to the dissatisfactory nature of compromises much like the example above ("I am not convinced this is the best place..."). There is common agreement that it is better to try and improve the ontology and make changes, even if in the short term a solution is not ideal, so that the ontology may become 'more true'.

Ontology requests can fail, and in some cases it can be the failure of a subordinate GO project member to convince a more authoritative Consortium partner that a change is necessary. Consider the message posted in January 2001 by a Curator under the title 'Caspase':

> "Is it proper to list the caspases as separate functions or are we falling into the trap of gene-level terms?"

Historical versions of the ontology files indicate numerous terms in the Molecular Functions Ontology of the form Caspase-1, Caspase-2 and so forth. These were to be obsoleted from the ontology in May 2003, but the original response from a Senior Editor denied any need to address the issue:

> "I do not think these are too gene levelish."

Gene products, explored previously as a problem for the Gene Ontology, take on several different names with speakers on the mailing list. In this simple exchange 'the trap of gene-level terms' and 'too gene levelish' are examples of overlexicalization in relation to the one concept. At this early stage of the project, only two years after its inception, the Gene Ontology is erasing the means of

describing genes beyond a certain level. The motivation is predicated on a reductionist assumption, that there must be molecular functions common to gene products across different species. Beyond a certain level, these generalisations break down and so caspases must be grouped under single term 'caspase activity' and the detail of these capases' potentially varied forms and functions is subsumed by a broader category.

#### *3.4.2.3 Detailed discourse analysis of selected texts*

I will now present four detailed discourse analyses of exchanges between different types of speakers on the Gene Ontology mailing list. The messages are problem communications and ontology requests; detailed comments on each turn and additional notes are included in Appendices 6.3-6.6.

In each sample, I am looking from a critical perspective for evidence in language that speakers are exercising, or submitting, to power in the process of resolving problems and issues with the Gene Ontology Consortium. Passages are included in the results below to aid in understanding my interpretations of the texts, but full text for each message thread is presented in the appendices as speakers, turns and content.

##### 3.4.2.3.1 Missing term

The first text is from a message thread entitled 'Missing term' from September 2007 (see Appendix 6.3). A Commercial Partner, BS, queries a difference between two ontology files, since a GO term identification number for an obsolete term cannot be found in the latest file versions. This runs counter to the speaker's expectations – deleted or deprecated terms should still be recorded in subsequent versions of GO files. A Senior Developer, MH, responds, explaining that the missing terms should indeed still be secondary identification numbers for two existing GO terms, but have somehow been deleted in the latest edit of the Gene Ontology files.

There follows a short exchange of two more messages, in which BS asks when a system for tracking historical changes to the ontology files will be implemented. MH confirms that when a shared database of GO files goes live, this feature will be a part of the database.

This short text demonstrates how much technical language has built up around the Gene Ontology project in a relatively short time, which lots of jargon substituting for various tools and practices used in editing ontology files. This language has largely been invented to serve the GO developers' work, but its effect is also to act as a barrier to non-expert users who may otherwise contribute to the ontology.

Transitivity in this text also shows how terms are treated as material objects, to be found, moved and, potentially misplaced. In this discussion about a term which is temporarily 'lost' in the ontology files, BS and MH use several different phrases which carry the same meaning which is 'GO term'. Overlexicalization is a linguistic feature which indicates objects in a community of practice which are important to that community, and clearly GO terms as things to be moved and manipulated, created and deleted, are central to the speakers on the GO mailing list.

What is interesting is how the two speakers in this 'Missing term' thread move between personal pronouns ('I have no idea...') and collective pronouns ('Hope we can...'). This is an example of users invoking group authority to press for changes, and is a common feature of discourse on the GO mailing list, especially when a speaker wants to achieve a goal. And, when it suits the aim of the

speaker, agency can be entirely removed, occluding responsibility for ontology changes, and making it unclear who has the authority to determine the specifics of ontology content. This is a classic demonstration of power in discourse, where it serves an authority agency to keep its work hidden and obscured (since this renders it more difficult to criticise).

The bureaucracy of the GO project removes the power to change the ontology from the hands of users and reserves this power for a small cadre of ontology developers. In this text, BS is the outsider, querying a small flaw in the error – BS has no authority to edit and change the ontology files to fix this error. MH does. GO Consortium ideology is one of control, since it is seen that only through control can semantic problems in biology be solved. The administration of the ontology itself is 'open' in the sense its products are not copyrighted, but 'closed' in the sense that users are absolutely not entitled to edit the ontology at will. This ethos contrasts strongly with projects like Wikipedia or resources like blogs where users are empowered to create content as they choose. Of course, resources such as Wikipedia and multi-authored blogs have devised mechanisms for the control of open user contributions [241-243] but the level and nature of that control in relation to the Gene Ontology is of a rather different order.

#### 3.4.2.3.2 Ubiquitin removal

The next text for analysis is entitled 'Ubiquitin removal' and dates from March 2000 (see Appendix 6.4). This text is more complex than the previous example, and includes more speakers contributing to the thread, and a technical discussion surrounding the deletion of a class of terms from existing versions of the ontology files.

Ubiquitins are proteins common to all eukaryotic cells. When attached to other proteins, they mark that molecule for degradation within the machinery of the cell. Here, the GO developers decide that the word 'ubiquitin' does not describe a molecular functions according to the vocabulary standards for the Gene Ontology, and discuss alternatives. The suggestion to remove ubiquitin-related terms is made by a Senior Developer, MH, on the grounds that these terms refer to the name of a class of gene products, and do not state what the function of these gene products actually is.

A Database Partner, JR, agrees with this reasoning, and the change is supported by several other Senior Developers. Since there is no canonical name in the biological terminology for the function of ubiquitins, two suggestions are made, and the ubiquitin terms are subsequently obsoleted.

This major change to the ontology is rendered agency-free by transitivity in the original proposal by MH to these terms from the Molecular Function Ontology. The person responsible for the addition of these terms is anonymous, and instead the ontology is portrayed as a structure which needs to be altered.

MH's justification rests on a standard in the ontology, which is that gene product names cannot be used as surrogates for functions, and the choice of phrase 'get rid of' is interesting for its implicature that these terms are unsightly. MH does not actually explain why these terms need to be removed, choosing instead a synctactic arrangement which assumes fellow GO Mailing List readers will implicitly understand the reasoning.

Several instances of collective pronouns and an invitation for consensus indicate that MH believes this decision meets GO standards and needs to be implemented. The only comment is from a

Database Partner, JR, who offers a complicated response with several qualifications and interrelated statements, essentially agreeing with MH. JR's comment is actually an argument against a long-standing linguistic norm in biology, which is the way biologists use gene product names when they could, and perhaps for clarity should, be talking about the functions these gene products carry out.

Regardless, several Senior Developers support MH's proposal, and ubiquitin terms are obsoleted from the ontology and replaced with newly created phrases to represent their existing functions. This text illustrates how Senior Developers can exercise their power to make ontology design choices without objection or criticism.

#### 3.4.2.3.3 Proposed definitions for reproduction terms in process.ontology

A third discourse analysis addresses a lengthy discussion from December 2001 about a high-level change in the Biological Process Ontology (see Appendix 6.5). A Senior Developer, MH, 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 effect 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, MA, eventually attempts to resolve the dispute.

In contrast to the previous discourse text in which senior developers reach consensus on a change to the ontology with relative ease, this next 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 included in the Appendix. 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.

#### 3.4.2.3.4 Less than two months. please?

The final text analysed in this section is 'Less than two months. please?' and dates from December 2001 and January 2002 (see Appendix 6.6). This short message thread is a result of the previous discussion about a definition for sexual reproduction, and a Curator, ES, voices dissatisfaction with how the Gene Ontology implements changes.

A Senior Developer, JB, responds to ES's concerns, and defends the GO Consortium's approach.

The final discourse text for analysis follows on from the previous discussion about reproduction. ES, a curator and newcomer to the GO project, questions the requirement that contentious decisions are resolved at GO meetings, rather than via the mailing list.

ES's language suggests deference to authority. Uncertainty, irony, and ES's omission of agency with respect to who actually makes changes to the Gene Ontology are indicative of a curator struggling to get changes made when confronted by a bureaucracy. ES is appealing for email as a medium for resolving contentious issues because it enables all users to voice opinions. Yet in the way this issue is phrased, ES is cautious and avoids directly accusing GO developers of keeping important discussions for major meetings.

JB's response claims that consensus is needed to make changes to the ontology, and turns ES's problem into a personal frustration. No solution is offered to ES. The discussion will be escalated to the next GO meeting, where ES will possibly be a marginal voice amongst many GO partners contributing to a definition for reproduction.

No discussion of a definition for reproduction is recorded in the minutes of meetings for 2002, and the definitions MA suggested remains today. Of note is that at the time, only 10% of GO terms were defined. The previous, extended discussion around a handful of definitions for major reproductive terms contrasts a major issue JB alludes to but does not explain, which is the need for a rapid expansion in definition coverage if GO is to garner acceptance as a standard by the National Library of Medicine.

### 3.4.3 Discourse analysis: conclusions

The mailing lists are texts created by GO developers and users as they try to resolve disputes about Gene Ontology content. Present versions of the Gene Ontology guidelines suggest that ontology development is the identification of objects and processes in reality. An ontology is intended to be objective or value-free in its representation of reality. Disagreements between developers similarly ought to be objective and value-free discussions about what exists in reality. Resolutions to disagreements ought to be the resolution of scientific errors, and not the exercise of power, by an authority or an ideology, to prioritise a set of values and beliefs.

The results to this discourse analysis provide evidence that values, subjectivity and the exercise of power do in fact play an important role in determining the content of the Gene Ontology.

Truth in the ontology is the result of a negotiation, constructed according to a consensus reached between a small group of developers and users. I have discovered 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 Gene Ontology 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.

The GO developer mailing list is a part of the scientific discourse dealing specifically with how to create classifications to support computer applications. As Foucault asks:

> "Posing for discourse the question of power means basically to ask whom does discourse serve?" (pp. 115, [244])

For whom does the GO mailing list discourse serve? The conclusion I draw is that the discourse serves the small cadre of senior GO developers orchestrating the Gene Ontology project, 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.

Yet as Foucault advises, we cannot seek to simply juxtapose ideology against truth when trying to understand the question of power in a discourse:

> ".. the problem does not consist in drawing the line between that in a discourse which falls under the category of scientificity or truth, and that which comes under some other category, but in seeing historically how effects of truth are produced in discourse which are themselves neither true nor false." (pp. 118, [244])

What we see in the GO mailing list discourse are how these 'effects of truth' are in accordance with the assumptions of ontological realism at the expense of other, potentially powerful ways for representing concepts in the sciences, as discussed in Chapter 3.2.

Although objectivity, guided by a commitment to the doctrine of realism, is an ongoing philosophical commitment in the Gene Ontology design, it is social power, exercised by senior scientists in the GO Consortium, which has disenfranchised the individual biologist in order to contrive a unified view of knowledge in the domain. This exercise of this power, legitimated by the GO Consortium, has at times superseded the doctrine of objectivism because disputes, such as over what sexual reproduction *is,* need to be resolved. There is a rhetoric at work here, a rhetoric that maintains the image of objectivity in science [245], and marginalises the idea that conceptual inventiveness and non-paradigmatic thinking are part of the growth of scientific knowledge.

In the next chapter, I will be exploring how term obsoletion acts as the clearance of old ideas from the Gene Ontology, a fresh start that draws a line under many of the seemingly out-dated and idiosyncratic classification approaches to gene products in the domain. This discourse analysis demonstrates a different feature of classification construction whereby ontology requests act as the process of erasure [246]. The addition of terms or changes to ontology relationships validates knowledge, providing a means for database curators to annotate gene products. Unsuccessful ontology requests render knowledge invisible in the schema of the Gene Ontology classification, erasing potentially relevant ways of thinking about gene products.

The Gene Ontology Consortium, in constructing a vocabulary to serve the molecular biology domain, clears and erases problematic issues and ideas. In doing so, it is manufacturing a better ontology to serve the needs of its user community. At the same time, the GO Consortium entirely obliterates categories of concepts, an act legitimated through increasingly complex ontology rules and standards.

One might argue that work practices in the obsoletion of terms are simply a reflection of the fluid state of knowledge in the sciences. Biological knowledge is in a constant state of flux. Hence the classes and relationships between classes in a knowledge representation like the Gene Ontology necessarily change as new discoveries are made by molecular biologists. The results of this chapter though strongly suggest that term definitions and relationships in the Gene Ontology are being socially created.

Bowker and Star have explored the idea of scientific categories which are socially created [246]. The concept of AIDS as a condition with an identifiable cause was initially constructed as a disease limited to homosexual men. Subsequently, the condition was extended to being a disease afflicting heroin users and hepatitis sufferers, and finally to being a condition transmissible by exposure to the HIV virus. A detailed set of codes in the International Classification of Disease (ICD) now describes the various associated diseases caused by an HIV infection. These codes are the product of a complex legal, moral, social and scientific debate about the nature of AIDS. Before these ICD codes were created though, there was no record of AIDS. Bowker and Star stress that without a classification, we cannot even retrieve information about a disease like AIDS from the historical record.

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, individual 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.

## 3.5 Term obsoletion

In the next results section, I apply a semi-quantitative, content analytical methodology 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.

### 3.5.1 Term obsoletion: how GO terms are obsolete from the ontology

#### *3.5.1.1 The process of term obsoletion*

The Gene Ontology Consortium has long had a system in place to remove terms from the ontology [247]. 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 principle reasons the GO Consortium cites for the obsoletion of ontology terms are:

1. A term is outside the scope of GO and is no longer to be used to annotate gene products
2. A term is redefined and changes meaning

GO curators propose terms for obsoletion through several different mechanisms. In the early development stages, GO curators suggested terms for obsoletion via the GO mailing lists, or directly at GO developer meetings. Once management of ontology requests was consolidated into the Sourceforge tracker mechanisms, most term obsoletion requests were submitted and reviewed by ontology editors on Sourceforge.

Obsoletion occurs through a standard protocol. Firstly, Consortium members are required to be alerted of an obsoletion proposal; this is satisfied through the Sourceforge tracker. A 14 day time limit exists for raising objections, to give Consortium members time to prepare arguments for or against obsoletions. A standard method for obsoletion and handling out-dated annotations exists, to ensure the consistency of the ontology and the amendment of annotations.

Proposals are normally discussed between the ontology editors and database curators with responsibilities for annotation. Discussions are sometimes opened up to invited domain experts at special curation meetings, and ontology users may also contribute to discussions.

If an ontology term is to be obsoleted, it is tagged as such in the ontology files and can no longer be used to annotate gene products. Gene product annotations to an obsolete term can be migrated to a new, exact alternative using the 'replaced_by' tag in the OBO file format. Alternatively, obsolete terms may have suggested, inexact alternative terms, and these are indicated by a 'consider' tag in the ontology files.

If a term changes, it is obsoleted and the new term, often with exactly the same term string, is assigned a new GO id; existing gene products annotated with the obsolete term are curated again as necessary.

#### *3.5.1.2 Why look at term obsoletion?*

The Gene Ontology project has a limited and defined scope, outlining what may or may not be included in the vocabulary. As a representation of the current, best knowledge of molecular biology, it excludes other sub-disciplines in biology. It does not represent gross anatomy, cell types or tertiary protein structure. The body of knowledge the Gene Ontology does represent is demarcated by the rules and standards governing the three GO sub-ontologies, as previously described. The scope of the project is legitimised by the assumptions of ontological realism: GO describes classes of processes, functions and cell components in reality. These classes represent but a few of the major classes of entities that biologists can potentially study, and potentially incorporate into theoretical arguments about biology. This is reductionist thinking.

The scope of the Gene Ontology is limited to the kinds of activities and structures gene products can contribute to at the sub-cellular level, across different species. Activities and structures at the sub-molecular level or above at the multi-cellular level are outside the part of reality GO describes. Term obsoletion provides a procedure for eliminating concepts outside this scope, and for removing concepts which undermine the logical consistency of the ontology, as designed by its authors.

The project's scope and the rules governing term obsoletion have developed organically with the project. As discussed previously, the Gene Ontology has never subscribed to any special standards in controlled vocabulary construction, and many of the ontology standards, file formats and partner ontologies which contribute to the look and feel of GO were designed by GO collaborators [99, 106, 161, 218, 248-251].

How then has the Gene Ontology Consortium devised rules and justifications for term obsoletion?

A GO meeting in Chicago 2004 discussed the problem of obsoleting terms [252]. The meeting minutes describe term obsoletion as a problem centred upon a "...tension between 'ontological purity' and 'scruffy necessity'". There is a sense in the minutes of the October 2004 meeting that this tension developed from the requirements of the information systems designers and the needs of users in the biosciences community. The systems designers created a logically consistent ontology, built in such a way as to permit reasoning across the GO graphs. Reasoning in this context means the derivation of new facts not explicitly stated in the ontology. In the case of the Gene Ontology, this is the derivation of new facts about gene products or predictions about the meaning of gene expression study results. In order to reason, the ontology must be logical and follow basic formalisms, what the GO Consortium refers to in the quote above as 'ontological purity'.

On the other hand there are the users looking for terms they understand in order to annotate gene products in species databases, and to use these annotations to test hypotheses. The same GO meeting discusses how important it is that these users accept the Gene Ontology if the project is to be demonstrably useful and secure future funding. The 'scruffy necessity' is not only a practical necessity, of the ontology reflecting the knowledge of the molecular biology domain in forms and language its users find familiar. There is also a financial necessity. The GO Consortium is conscious that only by relaxing the ontology standards can the Gene Ontology assume a form which will attract a wide enough audience to guarantee the grants it needs to survive.

The issue of term obsoletion, of what is removed from the Gene Ontology, was then and is still now, controversial. 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.

Ceusters proposes that term deletions are one way in which ontology quality can be audited [253]. The ontological realism to which Ceusters and the Gene Ontology adhere interprets obsoletion to mean a change in reality, a change in current scientific understanding, or a correction of a mistake in the ontology. Rather than being the correction of mistakes, this chapter will address term obsoletion from a sociological perspective and apply a semi-quantitative technique to analyse the reasons why terms get obsoleted from the Gene Ontology.

### 3.5.2 Term obsoletion: methodology details

All previously obsoleted terms are included in newer versions of the Gene Ontology files. Therefore, each of the three sub-ontologies (Cell Component, Molecular Functions and Biological Process) can be searched for previously obsoleted terms. This provides a simple means for users to track obsoletions, and this is especially useful when the ontology changes and old annotations are moved to new or alternate GO terms.

The methodological approach aims to ascertain the reason for individual term obsoletions by searching all available Gene Ontology documentation. 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 Sourceforge tracker files, the Gene Ontology mailing lists and the minutes of GO meetings.

For each term obsoletion included in these results, the reason for the term removal was categorised, along with notes behind the obsoletion and the dates for proposal and implementation (see Table 15). 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.

**Table 15: Term obsoletion analysis workflow**

1. Downloaded 'Terms, IDs, secondary IDs, obsoletes' file from the 'Ontology related-files' section of the GO website (retrieved 21 October 2010). This file is a simple table generated by an SQL query of the GO database designed to retrieve all terms which have been merged or replaced
2. Imported this data to an Excel spreadsheet, filtered for Function Ontology terms only, identified all obsolete terms as those marked with an 'obs' tag
3. This subset of obsolete terms was sorted by GO ID number, and each term was analysed in turn
4. In order to determine why each term was obsolete, searched for the term string or GO ID in the following resources:
    a. AmiGO online search
    b. Sourceforge Ontology Tracker
    c. GO Mailing lists 1999 – 2010
    d. GO Consortium full minutes of meetings
    e. Archived ontology files
5. Resources were searched for discussions or mentions of why a term was obsolete, and detailed explanations of the reasoning behind the obsoletion
6. Recorded the following metadata about the obsoletion:
    a. Date obsoletion was proposed
    b. Date obsoletion was implemented
    c. Category for obsoletion
    d. Short free text notes on why the term was obsoleted

### 3.5.3 Term obsoletion: analysis of reasons for why terms are obsolete

As of 21 October 2010, the Gene Ontology consists of 32826 terms divided across the three sub-ontologies as shown in Table 16 below.

Table 16: Number of terms per sub-ontology (October 2010)

| Sub-ontology | Cell Component | Molecular Function | Biological Process |
|---|---|---|---|
| **Number of terms** | 9684 | 2892 | 20250 |
| **Number of obsolete terms** | 123 | 806 | 519 |

A total of 1448 GO terms are marked as obsolete, with 8% of obsoletions made in the Cell Component Ontology, 36% of obsoletions made in the Biological Process Ontology and the majority of obsoletions, 56%, made in the Molecular Function Ontology.

The Molecular Function Ontology consists of a total of 9684 terms, of which 806 have been made obsolete. This represents an obsoletion rate of 8.3%. The obsoletion rate across the whole Gene Ontology is an average of 4.4%, meaning that for every 100 terms added to the ontology, over 4 terms will eventually be obsoleted.

#### *3.5.3.1 Reasons for term obsoletion*

Working sequentially by GO identification number (a seven digit number preceded by 'GO:'), through the Molecular Function obsoletions, 403 obsoleted terms were categorised. Eight distinct categories for why terms are obsolete were identified (see Table 17 below).

Table 17: Categories for Molecular Function Ontology obsoletions, and reasons

| Category for obsoletion | Reasoning |
|---|---|
| **Cell component term** | The term refers to a structural component in a cell |
| **By structure, not ligand** | The term defines a class according to common structural feature and not the ligand a product may bind |
| **Complexes** | The term refers to a complex of two or more protein subunits |
| **Compound of more than one other ontology term** | The term is a combination of more than one concept |
| **Gene product** | The term is the name of a single gene product or a class of gene products |
| **Non-existent function** | The term is not a function |
| **Process term** | The term is a process |
| **Protein family** | The term describes a class of proteins sharing a common goal |

The most popular reason for term obsoletion in the Molecular Function Ontology is 'gene product' with 66.00% of all terms obsoleted for this reason (see Table 18). Terms which needed re-categorising as Biological Process terms represent 9.93% of the total term obsoletions in the Molecular Function Ontology, whilst compound terms were the third most popular category for obsoletion, representing 8.19%.

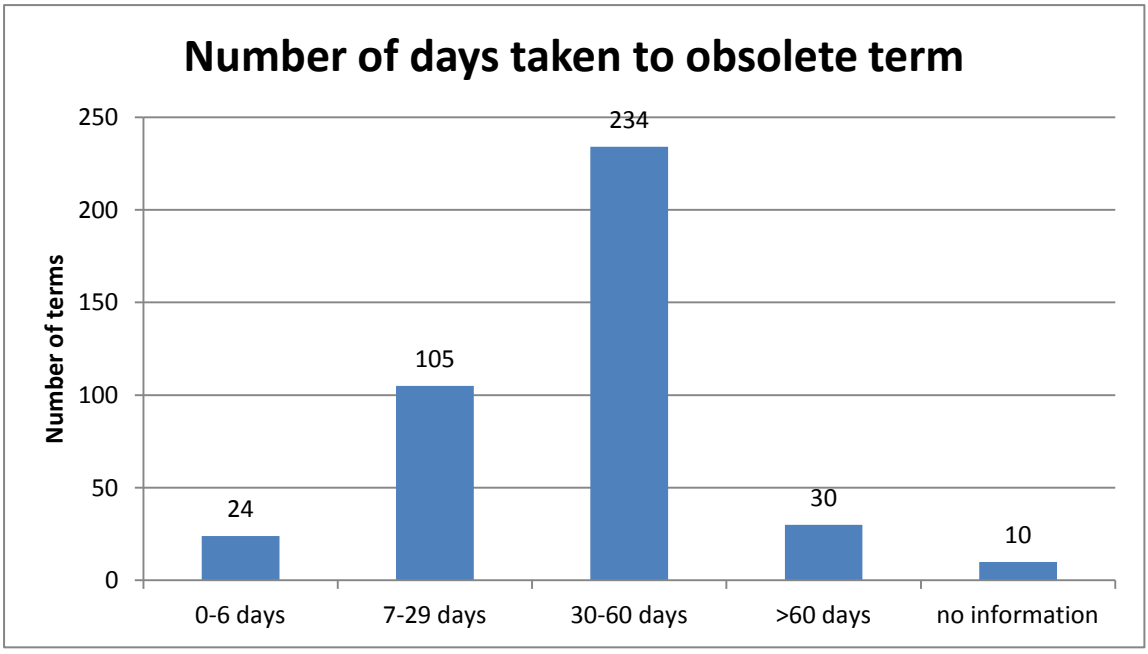
Table 18: Reasons for obsoleting molecular function terms

| REASON FOR OBSOLETION | COUNT | % TOTAL |
|---|---|---|
| Cell component term | 16 | 3.97 |
| By structure, not ligand | 12 | 2.98 |
| Complexes | 6 | 1.49 |
| Compound of more than one other ontology term | 33 | 8.19 |
| Gene product | 266 | 66.00 |
| Non-existent function | 26 | 6.45 |
| Process term | 40 | 9.93 |
| Protein family | 4 | 0.99 |
| TOTAL | 403 | 100 |

#### *3.5.3.2 Time taken to obsolete terms*

Most terms (59.6%) are made obsolete in 30-60 days from a term obsoletion proposal first being made (see Figure 12). 6.11% of term obsoletions are proposed and implemented in less than a week, and 7.63% of term obsoletions in the Molecular Function Ontology took over 60 days to be put into action.

Figure 12: Number of days taken to obsolete term




#### *3.5.3.3 Procedures for term obsoletion*

Early obsoletions, before the implementation of the Sourceforge tracker for handling obsoletion proposals, were principally recorded and discussed on the Gene Ontology mailing list.

For example, GO:0001597 'apelin-like receptor' was obsoleted in December 2002. It had no definition prior to obsoletion, a common problem with many GO terms in early versions of the ontology. AmiGO, the ontology browser, reports that the term was obsoleted because "...the function it represents does not exist" and hence this term was categorised in the results as 'Non-existent function'. For further information about why 'apelin-like receptor' is not a molecular

function, a detailed search of the GO mailing list files is necessary, and reveals a message circulated by a GO editor at the EBI in December 2002, announcing a series of obsoletions. In this message the editor reports the following:

> "I'm not quite sure what was intended with this term, but apelin is a *ligand* of the APJ receptor, so it doesn't make sense. As far as I can tell, there are no receptors that are 'apelin-like' (apelin is a peptide) so I vote we make this term obsolete - it has no annotations to it."

A valid molecular function according to the Gene Ontology standards is the binding of a ligand species by a gene product. In this example, 'apelin-like receptor' breaks this condition, for the term 'apelin-like' is not specific enough and may only refer to an imaginary class of peptides which resemble apelin. Apelin was discovered in the late 1990s, and binds to a signalling receptor found on cells throughout the human body. Apelins are not a family of peptide molecules – there is only one apelin, and the molecular function of apelin orthologs in different species is now represented by the GO term, GO:0031704 'apelin receptor binding'.

Another example of a term obsoletion, this time using the Soruceforge tracker, is GO:0019040 'viral host shutoff protein'. The Sourceforge tracker indicates that a new request was opened on 13 February 2003 by a GO editor, suggesting that this term, amongst several others, ought to be removed because it is "...not a function" – this is the extent of the explanation. A week later the term is obsoleted without any comments from users, and several GO terms suggested as alternatives for annotating relevant gene products. The reason for obsoletion is therefore categorised as 'Non-existent function' with no further information on the explanation.

A GO user later opens another GO tracker request, asking why 'viral host shutoff protein' was obsoleted. The user is directed back to the previous tracker thread and the reason that the term is not a function.

Through reading messages and discussions like these, three broad mechanisms for processing an obsoletion are apparent.

Firstly, individual terms may be suggested as candidates for obsoletion, normally via the 'GO Requests' Sourceforge tracker. 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.

Secondly, larger groups of terms may be proposed for obsoletion. These groups are normally suggested by GO editors or annotators, and the proposal recorded via the Sourceforge tracker. 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.

Thirdly, 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 Gene Ontology website [254].

#### 3.5.3.4 Who proposes GO term obsoletions?

Sourceforge offers some limited means for determining who proposes GO term obsoletions, and their likely institutional affiliations. Everyone submitting a request to the tracker is required to register for a Sourceforge account, and if these users are involved as developers on the Gene Ontology, it will be listed on their profiles.

In addition, many Sourceforge users register using their real names, allowing simple literature searches of PubMed and the Internet to ascertain their degree of involvement in the Gene Ontology project when they submitted their ontology request.

Reading the requests for obsoletions, it seemed likely from recognisable names and user accounts that most requests for removal of terms are made by GO developers and annotators. As a short test of this hypothesis, all requests on the Sourceforge GO ontology tracker for the year 2004 were searched for any mention of obsoletions. This retrieved 102 message threads. Each thread was then categorised according to the likely affiliation of the ontology request author, based on whether they were GO Developers (according to their Sourceforge profile and attendance at GO Consortium meetings in 2004), annotators using GO to index gene product databases (based on attendance at GO annotation camps in 2004 and their publication record) or if the requestor was likely to lie outside the main Gene Ontology development or annotation teams. The results are shown in Table 19 below.

**Table 19: GO project affiliation for term obsoletion requests listed on Sourceforge for 2004**

| Relationship to the GO Project | Number of requests |
|---|---|
| **GO Developer** | 65 |
| **GO Annotator** | 32 |
| **No affiliation to GO** | 5 |

This simple test validates the impression gained whilst reading the term obsoletion threads on Sourceforge that most proposals for deletion of GO terms came from the developers or annotators closely affiliated with the project, rather than the wider biosciences user community. Rarely were objections raised to term obsoletions by individuals who clearly had no affiliation to the GO Consortium as a developer or annotator, and this pattern extended to the general discussions about whether terms should be kept or obsoleted.

The response of the GO developers to ontology proposals made by users with no direct affiliations to the project will be explored in more detail in later chapters.

#### 3.5.3.5 Trends in obsoletion categories

Eight main categories for why a GO term is obsoleted from the Molecular Function Ontology were identified.

**Restricting the data to the first five years for term obsoletion information (2000-2004), the proportion of term obsoletions attributable to different categories of reason can be plotted. This produces**

Table 20 below: the percentage of term obsoletions attributable to each category is shown for each year.

Table 20: Reasons for term obsoletion from the Molecular Function Ontology, 2000-2004

| year | 2000 | 2001 | 2002 | 2003 | 2004 |
|---|---|---|---|---|---|
| **non-existent function** | 0 | 45 | 17 | 15 | 3 |
| **gene product** | 0 | 18 | 52 | 36 | 40 |
| **cell component term** | 0 | 36 | 7 | 9 | 0 |
| **compound of more than one other ontology term** | 100 | 0 | 3 | 28 | 3 |
| **process term** | 0 | 0 | 14 | 2 | 49 |
| **classed by structure, not ligand** | 0 | 0 | 7 | 9 | 0 |
| **complexes** | 0 | 0 | 0 | 0 | 0 |
| **protein family** | 0 | 0 | 0 | 0 | 6 |

The simple message from this data is that the reasons for obsoleting terms are not distributed evenly year-on-year. In 2001, most obsoletions were attributed to non-existent functions, whereas in 2002 the fact that a term looked like a gene product name had become the most popular explanation. In 2003 a large number of compound terms were removed from the ontology, and in 2004 the most popular reason for a term obsoletion was because it resembled a biological process rather than a true molecular function.

#### *3.5.3.6 Justifications and objections to obsoletions*

What kinds of discussions do Gene Ontology developers and users have when terms are proposed for obsoletion? 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 on the GO mailing list and Sourceforge trackers.

GO developers frequently encounter problems surrounding whether a Molecular Function Ontology term might better be moved to the Biological Process Ontology. This requires initiation of the term obsoletion procedure. GO:0003685, 'DNA repair protein' was one of several DNA repair terms obsolete in September 2002. Annotations to these terms were moved to a Biological Process Ontology node, GO:0006281 'DNA repair' because it was felt by the developers that repairing DNA did not constitute a single function, but was rather a composite of several, quite separate molecular operations that brought about repair in DNA strands.

Similarly, GO:0008435 'anticoagulant activity' was obsoleted because it was felt to be better described as a set of processes. Objections were raised to the semantics of 'anticoagulant' implying that coagulation was stopped, rather than more appropriately being down-regulated. A substitute term in the Biological Process Ontology was therefore created, GO:0050819 'negative regulation of coagulation' and 'anticoagulant activity' retained as a related synonym.

A recurring problem in obsoleting terms on the grounds that they are processes is that the Gene Ontology standards do not clearly define the difference between functions and processes. The question of what a function is in biology is a long-standing philosophical problem which has never been properly broached by the Gene Ontology development team. The implications of this failure to

confront the problem of what constitutes a function in biological terms will be explained later in the thesis.

Likewise in the anticoagulant term obsoletion above, curators rarely deal with question of how an obsoleted Molecular Function Ontology term like 'anticoagulant activity' acts as a related synonym in the Biological Process Ontology. Ontological realism dictates that there is but one referent process in reality, the negative regulation of coagulation yet its relationship to 'anticoagulant activity', especially with respect to automated reasoning across the ontology, is never explained. Here we see the tension between 'ontological purity' and 'scruffy necessity' played out in the process of obsoleting GO:0008435.

The removal of Molecular Function Ontology terms on the grounds that they resemble gene products is another ongoing obsoletion issue in the Gene Ontology. Biologists commonly use gene product names as a synonym for the molecular function or functions carried out by that product. An ubiquitous example is that of enzymes. Enzymes catalyse biochemical reactions. An extensive classification system exists for enzymes called the Enzyme Classification, which divides enzymes up into different families according to the type of chemical reaction they catalyse. EC 3.4.24.7 for example corresponds to 'interstitial collagenase' which cleaves specific chemical groups found in collagens. MMP-1 is one of a number of gene products in humans which would be classified according to the Enzyme Classification as a kind of interstitial collagenase.

However the Gene Ontology annotation of MMP-1 is quite different. The word 'collagenase' is inadmissible according to GO standards because it is a gene product name. It is the name for a kind of enzyme, but does not explain what its purpose is. GO:0008133, 'collagenase activity' was obsoleted without objection from the Gene Ontology in August 2008 and all annotations migrated to the Molecular Function Ontology term GO:0004222, 'metalloendopeptidase activity'.

The Gene Ontology forbids the function of a gene product like MMP-1 to be 'exhibition of collagenase activity'. A collagenase like the human protein MMP-11 must be annotated to the GO terms metalloendopeptidase activity, calcium ion binding, zinc ion binding and peptidase activity. These different terms together describe the various functions of MMP1.

This then is an example of an important task as seen by the Gene Ontology, of specifying precise molecular functions that go beyond the general, conceptual handles that many common gene product names have traditionally served in the biosciences domain. This is one of the tasks of term obsoletion: of deleting these gene product names from the vocabulary, and substituting them with specific functions.

Another category of obsoletion identified is the removal of terms defined by a structural feature and not by the ligand bound in a process. The most common type of term obsoleted for this reason is an electron carrier. Many different types of enzymes catalyze biochemical reactions through the transfer of electrons from one molecule to another. Traditionally, protein classifications have grouped these carriers based on major structural features. For example, flavoprotein electron carriers contain a nucleic acid derived from vitamin B2, whilst iron-sulphur electron carriers, unsurprisingly, contain iron and sulphur.

The Gene Ontology opted to remove all terms which make reference to structural features, choosing instead to insist that molecular species be classified by the ligand they bind. In the case of electron carriers like the flavoproteins, the same entity is being transferred in every reaction – an electron - irrespective of whether a gene product is a type of flavoprotein or a type of iron-sulphur electronic carrier. Here is a good example of clearance in the action of term obsoletion. By refusing to admit the categorisation of molecular functions according to the physical structure of a molecule, the GO Consortium prevents the classification of gene products according to an old fashion. One can understand this action through an analogy. All the products in the dairy section of a supermarket contain milk. Yet we might argue that the chief structural component of these cheeses, butters and yoghurts is irrelevant. Their redistribution to alternative locations throughout our store is legitimate, if we choose an alternative logic for classification. Butters may be placed beside flour because they are principally used in baking, or cheeses moved near to breads because they are commonly used together to make sandwiches.

This example illustrates how a design choice in the Gene Ontology is arguably an arbitrary one.

In reading the reasoning behind why terms are obsolete, a recurring justification was that an existing term was either undefined, or had no annotations. Poor history notes often make the reason why a term was added to the Gene Ontology in the first place quite difficult to determine. As the ontology is continually edited, 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.

The GO term GO:0001615, 'thyrotropin releasing hormone and secretagogue-like receptors activity' is a representative example of an obsoletion justified by a lack of a definition. It was removed from the ontology without comment in 2008, other than an editor highlighting that the term name contained a gene product. It was an orphan term without annotations, and therefore no alternative GO nodes were suggested as replacements for GO:0001615. In the logic of the Gene Ontology, this term made no sense – it did not refer to any recognisable molecular function.

The retirement of relatively anonymous, orphan terms can be contrasted with proposals to obsolete terms with many existing annotations. In these cases, GO curators may engage in detailed discussions to determine alternative terms to which existing annotations may be migrated. The wholesale removal of many peptidase terms in July 2008 warranted a special meeting and discussions between a group of developers.

Although ontological realism would assume all nodes in the GO graph to be equal, some are clearly more important than others to the developers, and especially those term obsoletions which require considerable extra work in managing changes to annotation files.

Some term obsoletions can be described as matters of taste. In 2004, GO terms containing the word 'chaperone' were obsoleted from the ontology 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, the GO Consortium has 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.

Some users have gone so far as to suggest that it may be appropriate for the Gene Ontology to retain a system for reversing obsoletion (GO Mailing list discussion, February 2006). Several GO terms were discussed as possible candidates to be 'reinstated', since the GO Consortium view had changed: the concepts these terms referred to were unchanged, but they now fell once more within the scope of GO.

Unobsoletion was rejected as a viable option, and the terms were given new identities within the GO schema.

### 3.5.4 Term obsoletion: discussion, and how obsoletions are 'truth production'

At successive GO meetings Consortium members have agreed upon exclusions from the ontology, re-structured the graphs, and deleted existing GO terms. Term obsoletion has been an important component of the project, in order to establish its identity and to create a workable ontology that supports reasoning. Over 4% of terms added to the Gene Ontology since its inception are now obsolete, so what status do these terms now have within the paradigm of ontological realism?

Bowker and Star describe the work of deciding the content of a scientific classification system as a kind of organizational forgetting [246]. 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. Clearance 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 keeping with the adage 'Out with the old, in with the new', clearance is illustrated by Bowker and Star's work on the Nursing Information Classification (NIC).

The NIC authors attempted to construct and legitimate a more scientific foundation for nursing through a classification system for nursing work. As a profession, nursing had not traditionally enjoyed a strong theoretical basis. Clearance in the NIC required the selection of and commitment to very limited theoretical paradigms for the purposes of creating a formal classification system that could be used for describing the work of a nurse. Alternative paradigms or theoretical models for the nursing profession were necessarily marginalised in this process of authoring the NIC, as a single new taxonomic reality for the nurse's role in the health and well-being of a patient was written as a classification.

In the case of the Gene Ontology, I contend that ontological realism acts as a paradigm facilitating clearance in the construction of this controlled vocabulary for describing gene products. Old, unfashionable or disproven theories are not represented in the Gene Ontology: 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 Gene Ontology 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 acts 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.

Many terms marked for obsoletion are orphan terms – they were added to the vocabulary, but were never used to annotate any gene products in species databases. Orphan terms litter the Gene Ontology, and the GO Consortium assumes that, since these terms are created in accordance with the standards and are logically consistent with the rest of the graph, eventually gene products will be found that can be annotated to these nodes.

It is incongruous that in the early development stages, orphan terms were identified as potential targets for obsoletion, or that orphan status could be used as supporting evidence for eliminating a node from the GO hierarchy. Sections of the GO graph with very low numbers of annotations could represent areas that the developers may consider for term obsoletions, but as the project has grown, it becomes increasingly difficult to authorise large-scale changes to the ontology structure. Too much is invested, and the management of changes presents special challenges for GO developers. The consequence is that a reason for obsoleting terms early in the project is less acceptable later in the project. It is an open question as to whether term obsoletion can be recognised as an objective and scientific process in the light of this shift.

Most terms are obsoleted from the Gene Ontology because they contain gene product names. This represents a contradiction in GO Consortium thinking. One 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 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 have are gene product names cited so often as a criteria 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. Functions have long posed a problem for the philosophy of biology, and the GO project has made little effort to tackle the problem during the course of ontology development. The implications of this will be explored in later chapters, but the point I make here is that there is a strong argument for keeping gene product names as surrogates for molecular functions in the ontology, if this is what the community understands.

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? My reading of term obsoletions in the Molecular Function Ontology has discovered no clear answer to this question, and this is an interesting result.

The high proportion of senior developers directly responsible for initiating term obsoletions, the relatively short time in which most terms are made obsolete, the agreement upon a small number of justifications for obsoleting large groups of similar terms: all this is evidence of what Latour describes as the work of 'allies' in science [255]. As the Gene Ontology project progressed, key personnel recruited small numbers of like-minded individuals from species databases and with expertise in bioinformatics to join a decision-making body on the GO mailing list. Term obsoletions are quick to be proposed and consensus the norm because these GO experts are acting out their role as 'gatekeepers' to the ontology, authorising content and managing contributions from outside the project to keep the graph consistent, to keep it looking 'scientific' [215]. Diverse ways to represent gene product functions do exist, yet the process of term obsoletion in the Gene Ontology acts as a mechanism for different scientific actors to limit this diversity. The result is mutual agreement, despite the fact several important ways of representing knowledge in molecular biology are cleared in the process.

The very fact that hundreds of terms made it into the Gene Ontology in the first place, only to be removed at a later stage as new standards were introduced is significant. It implies that an obsolete term had some validity for the author of the term at some stage. That validity was undermined and eventually cleared away by the authority of the GO project.

In the course of this investigation, 50% of the obsoleted terms from the Molecular Function Ontology were analysed. This represents approximately 28% of all the term obsoletions across the three GO sub-ontologies, and therefore an extension of this work would be to look at the remaining terms which have been removed. In order to strengthen the validity of the data presented here, it would be ideal if my categorisations could be verified by a third-party.

Of special interest would be the term obsoletions in the Biological Process Ontology. Proportionally fewer terms have been obsoleted from this part of the GO graph, and it merits further investigation to qualify why this should be the case. Terms obsoleted from the Molecular Function Ontology often end up as new terms in the Biological Process Ontology, and as discussed previously the precise criteria for what constitutes a process over a function has not been clarified by the GO Consortium. The implications will be explored in a later chapter. Suffice to mention, the vague status of related synonyms in the Gene Ontology vocabulary parallels the expansion of terminology representing concepts in the Biological Process Ontology.

How do other controlled vocabularies handle retired terms or new, expanded terminology designed to expand on concepts previously represented by a single node in the hierarchy?

The Medical Subject Headings provided by the National Library of Medicine includes in its MeSH XML files a data element for a 'previous indexing' field which is intended "...to enable users of new Descriptors to find similar concepts indexed before the Descriptor was created" [256].

Since 2010, MeSH has produced annual releases of its vocabulary files, and lists deleted descriptors together with the terms they are replaced by. The previous descriptor becomes an entry term for the new concept, and all articles indexed with the redundant term are migrated to the new descriptor.

The Library of Congress will routinely cancel subject headings and suggests alternatives, such as deleting 'Die Hard films' and replacing with a genre like 'Action and adventure films' [257].

This is performed by the Policy and Standards Division, discussed in a briefing paper, and normally appears in the Cataloguing Service Bulletin. Most of these are scope adjustments, or correct culturally bias headings [258].

Like MeSH and Library of Congress subject headings, the Gene Ontology removes terms, but its procedures are slightly 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, are obsoleted in this way because, according to the GO approach, they do not represent instances in reality – they are gene product names or things in biology which are not molecular activities. But as previously described, 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.

In conclusion, the poor management of term obsoletion in the Gene Ontology, justified by ontological realism and the exclusion of certain classes of concepts, makes it very difficult for users to understand why terms have been deleted, and whether any alternative exists for indexing gene products under similar concepts.

It is worth mentioning the role of erasure in the Gene Ontology, as it is subtly different to this process of clearance I have described in the obsoletion of GO terms.

Erasure, in the manner described by Bowker and Star, is the deletion or omission of content from an information system. In nursing, the nursing record is often removed from a patient's medical records for the administrative reasons, to simplify and reduce the notes to a manageable size. In so doing, the role played by nurses is filtered from the institutional memory of how patients are cared for in the hospital environment.

Erasure in the context of the Gene Ontology can be seen in the process of annotating gene products. In creating an annotation, in choosing a Gene Ontology term to associate with a protein or transcript entry in a database, all the rich and complex empirical data created in establishing an association such as 'Protein X is a component of the nuclear membrane' is lost. The debates, contentions and theoretical context in which biological facts are discovered (or manufactured) are deleted in the implementation of the Gene Ontology. By omission, a scientific classification like the Gene Ontology 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 Gene Ontology model enables institutionalised forgetting via erasure.

Ontological realism has guided the Gene Ontology 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 Gene Ontology has chosen to ignore the rich and varied ways in which biologists, and indeed all scientists, understand theoretical problems and, more importantly, conceive new ideas. The Gene Ontology takes a critical rationalist approach to the problem of justifying scientific knowledge. It 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.

The Gene Ontology assumes beliefs can be justified by the empirical method. Critics of theories of justification like Karl Popper have argued that this kind of scientific knowledge is never reducible to undeniable, basic truths [259]. If one falls into the 'justificationist trap', a very weak system for validating new knowledge is established in the sciences; the accumulation of positive evidence is accepted as proof that a theory is good and true. Popper's solution to this weak system was to test new ideas against falsification – can a theory survive against strong tests which, should it fail, would require the theory to be dropped?

The Gene Ontology procedures for term obsoletion are intended to make the ontology 'more true'. 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 biologist understand concepts. Simple examples revealed by this analysis of term obsoletion include the usage of gene product names as placeholders for complex concepts, the necessity for compounds of functions and anatomical locations, and the crossover between what is understood as a molecular function and what might be a biological process.

The seemingly objective, empirical rules governing what ought to be excluded from the Gene Ontology are tentative at best, and biased at worst. Therefore the criteria determining the ontology content are constructed by social activities and interactions, and term obsoletion is a process akin to the manufacturing of scientific truth. The Gene Ontology Consortium is a gatekeeper to an authored representation of molecular biology, warranted and justified by the social allies it has recruited to support its work.

## 3.6 Content analysis

In this final results section a content analysis of Gene Ontology article abstracts and full texts assesses whether authors cite GO data in peer-reviewed papers according to GO Consortium data citation policies.

### 3.6.1 Description of GO papers dataset

Thus far several analyses and arguments to support the contention that a concept-based view for ontologies is a tenable basis for building useful vocabularies to support knowledge discovery in molecular biology have been presented. Furthermore, dropping ontological realism in favour of a concept-based view could make better ontologies for the biosciences.

Concept analysis showed that even a single concept in the Biological Process Ontology can be viewed from a plurality of equally valid philosophical perspectives, none of which can necessarily be resolved into a single node in the Gene Ontology. Analysis of the Gene Ontology vocabulary standards indicated that there are severe limitations to the way in which GO handles synonyms, variations in biological language, and in how GO represents the current state of biological knowledge.

Term obsoletions show that the Gene Ontology Consortium have contrived a limited set of reasons for removing terms from the vocabulary, facilitating a form of clearance which erodes what the Consortium views as outmoded ways of thinking about how to categorise biological functions. Through discourse analysis, it was shown that social relations do play an important role in determining the content of the ontology, and that the exercise of power and authority by senior editors risked erasing popular knowledge forms from the ontology. At the basest level is the refusal to admit tentative or hypothetical knowledge into the Gene Ontology. Only entities and relationships deemed to exist in reality, according to the consensus view of a majority, are permitted into the ontology. And even elementary concepts in the biosciences, such as reproduction, have proven to be exceedingly difficult to define in terms that biologists from any sub-discipline in the domain may agree upon.

In this final chapter of results, the ways in which molecular biologists use and report usage of the Gene Ontology is appraised, and the limitations of the classification as a universal vocabulary are illustrated. The Gene Ontology is designed to represent the best, current knowledge of the domain, and justifies its existence through the application of the vocabulary to the computer-aided analysis of genes, their expression, and their potential roles in living organisms. As discussed in the introductory chapter, the problem for the analysis of biological data is the extreme complexity of the systems under observation. The structure and logical consistency of the Gene Ontology, coupled with the many millions of annotations made between ontology terms and gene product records in species databases, is intended to make this task of understanding what is happening at a molecular level in biological systems much, much easier.

How then are molecular biologists using the Gene Ontology, and do they follow the GO Consortium's simple advice on how to report results supported by GO annotation data, to ensure that these results are reproducible and scientifically justifiable?

What follows in this chapter is a content analysis of a set of peer-reviewed scientific articles which report usage, in some form, of the Gene Ontology. These papers are a subset of a bibliometric dataset (see Appendix 6.7 ), and only include papers published in 2009.

The procedure for analysing papers is shown in Table 21. In brief, 374 GO papers were published in 2009. Of these, 163 papers had been cited at least once according to Web of Science metrics (last indexed 02 February 2010). The abstracts of these 163 papers were read in order to identify which papers 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.

The full texts of papers reporting a GO analysis were read for several attributes deemed to be important by the GO Consortium. The papers were searched for any reference to the GO ontology file versions used in the analysis; the Gene Ontology 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. If a specific software tool was used for the GO analysis, this was recorded, along with the software version, as developers sometimes provide GO file version information embedded in these software tools.

In order to appraise how authors were using GO terms in their reporting of empirical results, full texts were checked for several features. If analyses were restricted to a particular sub-ontology of the Gene Ontology (Biological Process, Molecular Function or Cell Component) then this was noted. For the most popular GO papers (according to citation metrics), the number of GO terms reported in the results of GO term enrichment analyses either in the body of the text, in separate tables, or in supplementary material files, was recorded.

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 [260]. In short authors are suggested to cite the canonical Gene Ontology paper by Ashburner *et al* [214], 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.

Table 21: Procedure for content analysis of GO papers

| STAGES FOR PROCESSING ARTICLES FOR CONTENT ANALYSIS |
|---|
| 1. Filtered bibliometric dataset for all GO papers published in 2009 – 374 papers identified |
| 2. All papers with at least once citation were included and taken forward to the next stage – 163 papers in total |
| 3. Each abstract for these 163 papers was read to identify whether a GO analysis had been performed on empirical data – 113 papers identified |
| 4. *Abstract* of each paper read and the following content variables recorded:<br>a. Number of sentences in abstract<br>b. Abstract sentence in which 'ontology' is mentioned<br>c. Categorisation of whether 'ontology' mention is specific or general<br>d. Species on which GO analysis is performed<br>e. Material on which GO analysis is performed: gene, RNA or protein<br>f. If 'ontology' mention from 'c' is specific, count the number of functional categories mentioned as enriched in GO analysis |
| 5. For each GO analysis, *full text* of paper read and the following content variables recorded:<br>a. Whether GO version is reported in the paper<br>b. What type of GO tool is used in the GO analysis<br>c. Whether version of GO tool is reported |
| 6. Finally, top cited GO papers published in 2009 (papers >3 citations) were read for the following attributes:<br>a. Sub-ontologies used in analyses<br>b. Number of GO terms mentioned in full text<br>c. Number of GO terms mentioned in tables<br>d. Number of GO terms mentioned in supplementary materials<br>e. Full text notes recorded for papers with marked deviations from GO data citation policy |

### 3.6.2 Results: how authors use and report GO terms

#### *3.6.2.1 Description of GO term enrichment analysis papers published in 2009*

In reading the abstracts for all the GO papers published in 2009, it quickly became apparent that not all articles reported the analysis of empirical data using Gene Ontology files. In my corpus of papers, some articles mentioned the integration of GO terms into existing bioscience databases, the application of ontology files to the modelling of biological systems, or the processing of GO terms as a small part of a larger bioinformatics software package.

My primary aim in this content analysis though is to try and understand how molecular biologists report the usage of GO terms in interpreting and understanding complex empirical data, a process known commonly in the domain as a 'GO analysis' or 'GO term enrichment analysis'.

These GO term enrichment analyses have become increasingly common in the biomedical literature since the inception of the Gene Ontology project. A GO analysis will use ontology files provided by the Consortium coupled with sets of annotation data (where GO terms have been used to index often many hundreds of thousands of gene products in species databases) in order to make inferences about the functions of sets of genes.

A simple example is an experiment performed by Stapleton and Chan [261] designed to understand the toxic effects of a compound on the brain. Rats were exposed to a toxic agent, their brain tissue removed and RNA extracted from this tissue. RNA levels indicate which genes are active in a tissue at a given time, therefore the authors subjected their RNA sample to a microarray analysis to measure

the increase or decrease in RNA expression of many thousands of genes at the same time. They were left with a list of 277 genes which were differentially expressed. By employing the methodology of a GO term enrichment analysis, the authors used ontology and annotation information to discover the functions and processes these 277 genes were involved in, and used statistical methods to assess whether particular types of functions were over-represented. They concluded from this GO analysis that rats exposed to the toxic agent did undergo changes in genes associated with neuronal development and function, and based on this inference were able to propose mechanism for the damaging effects of this neurotoxin.

Table 22 below shows the proportion of papers published in 2009 which could be categorised as GO analyses like the Stapleton and Chan paper described above. The proportion of GO papers covering topics not describable as GO enrichment analysis in 2009 was 31.0%, a figure comparable to the 27.5% of non-analysis GO papers indicated in the bibliometrics results in Appendix 6.7 . This is important since a significant proportion of GO papers published every year are closely allied to the discipline of mathematics, bioinformatics and computer science. However these results focus on molecular biologists – I want to find out how laboratory scientists use and report usage of the Gene Ontology, since this controlled vocabulary is designed primarily to facilitate knowledge discovery work amongst this core group of users.

**Table 22: GO paper sample, categorised by type**

| Paper type | Definition | Count |
|---|---|---|
| **Analysis** | Analysing, profiling or classifying genes or gene products using GO terms | 113 |
| **Database** | A structured online resource which includes GO term and/or annotation data | 19 |
| **Model** | A computational model, algorithm or method dependent on GO terms | 24 |
| **Software** | A computer application, usually identified by the a proper name, which uses GO data to analyse empirical data | 7 |

The distribution of journals publishing GO papers in 2009 is indicated in the Table 23 below. The top two journals publishing Gene Ontology papers are BMC Genomics and Nucleic Acids Research, which together account for over 20% of all the publications. It is not surprising that a core set of journals are publishing a large proportion of papers related to Gene Ontology work, but what is interesting is the featuring of GO analyses in a very large number of individual journals. This suggests that users across the breadth of the molecular biology domain have engaged with and applied the products of the Gene Ontology project.

Table 23: Top journals publishing GO papers in 2009

| Journal | Count | Percentage of all papers |
|---|---|---|
| **BMC Genomics** | 17 | 10.4 |
| **Nucleic Acids Research** | 17 | 10.4 |
| **Bioinformatics** | 7 | 4.3 |
| **BMC Bioinformatics** | 5 | 3.1 |
| **PLoS ONE** | 5 | 3.1 |
| **Genomics** | 3 | 1.8 |
| **Journal of Proteome Research** | 3 | 1.8 |
| **Journal of Biomedical Informatics** | 3 | 1.8 |
| **Plant Physiology** | 3 | 1.8 |
| **Other (2< instances in set)** | 100 | 61.3 |

To gain a feel for the kind of subjects covered by these GO papers, all major MeSH headings for the 163 GO papers were extracted and counted. A total of 554 major MeSH headings were found to categorise this document set, of which 338 were unique. The most common headings are shown in the Table 24 below. This set of MeSH headings accounted for over 22% of all the subject headings describing these GO papers, and give some indication of topicality for the documents.

'Gene Expression Profiling' was the top MeSH heading which is consistent with the normal application of Gene Ontology files in the molecular biology domain. Software and database headings also featured strongly, and corresponded to the type of non-GO enrichment analysis paper described above, where GO files are reported as being used as part of a software package or online database for gene products.

Interestingly, 'Arabidopsis/ge [Genetics]' features high in this table. *Arabidopsis thaliana* is a model organism in plant biology and The Arabidopsis Information Resource (TAIR) Database [262] has been a long-standing partner organisation within the Gene Ontology Consortium. Papers exploring Arabidopsis biology using the Gene Ontology are a reflection of strong uptake of GO analysis methodologies within this specialist community of the molecular biosciences.

Table 24: Most popular major MeSH headings used to categorise GO enrichment analysis papers

| Major MeSH heading | Number of papers with this heading |
|---|---|
| **Gene Expression Profiling** | 22 |
| **Computational Biology/mt [Methods]** | 13 |
| **Databases, Genetic** | 11 |
| **Software** | 9 |
| **Gene Expression Profiling/mt [Methods]** | 9 |
| **Databases, Protein** | 8 |
| **Gene Regulatory Networks** | 7 |
| **Gene Expression Regulation** | 7 |
| **Proteomics** | 6 |
| **Arabidopsis/ge [Genetics]** | 6 |
| **Genomics/mt [Methods]** | 5 |
| **Gene Expression/de [Drug Effects]** | 4 |
| **MicroRNAs/ge [Genetics]** | 4 |
| **Expressed Sequence Tags** | 4 |
| **Genome** | 4 |

#### *3.6.2.2 Results from content analysis of abstracts*

The next stage of the content analysis looks at the text of abstracts for each GO analysis paper.

All GO analysis papers with at least one citation (113 papers) were coded according to whether a reference to the Gene Ontology project in any sentence of the abstract was either:

- GENERAL : The Gene Ontology is mentioned, but no clear reference is made to how the ontology is used to analyse experimental data, nor are enriched GO terms stated in the description of the results
- SPECIFIC : A clear reference is made to using the Gene Ontology in the interpretation of experimental data; a specific GO tool may be mentioned, or the results of GO term enrichment analysis cited in informing the main conclusions of the paper

Based on these criteria, 29 GO analysis paper abstracts made a general reference to the Gene Ontology in the abstract with no further information regarding precisely how ontology or annotation files were used to inform the results. 84 papers were identified where a clear methodological statement was made declaring how the Gene ontology was used in the interpretation of the empirical data related to the experiment.

For each GO analysis paper, the abstract and, if necessary, the full text, was read in order to determine the principal species of the model organisms used in the experiment. Are GO analyses limited to a small number of species, reflecting the specialized nature of the field, or are authors from across the spectrum of potential model organisms finding uses for the Gene Ontology?

Table 25 below shows the species tissue type used for analysis. Experiments using GO analysis of human tissue are the most popular, with mouse, plant and rat ranked below. The size of the 'other' category is significant, with GO analyses covering a diverse range of model species including coral, cockroaches and bees. These species have yet to be manually annotated, yet algorithms employing sequence similarity with gene products in other species allow researchers to perform GO enrichment analysis even for more obscure model organisms.

Table 25: Main species type for GO analysis

| Species | Count |
|---|---|
| **human** | 47 |
| **mouse** | 19 |
| **plant** | 8 |
| **rat** | 8 |
| **fish** | 5 |
| **cow** | 4 |
| **bacteria** | 3 |
| **other** | 19 |

A GO enrichment analysis may take one of three main types of biomolecule as its source of empirical data. Some analyses use genetic information, this being the primary sequence of the genomic code, as the source for a GO enrichment analysis. This approach is often used in linkage analyses to study the differences between two different genetic backgrounds, such as individuals with susceptibility for a disease and a control group. Other GO analyses use protein expression data to study biological pathways in cells and tissues. Protein data is typically more expensive to generate, but is viewed in the biosciences as a stronger source of evidence in analyses since the actual levels of expressed proteins are being measured. Thirdly, GO enrichment analysis may look at RNA expression data. This kind of data is very common in the biosciences domain now due to the dropping cost of microarray chips and the ease with which software packages can be used to interpret even very large, complex datasets.

Table 26 below shows the distribution of source material for GO enrichment analyses identified in the 2009 paper set. Analyses of RNA expression data were the most common, followed by protein material and finally analyses looking at genomic sequences.

Table 26: Type of biomolecule analysed using the Gene Ontology

| Biomolecule analysed | Count |
|---|---|
| **Gene** | 9 |
| **Protein** | 17 |
| **RNA** | 87 |

#### *3.6.2.3 Results from content analysis of full texts*

Top cited (>3 citations) full text articles on GO enrichment analysis papers were read for any citation information for the specific Gene ontology file versions used in the analyses.

Of 32 full text papers read, 2 papers were found which gave enough citation information to ascertain the ontology file versions used in the GO analysis. This equates to a data citation compliancy rate of 6.3% for authors using the Gene Ontology.

This same set of 32 papers were read for reporting of the software, algorithm or bioinformatics tools used to analyse empirical data for the enrichment of GO categories. A total of 34 mentions of tools were found in the full text papers. The most popular GO analysis tools are shown in Table 27 below, with in-house solutions ranked as the most popular, closely followed by the DAVID analysis suite [263].

In-house solutions accounted for 20.6% of GO data analysis tools. Authors provided very little information of precisely how they used be-spoke statistical methods for generating GO term enrichment lists. Several other popular GO analysis tools 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 the GO versions utilised by the software.

**Table 27: Gene Ontology analysis tools reported in papers**

| Tool used | Count |
|---|---|
| **In-house** | 7 |
| **DAVID** | 5 |
| **GeneSpring** | 3 |
| **Ingenuity Pathway Analysis** | 2 |
| **EASE** | 2 |
| **Other** | 15 |
| **Total** | 34 |

Authors may restrict GO analysis to one or more of the three Gene Ontology sub-ontologies. 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 in Table 28 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.

**Table 28: Sub-ontologies used in GO analyses**

| Sub-ontology used for analysis | Count |
|---|---|
| **Function, Biological Process and Cell Component Ontologies** | 22 |
| **Biological Process Ontology only** | 6 |
| **Function and Biological Process Ontologies** | 2 |
| **Biological Process and Cell Component Ontologies** | 1 |
| **Cell Component Ontology only** | 1 |

Full texts articles were read for mentions of GO categories in the results or discussions sections. For the 32 papers analysed, the average number of GO terms cited in the full text is 6.4 (standard deviation 7.4) with 6 articles failing to give any information about the specific names of enriched GO terms. No relationships were found between the average number of citations received by a paper and the number of GO terms cited in the full text.

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 of GO category names 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 (see Table 29 below).

Table 29: Typical errors and failures to comply with GO data citation policy

| PMID | Error in GO usage |
|---|---|
| **19615732** | GO categories not reported in results or discussion section, with authors re-wording enriched GO terms into domain specific language forms such as 'protein turnover' or 'ubiquitin biology' |
| **19615732** | Bespoke categorisations created to group several hundred Biological Process Ontology terms under 14 author-generated subject headings |
| **19018450** | Authors refer to the GO category 'chemokines' which is non-existent in the Gene Ontology vocabulary; the closest synonym is the GO process term 'chemokinesis' |
| **19435504** | GO term 'vacuole organization' mis-quoted as 'vacuolar organization'; although the meaning of the term is clear in context, there is no synonym control in the Gene Ontology for 'vacuole' and 'vacuolar' |
| **19148281** | Diagram makes reference to a non-existent GO term 'cytochrome C release' |
| **19372578** | Selective reporting of only some enriched GO terms in the results section, with emphasis placed on the significance of categories the authors feel are most relevant to their argument |
| **19372578** | Failure to distinguish terms derived from the three GO sub-ontologies, with Functions, Processes and Cell Components merged into a single figure |
| **19148281** | Authors do not make clear whether all three sub-ontologies are used in an analysis; same paper makes reference to now obsolete GO term 'caspase activity' |
| **19155294** | A restricted set of Gene Ontology terms (much like a GO Slim) are used in a GO analysis, yet the authors provide no details on the ontology branches used, nor the reasoning behind the term selections |
| **19240082** | Selective presentation of particular enriched GO categories |
| **18814146** | Gene Ontology terms are presented alongside terms from alternative controlled vocabularies yet no distinction is made between the different sources |
| **18814146** | Biological Process Ontology terms conflated with Molecular Function Ontology terms in a table of results |
| **19029062** | Annotations to Gene Ontology terms created on an *ad hoc* basis by authors, and included in source data for a GO enrichment analysis; same paper reports a disclaimer for all results, based on incomplete database coverage by the Gene Ontology |
| **19211887** | Functional categories added to the results of a GO analysis on the basis of literature searches intended to substantiate the association |
| **19333917** | No information given about enriched GO categories or GO analysis methodologies |
| **19328199** | GO enrichment analysis performed, by only result given is that two samples correlated |
| **19111481** | Reporting of now defunct *sensu*-qualified GO terms |
| **19077055** | Citing of only the broadest GO categories in the results of a GO enrichment analysis, and preference given to listing protein names rather functions in tables of results |
| **19228845** | No statistical control for the bias involved in selecting small sets of GO terms to analyse for enrichment in gene sets |
| **18388159** | Molecular Function Ontology confused with Biological Process Ontology |
| **19049829** | Author-derived categories for functions of genes presented as authentic Gene Ontology terms |
| **19037624** | Source and identity of GO terms selected for an analysis not given by authors |

### 3.6.3 Discussion: comparison between GO rules and GO usage

The results of the content analysis reveal several interesting features of Gene Ontology usage in the molecular biology literature. 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.

In addition to poor data citation practices, authors use a range of analytical tools in performing GO analyses. Again, this would not normally be a problem if authors appropriately cited software versions and ontology files used by the software. However, GO analysis tools are notoriously opaque in terms of how they are manipulating annotation files and ontology records in order to achieve weighted significance scores for lists of expressed genes. Anecdotal evidence from my interviews with several GO curators suggests that authors "...choose the GO tool that gives the best result".

GO enrichment analyses are therefore acting as a 'black box' in the published literature. It is difficult to determine which ontology file versions are used in analyses, and the diversity of poorly referenced GO analysis software tools and in-house solutions for enrichment scoring translates into a large corpus of Gene Ontology-derived data in the domain which is difficult to reproduce, or even to understand its derivation.

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. This has long been recognized as a problem in indexing generally - for overviews see Bawden and Robinson [264] and de Keyser [265], and other recent research studying the issue [266-269]. de Keyser ([265], p. 47) calls it a 'platitude' that indexers cannot agree with each other. It is perhaps somewhat surprising that it persists as an issue in such the seemingly well-defined and objective molecular biology domain.

The questions remains as to why GO users are reporting Gene Ontology file usage in data analysis in this manner.

#### *3.6.3.1 Weaknesses of the technique*

In terms of a methodology for determining GO usage practices in the molecular biology domain, content analysis of published papers is open to several criticisms.

Ideally, the reading, interpretation and coding of content in papers ought to be verified for consistency with at least one other third party. Reproducibility is a concern here – is the way in which I have read and categorised the content in the corpus objective in the sense that any other coder would infer the same results? For example, I have divided my initial set of papers from 2009

into four broad categories. This was a necessary first step, to permit analysis of only GO enrichment analysis papers at a later stage, although arguably all papers mentioning the Gene Ontology could be read and coded for data citation policies or behaviours in re-wording GO terminology.

An alternative, more objective approach to dividing GO papers into types would be to rely upon MeSH major headings derived from MEDLINE records. These subject headings have at least been applied by an independent domain expert. As with most content analysis approaches though, the iterative process of reading text, inductively noting categories and themes in the corpus and coding other content in a manner consistent with these categories is open to the bias of the coder.

In the approach presented here and the semi-quantitative nature of these results, it is hoped that any research bias will have been adequately limited.

The very time-intensive procedure of reading scientific articles and associated data files for mentions of GO terminology and its presentation necessarily limited the data-set size used in this content analysis. Only papers from 2009 were studied, and it would prove interesting to extend this approach to publications from more recent years. One might expect to see better data citation policy compliance in later years as the GO project has marketed itself to users in the domain. Preliminary reading of recent publications would seem to indicate though that there has been little change in the reporting of Gene Ontology analyses over the results presented in this section.

Informal interviews with GO curators and users, presented in the Appendix, hint at more complex user behaviours in operation behind the findings in this content analysis. Semi-structured interviews and user group discussions would offer potential extensions to the methodology presented here that could offer further insights into how GO users handle ontology terminology and why GO enrichment analyses are conducted and reported according to specific conventions in the scientific literature.

#### *3.6.3.2 Main features of GO papers for 2009*

Content analysis revealed certain distinctive features of publications citing Gene Ontology files and applications. Although most articles retrieved via a literature search were applications of GO data to the analysis of gene expression studies, many publications covered wider applications. Many of these were species-specific or sub-domain specific genetic databases, collating gene, transcript and protein level sequence data on particular themes and then incorporating Gene Ontology terminology as one of several means to index content in these databases.

In addition, a significant proportion of publications reported usage of the Gene Ontology in theoretical models, normally using statistical techniques to leverage GO annotation data in predicting biological pathway involvement in uncharacterised sets of genes. In parallel with these modelling papers, numerous articles described new software supplications available for molecular biologists to integrate GO terminology data into analysing empirical gene data. These papers represent an information need in the molecular biology domain, which I would describe as the need to exploit gene product function indexing to give meaning to empirical data, in both validating pre-existing hypotheses, and in suggesting new, potentially interesting explanations for biological phenomena. The one information need is to verify theories, the other being to facilitate serendipitous discovery.

A small core of journals was found to publish the greater proportion of Gene Ontology-related publications, and this was consistent with other areas of specialist science. Of special note is the feature that 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 Gene Ontology 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 Gene Ontology.

#### *3.6.3.3 Analysis of abstracts*

The content analysis of texts from abstracts suggests that authors do make specific indications of how the Gene Ontology and GO analysis has contributed to conclusions in the research work. A large proportion of abstracts however did make only a passing reference to the Gene Ontology, 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 Gene Ontology 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.

Most abstracts relate to work in the molecular biology domain relating to mammalian model organisms such as humans and mice. The clinical relevance of these species explains their popularity as the objects of research in the domain. Quick, laboratory tests for analysing the expression of many thousands of genes using microarray chips are becoming cheaper in clinical hospital settings, and ontologies are important tools in the information systems necessary to make sense of these data sources in disease diagnosis.

The application of GO analysis to tissues derived from many other, less popular model organisms highlights the broader appeal of the Gene Ontology to biologists trying to make sense of complex genetic and gene expression data.

Content analysis of GO paper abstracts in general suggests good uptake of the Gene Ontology across a diverse number of applications and model organisms in the molecular biology domain.

#### *3.6.3.4 Analysis of full texts*

Content analysis of GO analysis full text papers paints a less positive picture of Gene Ontology 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 absent in the molecular biology domain. 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 anecdotal evidence from contact with GO experts has indicated, researchers choose whichever tools give a good result in a 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 Gene Ontology 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 will be discussed in some detail in Chapter 4.1.2 .

The content analysis here 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.

#### *3.6.3.5 Typical errors in reporting GO data*

Arguably, authors are guilty of choosing those enriched GO terms which best fit the theoretical context of their research argument, and marginalising other statistically significant GO terms 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-Gene Ontology 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 to litter 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. 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.

These observations in only a small set of GO papers are alarming in that the entire Gene Ontology 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 contrary picture. Whilst GO term enrichment analyses *might* be performed according to the exacting GO standards, it is impossible for the rest of the research community in the molecular biology domain to reproduce GO analysis results based on the information presented in most GO papers. Non-existent data citation, 'black boxed' software systems for performing analyses, confusing presentation of enriched GO term results and repeated instances of errors and re-wording in the reporting of the very Gene Ontology terminology itself present serious problems for users to have confidence in GO-derived results.

Publishers, editors and peer-reviewers could play a role in demanding that authors improved their description of GO enrichment analyses and the associated GO terminology reporting. Indeed, the GO Consortium and wider ontology community has appealed to publishers to require more of authors both in terms of using the Gene Ontology and in proposing annotations for novel gene products. These appeals have had little effect, and the molecular biology community either does not realise the importance of appropriate citation of Gene Ontology data, or finds the task overly onerous given the myriad of other requirements when publishing a peer-reviewed article.

Marketing and education is another solution, and the Gene Ontology has made extensive efforts to publicise its resources and to raise awareness of issues such as annotation and the limitations on empirical data analysis given the design of the ontology. The difficulty is incentivising the necessary investment of time, money and effort into learning about the Gene Ontology, using ontology data according to the standards, and contributing to the project.

The GO Consortium has also always tried to keep a distinct separation between the development of the ontology, the ongoing effort to annotate gene products with GO terms, and the software tools needed to analyse data using ontologies. The third issue here is of particular relevance to the results of this content analysis. Ontology files and annotation data are available 'as is' to biologists, and no special restrictions or requirements are placed on how bioinformatics applications exploit this data. Hence there has been a proliferation of rival software of variable quality deploying functional analysis features that draw on a variety of statistical methods. The multitude of analytical tools, coupled with poor reporting of those tools in peer-reviewed papers risks rendering GO analysis data untrustworthy in the first instance, and potentially wildly erroneous in worst-case scenarios. If the GO Consortium, or perhaps even species databases themselves were to offer a merged package of ontologies, annotation data and high-quality bioinformatics software packages to use these sources,

it may do much to improve the quality of ontology analysis data and reproducibility of this data in the primary literature.

# 4 Discussion

The discussion is split broadly into two main topics.

Firstly, I will address how the failure of the Gene Ontology to provide an adequate definition and philosophical framework for the concept of a biological function has caused several problems in the structure and management of the ontology.

Secondly, I will argue a case for pluralistic classifications in the molecular biology domain as a counter to vocabularies like the Gene Ontology which are built around a commitment to ontological realism.

## 4.1 Treatment of functions in the Gene Ontology

### 4.1.1 Overview of philosophy of functional explanations

#### *4.1.1.1 Why look at biological functions?*

In this next chapter, I intend to consolidate the results thus far by offering a philosophical analysis of the treatment of functions in the Gene Ontology. Aside from the Cell Component Ontology, all Gene Ontology terms are representations in one form or another of one or more *biological functions*. Every term in the Molecular Function Ontology describes the action of biomolecules performing tasks at the molecular level, whilst the Biological Process Ontology is ostensibly compounds of more than one molecular function directed towards the achievement of a biological goal.

What do I mean then by the word 'function' and why is an understanding of biological functions relevant to the design of the Gene Ontology and other vocabularies in the molecular biology domain?

The Gene Ontology takes functions for granted. It assumes the identification and description of functions to essentially be a trivial task. Where there may be technical problems in writing good definitions for functions, creating meaningful names for these functions, or in creating hierarchies of functions in ontological languages which are logically consistent, the job of spotting a biological function in the first place is considered to be obvious.

Yet functions present very complex philosophical problems. Take for example the function of a simple mechanical object like a door bell. What is the function of a door bell?

1. To ring
2. To be pressed
3. To act as a connection between a finger and an electrical circuit
4. To bring people to the front door
5. To stimulate the family dog to run downstairs
6. To regulate entry into a house
7. To transmit information that someone might be waiting outside

These are but a few potential functions for a door bell, all of which are arguably valid, yet each being subtly different. Functions for objects are often tied to purposes – what is the purpose of a door bell? To say that a door bell has a purpose is to imply that there lies behind its function some form of intelligence which has designed the object to behave in a certain way. Furthermore, we can talk about the kind of goals we want to achieve with a door bell, and this kind of talk is known as a *teleological* argument. If we intend to write logical statements describing the behaviour of a door bell given specific starting conditions, if we were aiming at making certain predictions about how door bells work, then these statements can start to look like teleological arguments. The word teleology comes from the Greek *telos* meaning final causes, and to adopt a teleological argument is to assume that there is a purpose or final cause to events in nature – and this is a distinctly 'unscientific' viewpoint to adopt.

Biologists frequently talk about the function of something in a biological process. Take for example the growth of a blood vessel, either during vasculogenesis (the growth of new blood vessels in the development of the circulatory system) or angiogenesis (the growth of blood vessels from pre-

existing vasculature). Various receptors and signalling molecules engage in a complex dialogue during the growth of new blood vessels, instructing cells to divide, move, and change shape in order to create the tube like structures that will conduct blood to tissues requiring vascularisation. An important family of receptors involved in this process are the Vascular Endothelial Growth Factor (VEGF) receptors and members of this family like the gene FLT1 will be expressed on cells in blood vessels as they are growing.

What is the function of the product of a gene like FLT1? The Gene Ontology annotates FLT1 to GO:0019838, 'growth factor binding' in the Molecular Function Ontology. As per the term definition, FLT1 gene products do indeed interact selectively and non-covalently with growth factors that stimulate a cell to grow or proliferate.

Yet FLT1 is also annotated to a number of other molecular functions including ATP binding, nucleotide binding, protein binding, receptor activity, transmembrane receptor protein tyrosine kinase activity and vascular endothelial growth factor-activated receptor activity. In addition, there are more than twenty Biological Process Ontology terms to which FLT1 may be associated including cell migration, female pregnancy and the regulation of smooth muscle contraction. An FLT1 protein embedded in a membrane in a cell is not performing all these functions and participating in all these processes at the same time. What counts as a function for FLT1 and what does not? Are there criteria for determining the identity of a function in a biological context?

Despite the fact that most of the Gene Ontology terms describe biological functions in one way or another, the GO Consortium has never tackled the long-standing problem in the philosophy of biology regarding what constitutes a function. I will therefore provide a short review of functions and functional explanations in the biological theory-talk and use this as a basis for explaining how the Gene Ontology handles biological functions.

I will pre-empt this section now by stating that the Gene Ontology offers no satisfactory definition for biological functions. Nor does it offer any robust procedures for identifying, naming or eliminating functions from the ontology, the GO approach being an *ad hoc* synthesis of several different potential treatments for the difficult problem of stating what the properties of biological functions actually are. My argument is that there are no criteria for objectively determining the identity of biological functions, but this is not necessarily a problem for creating good classifications of functions that will support e-science-led knowledge discovery.

However the failure of the Gene Ontology to adequately define biological functions severely undermines the mission of the GO Consortium of creating a unity to solving of theoretical problems in biology, and provides a strong argument against the philosophical doctrine of ontological realism. The implications of the GO approach to functions for existing theories of biological functions will be considered at the conclusion to this chapter.

#### *4.1.1.2 Theories of biological functions*

A number of solutions to the philosophical problem of functions in biology exist, solutions to what functions are, how one can resist teleological arguments in defining functions, or the role of functions in explaining biology. In the next section I will be looking at five different kinds of approaches to functions in biology:

1. Functions as parts of teleological or 'goal-orientated' explanations in biological theories
2. Functions as 'dispositions' in biological systems
3. A definition for functions grounded in the analysis of concepts in the biosciences domain
4. Millikan's theory of 'proper functions'
5. Functions as selected effects, selected for by evolution

#### 4.1.1.2.1 Teleology and Nature's plans

Teleological explanations are an important part of evolutionary biology, and are used to explain why traits have evolved [204]. Ayala describes purpose-driven biological explanations being deployed in two broad senses:

a) To explain the presence of a trait in terms of the end-goal it serves, in terms of its functional role within a goal-directed system and,
b) To explain a trait in terms of the extent to which the trait has contributed to the past reproductive success of the species

Therefore the function of a fish's gills can be explained by:

a) Reference to the capacity of the fish to absorb oxygen from water, in order to support respiratory processes or,
b) By explaining how gills play a role in the reproductive fitness of fish in the evolutionary past

The former kind of explanation is largely mechanistic, whilst the latter incorporates ideas about adaptation by natural selection.

Functions in the biological sense are usually associated with an aim, rather than being merely accidental consequences of a trait. So the function of the heart is identified by the aim of the heart which is to pump blood. The aim of the heart is not the production of thumping sounds in chests, therefore the heartbeat is not a function [270]. Larry Wright also makes an important distinction between functions which have been consciously designed ("Door knobs for opening doors") versus naturally-occurring functions, created by the action of natural selection ("Kidneys for removing wastes products from the bloodstream"). The example of the door-bell above lists functions for a consciously designed object or 'artifact', whereas to ascribe 'growth factor binding' to the gene products of FLT1 is to identify a naturally-occurring function. In biology, long-standing debates regarding the origins of life have conflated artifacts with natural functions. This criticism of the Gene Ontology's handling of functions is dealing with the identification of natural functions, rather than the role of artifacts in functional explanations.

In addition to functions in biology usually being associated with some sort of aim or goal, they are also often *etiological*.

To say that functional explanations in biology are etiological is to say that functions can be explained with reference to their origins, to the antecedent events in time which led to functions existing in the present. Evolutionary theory in biology makes this etiological aspect to functional explanations almost a prerequisite: one cannot separate the evolutionary origins of traits from explanations regarding why a function exists [270]. Yet in practical terms, it is exceedingly difficult to examine the fossil record and explain the functional reason why specific characteristics appear in organisms in the past: not all wing-like structures are used for flight. The inference of function or, in other words,

adaptive usefulness from structures discovered in the fossil record has long been problematic in palaeontology. As Rudwick writes, there is "...no positive criterion by which non-adaptedness can be recognised" [271]. The best an observer of the fossil record can do is to try and offer reasons by which a structure may have supported a conceivable 'operational principle', though such a reason cannot be falsified.

For the molecular biologist, this task of inferring functions has been somewhat simplified by sequence analysis. Similar genetic sequences are assumed to encode similar protein structures which themselves share common functions. Yet the molecular biologist is acutely aware that this is but one small step in the long-process of identifying the true functions of proteins in complex biological systems. Similar protein can have entirely different functions in the complex milieu of living organisms. The long, expensive and complicated process of bringing a new pharmaceutical product to the marketplace is emblematic of just how difficult it is to say what a chemical structure will do against the varied genetic and environmental backgrounds of different individual persons.

The challenge for functional explanations in biology is to identify valid functions in theoretical arguments which separate the etiological from the mechanistic. It makes better sense in a logical argument for explaining why humans have feet to refer to the structural and locomotory properties of feet rather than saying the existence of the ground ultimately required the evolution of something to step on it with. Wright teases the etiological from the mechanistic as follows: "The function of *X* is *Z* means (a) *X* is there because it does *Z*, (b) *Z* is a consequence (or result) of *X*'s being there". Wright's reasoning would correctly explain the presence of haemoglobin to the oxygen-binding capacity of the protein, rather than getting caught in the odd situation whereby haemoglobin exists because there is oxygen in blood.

Teleology is bound with etiology in functional explanations for biology. The aim or purpose of a gene product is normally described as the functional end to which the genome has changed and evolved over time to achieve. So the reason why I produce the hormone insulin, the goal to which the production of insulin in humans is aimed at, is to regulate the storage of fat and carbohydrate in my body. Insulin possesses this function because over successive generations, evolution has selected for this useful function encoded in the genome.

However Brandon reminds us that evolution by natural selection is a theoretical assumption in most philosophical discussions about functions in biology [206]. If the theory were proven wrong, it would severely undermine common teleological thinking in biology. Brandon cites Mayr's distinction between functional biology, which seeks the proximate cause or mechanism for a feature, and evolutionary biology, the aim of which is the ultimate cause or explanation for why an organism or trait exists [206]. This distinction is much like Ayala's above. It is one thing to assert a functional explanation for the mechanism of action of insulin, but quite another to claim its function explains *why* insulin exists.

#### 4.1.1.2.2 Cummins functions and 'dispositions'

Cummins points to two assumptions in functional explanations which have created special problems for biology.

Firstly, it is assumed the purpose of functional explanations is to give reasons why an organ or process exists. The answer to question 'Why do we produce insulin?' is a reason why there is a biochemical structure we dub 'insulin' instantiated in reality.

Secondly, if anything is to have a function, it must contribute to the performance of some 'containing system' [272]. If I have a child with webbed feet, this webbing can be considered as functionless – the mutation makes no apparent contribution to the containing system which is my child, based on what a biologist considers to be a normal, working human.

Cummins' objection to the first assumption is that it is difficult to see how statements invoking functional explanations can act as explananda for the item, process, mechanism or whatever is sought to be explained. Functional explanations for heartbeats which take for granted the first assumption really do not prove why an organism has a heartbeat. Cummins proposes a much healthier alternative which is to drop any assumptions that functional explanations are required to explain the presence of traits. Instead, one ought to consider the role say, a heart, plays in the system in which it is contained – a circulatory system.

Cummins' objection to the second assumption above hinges on how biologists, intent on offering a functional explanation for a trait, can identify those effects of a function which contribute to the what Hempel calls the 'proper working order' of a system. In evolutionary biology, it is well understood that traits can be detrimental to the survival of a species but still have a function begging to be explained in mechanistic terms. The example Cummins gives is that even if wings were to become detrimental to the survival fitness of a species, the structure of that wing in terms of its skeleton, aerodynamics and so forth would still be crying out for a functional explanation. To assume a performance criterion in functional explanations would be to ignore all those normal, mechanistic functional explanations biologists regularly rely upon. Such a performance criterion is inevitably relative to a preconceived image of a normal, working system. Just what constitutes 'proper working order' is what Cummins takes issue with.

Cummins favours functions as dispositions: "...to attribute a function to something is to, in part, attribute a disposition to it" [272]. Dispositions are regularities therefore, which conditions are necessary to precipitate said disposition? We are no longer trying to explain the presence of functions, nor relying on preconceived notions of 'proper working order' for systems. If functions are conferring dispositions on a system, then how are these dispositions brought about?

Functional analysis in biology is therefore the analysis of system capacities. Anything we care to identify as a system will have components with dispositions, much like the parts of an electronic system with identifiable properties. Functions take meaning from the context of the system we choose to explain, and by a reductionist strategy we can decompose complex systems into combinations of simpler systems and the capacities of these more elementary units. Capacities are sensible only against the analytical background of the system. The function of the heart is not to create thumping sounds when we are trying to explain the functional capacities of the circulatory system. The disposition to move blood around the circulatory system is not brought about by thumping sounds. Cummins' strategy permits the biologist to exclude trivial or accidental 'functions' and to focus on an analytical approach which offers true explanatory power for any system being articulated, the penalty being that the analyzed system is only ever a contrivance.

#### 4.1.1.2.3 Conceptual analysis of functions

Biologists use the words 'function' and 'goal' interchangeably: Nagel describes how the paradigm of the genetic code as a program for the goals of an organism have contributed to, or not, to the identification of functions [207].

Nagel discounts the likelihood that biologists could discover "...the specific physicochemical structures that corresponds to a given process". Goal-directed processes cannot be identified by consulting the genetic code, yet Nagel does not fully qualify how such a correspondence is precluded. Despite writing in 1979, before the recent advances we have seen in genomics the point Nagel makes is that there are no necessary conditions by which a process can be said to be 'goal-directed', and hence the biologist cannot point to DNA and state that there, inherent in the code, are those requirements to state what a function is. Coded genetic programs are not predictive of goals since they encode natural functions for biomolecules, unlike a computer program, designed as an artifact to achieve some purpose.

Nagel gives an example of the maintenance of blood volume as a homeostatic process in organisms as compared to examples where equilibrium states are achieved in non-living examples, such as a ball dropped in a bowl coming to rest at the bottom. How can we justify the privileged status of system properties in the first instance, such as the role of hormones in maintaining the blood volume, against applying similar 'goal-directed' explanations to the structure of our bowl? He agrees that to understand the 'goal-directedness' of a system, one must admit a certain relativity, but this is only ever proportionate to our assumptions about what constitutes a goal being empirically testable against observational data. This is an important conclusion: are there models for biological functions that completely resist all charges of relativity? Here Nagel is suggesting that it is reality that can be the arbiter which reduces the multiplicity of potential, trivial functions down to a handful of 'proper' functions.

Bigelow and Pargetter expand the debate on functional explanations to incorporate the sense that "...in describing a present structure in terms of its [biological] function, we mention a future outcome of some sort" [273]. Functions are potentials, but possessing a function is not predictive of the exercise of that function. Prior causes for a structure are a given but existing functions, as commonly understood in the sciences, tell us nothing. In Bigelow and Pargetter's words "...reference to future events is explanatorily redundant".

A bee may possess a stinger, and this stinger structure has a function. Yet if it never fulfils this function, what explanatory power did ascribing a function to the stinger serve in the first place? Bigelow seeks to defend functions against three criticisms that:

a. Eliminativism is better (functional talk will eventually be dropped as biologists get better at explaining what they mean),
b. Representational theory is sufficient (functions are based on prior representations of what will achieve a goal, representations perhaps conceived by God) or,
c. Etiological theory (functions are selected effects, and so history tells us everything we need to know about the use of functions in scientific explanations)

Bigelow intends to offer a more 'forward-looking' defence of functional explanations than etiological alternatives offer, based largely on Cummins' proposal to think of functions as 'dispositions'. The

advantage of this stance is that future effects can be hypothesised based on considering new biological structures or processes in the context of a particular habitat and situation. Arguably, heart sounds do perform a function in the context of humans living in societies with sophisticated medical technologies. If a disposition makes a productive contribution to species survival, then it has a function. As Cummins' approach implies, the context or environment in which said disposition is found inevitably plays an essential role in determining an outcome like 'species survival', hence his insistence that functions cannot be discussed without defining the system in which they operate.

Related to eliminativism is the idea that functions only serve the biology discourse as 'interests' relative to speakers, rather than inhering in reality as properties of structures. Values are considered in the light of functional explanations by Bedau, who considers the range of value-centric thinking when biologists argue that a function fulfils a purpose for an organism, or that doing a particular function 'is good for' an organism [205].

#### 4.1.1.2.4 Millikan's proper functions

Millikan devises her own response to functional explanation by suggesting what she terms 'proper functions' [274]. Proper functions are historical: they look to the history of a structure to denote an appropriate function.

Millikan defines proper functions by appeal to two conditions. The first is that an item *A* should be a reproduction of other item(s) which, in the past, performed some function. The second condition is that item *A* is the product of a copy in the past which, under given circumstances, performed a function by means of producing *A*. Items falling under the first condition have proper functions, and Millikan considers simple biological examples to be organs and instinctive behaviours. Items falling under the latter condition are termed 'derived proper functions' and encompass items produced by malfunctioning systems.

Millikan stresses that her theory, unlike Wright [270], Nagel [275], or Bigelow and Pargetter [273], is not conceptual analysis of what biologists understand by 'function'. She argues that conceptual analyses, such as arguing that etiological theories of function are impossible if biologists' understood what was meant by a function *before* the theory of natural selection was ever conceived, are misleading. Her theory of proper functions is an attempt at what she terms a 'theoretical definition' quite distinctive from definitions for 'function' created to meet the purposes of functional explanations in biology (a 'stipulative definition'), or a definition constructed by the shared usage of the word 'function' in the biosciences domain (a 'descriptive definition').

Millikan devises a thought experiment in which an 'accidental double', a mysterious doppelganger which simply comes into existence, does not possess proper functions because it has no history. Dispositions and capacities in Cummins' sense may be properties of an item having a function but they do not constitute the function itself –items with proper functions have derived properties we understand as purposive. Millikan is trying to tie functions to our common notions of 'purpose'. When Cummins uses functional analysis to decompose an arbitrary system into capacities and functions, this approach does not handle the concept of entities being 'for something'.

Proper functions and derived proper functions attempt to reconcile the idea of function with items which may be defective, maladapted or never fulfil a potential function. A broken can-opener still has the purpose of opening cans; this function still inheres in its structure and our understanding of

the structure. If families of items can more or less have functions, if function is a fuzzy category, this undermines any attempt to explain the presence of functions in the here and now. Proper functions are therefore historical because they refer to an item's history in order to determine those dispositions which have led to the performance of recognisable functions; such a definition supports the kinds of teleological explanation predominant in biology.

Millikan does not believe functions can be understood with reference to 'normal conditions', since these are only relative to each other - unless one accepts that it is necessary to accommodate the actual history of an item. The alternative is to remain committed to the idea that functions are just one of many human ways of thinking incorporated in an epistemic consciousness, rather than meanings created by the interaction between the mind and reality. Millikan asserts that her theory of proper functions is much more favourable, since it relies on that which has passed before, on the reality of an item's history.

#### 4.1.1.2.5 Functions as selected effects or causal roles

Neander reject's Millikan's attack on conceptual analysis as a tool to understand functions [276]. She is committed to an etiological basis for 'proper functions' in biology, and argues that the normal understanding of what a structure is supposed to do, at least in the mind of the biologist, is grounded in the evolutionary history of the organism. Functions are therefore selected effects, features which make an active contribution to the survival of a species. Traits are selected effects, and since traits have functions which they are supposed to perform in the life history of a species, these goals and purposes are what create the categories of functions in biological explanations.

Neander's theory of proper functions, which borrows Millikan's language, ascribes functions to more limited situations that Cummins' style analyses. The cancerous cell therefore has no function. Dysfunctional structures are understood in biological explanations only through the 'norm' which is an etiological account of function: the norm which most biologists share. Dispositions or propensities, as outlined by Bigelow and Pargetter [273], are inadequate as they similarly fail to account for notions of, for example, a dysfunctional kidney. As Neander, for an item to be dysfunctional, it must have had a function in the first place, and this concept of 'function' is founded in a powerful norm in biology, the norm of functions as selected effects.

Amundson and Lauder remark that, given the popularity of etiological accounts of function like those variations proposed by Millikan and Neander, it is "...surprisingly difficult to find an unequivocal rejection of Cummins' alternative" [277]. They therefore defend Cummins' causal role account of functions against what has become largely the consensus view of functions in biology: the authors dub this consensus the *selected effect* accounts of function, whereby functions are those effects which have been selected for by natural selection.

Amundson and Lauder highlight the power of Cummins' approach to be that neither specific goals or purposes, nor evolutionary history need be invoked to ascribe functions to traits for scientists to "...choose capacities which they feel are worthy of functional analysis".

The main objection raised by critics of this method is that any causal properties we care to choose can be incorporated into a functional analysis, generating seemingly useless analyses. However, the authors stress that Cummins did set out limits on those properties that might generate greater epistemic content in an analysis, such as those revealing greater complexity than the analyzing

capacities. In many ways, these limits are much like those used in Popper's falsificationism in which we choose better theories by virtue of simplicity coupled with explanatory scope.

Causal role functions share a strong correspondence with how comparative anatomists actually understand traits and explain their functions. Amundson and Lauder use this fact to argue that selected effect accounts are of only marginal usefulness in this domain of biology compared to causal role interpretations of anatomical features in terms of their physical structure and how this structure contributes to Cummins'-style capacities.

Millikan and Neander are accused of treating causal role functions as concept analysis of natural language, rather than accounts for functions in biology which are grounded in the theoretical definitions (which selected effects supposedly are, given that they define functions with reference to the theory of natural selection). If treated as a theoretical definition of function, causal role accounts are immune to this criticism; they just use different biological theories to articulate functional explanations.

Can biological trait categories only be defined by selected effects? By this argument, causal roles cannot identify normal hearts and distinguish these from malformed organs because only selected effect definitions can correctly categorise hearts. Amundson and Lauder rubbish this claim, and argue that all manner of molecular, morphological, physical, and spatial data can be used to infer functions for traits without any reference to etiology and selected effects. Biologists can identify severely malformed, dysfunctional hearts by patterns of muscle striations, molecular markers and various morphological criteria which unambiguously categorise said organ as some kind of heart, without any recourse to proper functions and evolutionary history.

The authors also make an important distinction between homology and analogy. Homologous organs may share a phylogeny (history) whereas analogous organs share no such evolutionary past. When biologists put two analogous structures into the same functional group, they do so without any reliance on selected effect accounts for function, and instead are using causal role criteria to understand functions. This is akin to the Gene Ontology approach to classifying gene products, where analogous structures may share a function, regardless of phylogeny.

Causal roles also satisfy an important theoretical requirement in biology, in that they can be used in hypothetical models, or situations in which it is not possible to identify the trait a selected effect acts upon. Selected effect explanations are exceedingly weak when a biologist can point to an effect with relative ease – increases stamina for example – but cannot tease out, from the multiplicity of traits and features with potential involvement, just what was acted upon by natural selection to create this effect. The Cummins' approach lets the biologist analyse an effect in terms of whichever system properties and their capacities he believes to be involved. Evolutionary history is often a black box with pleiotropy (one gene change leading to many phenotypic effects) a norm: causal role functions let the biologist work with what can be see and measured, rather than relying on potential events in an evolutionary past it may now be impossible to reveal

Enc and Adams argue that the philosophical difference between etiological accounts of function and accounts of function which look to the future (the 'propensities' discusses by Bigelow and Pargetter), are not nearly as significant as first believed [278]. Enc and Adams seek a solution to the problem common to all philosophical discussions of functional analysis: understanding the role played by

function attributions in causal explanations. The 'goal directedness' they see as the solution is "...a broad taxonomy of properties (or behaviours) classified together by reference to a distal goal, and in the fact that such taxonomy makes it possible to generate non-causal explanations and predictive hypotheses about *types* of behaviours." [278]

Is then Gene Ontology one such 'broad taxonomy of properties' for explaining the molecular biology of the cell?

### 4.1.2 GO philosophy on functions in biology

The official documentation on the Gene Ontology website makes a single attempt at defining a function, with the sentence below, quoted from the Molecular Function Ontology guidelines:

> "A function is the potential to perform an activity, whereas an activity is the realization, the occurrence of that function; so in fact, 'molecular function' might more properly be renamed 'molecular activity'." [279]

The Gene Ontology therefore attempts to distinguish between functions for gene products, which are potentials, and activities, which are realised potentials. This distinction is drawn from Basic Formal Ontology (BFO) where an activity is an occurrent rather than a continuant. Gene products constitute constant entities in reality, the equivalent of things or objects. The functions of that gene product, or the processes it engages in, are the occurrents. Occurrents are limited in duration to a specific space and time. They happen and finish: the definition above alludes to this principle.

However all occurrents are established on a function which is best understood as a potential to perform an activity. For a gene product to be annotated to a GO term, it must be demonstrably true that the appropriate activity does occur at some point in space and time in a normal cell. It would be inconsistent with GO philosophy to create annotations between functions and gene products that never had, or potentially never would, realise these functions. But this stipulation that biologists ought to consider all the functions to be activities in the Gene Ontology, this still leaves one with the problem outlined at some length above. What are the criteria for objectively identifying molecular functions, that they may be described in and exploited by, ontologies in biology?

In the light of the different solutions to the problems presented by functional talk in biology and the brief definition for molecular functions above, I will attempt to discern which philosophy best matches the Gene Ontology approach to functions.

#### *4.1.2.1 Is the Gene Ontology classification of functions overtly teleological?*

Do the Gene Ontology treatments of molecular functions, and of biological processes as compounds of more than one function, correspond with teleological or 'goal-orientated' explanations for the meaning of functions in biological theories?

The Gene Ontology does subscribe to notion of 'goal-directedness'. High-level functions in the Molecular Function Ontology represent broad biochemical 'goals' in the context of biological systems , such as the catalysis of particular reactions ('ligase activity'), or the regulation of specific signalling pathways ('steroid hormone receptor activity').

Annotations therefore ascribe purposes to gene products based on these broad biochemical goals. The purpose or aim of gene product ABC in a species is to perform the function of XYZ. The GO

approach is mechanistic in its assumption of a teleological meaning for biological functions. The ascription of one or more purposes to gene products via annotations is an effort to explain the presence of that gene product in the context of specific biochemical aims, such as the joining together of two molecules with a new chemical bond by the action of a ligase enzyme, or the stimulation of cells in a tissue by the steroid hormones binding to receptors. The way researchers use the Gene Ontology in GO term enrichment analyses also corresponds with a teleological view of functions. Given the over-expression of a gene-set, can the biologist say to what final purpose or, more simply, *why* these particular genes are over-expressed in a system? This mechanistic deployment of teleological bases to functions is not an attempt at validating functions as adaptations over evolutionary time. Researchers using GO rarely use language describe functions as traits which have evolved according to a plan, as adaptations to identifiable problems in an evolutionary niche. Rather GO research is functionalist in philosophy, concerning itself with the proximate causes of biological processes – why are steroid hormones acting in this biological context? - rather than grappling with the bigger questions of why particular functions exist in the first place.

The Gene Ontology's dependency on electronically inferred annotations based on sequence analysis mirrors Rudwick's problem of accommodating a teleological perspective when interpreting the fossil record. When presented with a particular form, the paleobiologist must make assumptions, based on present knowledge, as to the purpose of these forms in the past. The wing-like structure is assumed to be for the purpose of flight, rather than possessing the function of attracting a mate, or acting as flippers in water. In the same manner, the Gene Ontology infers functions electronically from sequence data, assuming similar forms share similar functions. An underlying coherency to the purpose of particular sequences at the molecular-level exists across all species. Although the adoption of these assumptions has permitted the rapid expansion of GO annotations by algorithmic means, the GO Consortium, in placing such a great importance on manually validated annotations, is conscious that these data are only tentative. The very nature of life on earth has always been the capacity for organisms to adapt and change, and in the same way that a hand can function to grasp, or support, or communicate, or fight, so too can a single biomolecular structure potentially function, or not function, in a myriad of different manners.

Where does the establishment of goals for biological system create problems for the Gene Ontology in classifying different molecular functions? The main challenge to this approach is a by-product of ontological realism. Teleological thinking means that the purpose of a gene product is not necessarily realised. A ligase enzyme may not be joining two molecules together at a point in time for it still to be classified as possessing the function of a ligase. A realist perspective can struggle to deal with the sense that functions and processes as goals in biological systems are 'potentials', for realism is supposed to be about what happens in reality, not what *might* happen. Hence the GO definition for functions above re-phrases functions as the realizations of activities, and grounds the Gene Ontology in the representation of occurrences of functions in reality. Hypothetical functions are abrogated in favour of realized goals. GO assumes the material facts of final causes in biological systems to be true.

#### *4.1.2.2 Is the Gene Ontology an example of a Cummins -style 'containing system'?*

Cummins developed his perspective on the role of functional explanations partly in response to a tendency in biology to use functions as reasons for why traits or processes exist. The example given previously was the presence of a heart-beat explained by the need to pump blood around a system. As a functional explanation, this is problematic because one cannot prove that the beating heart exists in order to fulfil its function of pumping blood. Alternative traits with the same function, or features of organisms which have evolved over time to arguably manifest a series of functions, make it difficult to prove the existence of X by an appeal to function Y.

GO annotations are closely allied with this assumption. An annotation is a statement proposing the reason *why* a gene product is expressed. According to this type of explanation, the reason why FLT1 is expressed in developing blood vessels is to regulate the growth of blood vessels. This is the functional explanation for FLT1 but it does not prove, as per Cummins' reasoning that the reason why FLT1 exists is in order to perform this function. Yet the interpretation of the meaning of GO annotations in a GO term enrichment analysis does make this assumption: increased expression of FLT1 is interpreted as an increased likelihood that blood vessel growth is an active process.

Cummins also objects to functions defined in relation to the 'proper working order' of a system. In our example of the FLT1 gene product, the function of this protein can be explained by it normally binding growth factors and regulating vasculogenesis. A functional explanation for FLT1 in this sense is that the gene product contributes to the 'proper working order' of my body by helping to regulate the growth of blood vessels which themselves exist to distribute food and oxygen to my tissues. How do we define the 'proper working order' of a system though? Cancers are not considered to be part of the proper working order of the human species, yet over-expression of FLT1 might well contribute to the spread of a melanoma in the skin on my arm [280]. Here, FLT1 functions sense to both propagate cancerous cells and ruin my health: to form a single, coherent functional explanation that accommodates both contexts is difficult, if not impossible, using a 'proper working order' argument.

The Gene Ontology has indirectly grappled with the problem of defining a 'proper working system'. Much as it is difficult to explain the function of gene products in my healthy tissues *and* explain their functions in diseased tissues, so these contextual problems are found in describing host-symbiont interactions, or the relationship between a host and a virion. There are now hundreds of GO terms which describe cell parts and processes as belonging to a 'host', so that there is a 'cell nucleus' and a 'host cell nucleus' to enable annotations for symbiont gene product locations. What is a cell nucleus under a 'proper working system' for one species is exterior to a symbiont and represents a 'host' location.

My argument here is that if the Gene Ontology admits that there is a problem for defining host-symbiont interactions, or in describing processes from the perspective of a virion particle and the host cell it has infected, then the entire GO system admits that there are potentially any number of ways of defining 'proper working order' depending on context. Any functional explanations for pathological conditions could be made according to the healthy individual or from the contradictory context of the functional success of a diseased system.

Cummins realised this was a major problem for the logic of functional explanations in biological theories, and advocated his solution which was to define a containing system. We analyse functions

according to a specific capacity of this containing system C, and this capacity can be whatever we choose it to be. Any system, biological or otherwise, is rendered as something very close to an engineering problem. The advantage of selecting a capacity in C is that we avoid the sense of a purpose or goal in our more teleological explanations, and we need not rely upon an over-arching notion of the 'proper working' of a healthy organism.

**Figure 13: Cummins-style decomposition of biological process into a containing system**

**X is a means to Y *in relation to the capacity of containing system C***

1. FLT1 gene product is a means to regulate vasculogenesis
2. FLT1 gene product is a means to regulate vasculogenesis *in relation to the capacity to vascularise new muscles of a regularly exercising adult*
3. FLT1 gene product is a means to regulate vasculogenesis *in relation to the capacity to invade healthy tissues of a melanoma*

The example in Figure 13 above explains the Cummins' approach. '1' is weak because in order to explain the function of FLT1, we must embark on an odd regress of functions such as FLT1 regulates vasculogenesis, and the function of vasculogenesis is to grow new blood vessels, and blood vessels support the life of an organism, and the goal of an organism is to be alive. '2' is much better because we define a containing system, state a capacity and explain the purpose of FLT1 by how it enhances this capacity. '3' demonstrates that by a Cummins' analysis, we can now distinguish different capacities in different containing systems, and articulate different explanations for the purpose of FLT1 by how it contributes to these different capacities.

#### *4.1.2.3 The identification of molecular functions via concept analysis*

The suggestion that the analysis of concepts in the biosciences domain might offer a basis for constructing ontologies would, on the surface, be anathema to the Gene Ontology project. This particular model for functions is at the furthest end of the spectrum from ontological realism, grounded entirely in the discourse of biology and the subjective minds of biologists creating concepts in theory-talk.

However, concept analysis is actually quite close to how the GO Consortium has created the ontology, via appeals to domain experts, negotiation of shared definitions, by the creation of new terms as they are needed in order to capture newly identified concepts, or the creation of terms to represent new ways of thinking about biological processes.

Dialogues I have investigated such as the function of gene families like caspases, which traditionally have been known only by their family names and not necessarily by the functions they perform in biological contexts, suggest that the GO developers do analyse concepts in the defining new GO terminology. As highlighted by the discussion of Nagel above, the idea of a genetic code does imply a measure of goal-directed thinking in the way biologists think about biological functions. The whole GO project is predicated on the assumption that there is a fundamental unity in the way genetic programmes in different species share commonalities. The assumption that there are common molecular functions across different forms of life is a basic precept of the Gene Ontology Consortium and across the biosciences today, and concepts for functions are grounded in this assumption.

Bigelow's suggestion that functions in biology follow an eliminativist tract, whereby biologist get better at talking about what they mean by functions, rings true for the Gene Ontology. The Ontology itself is seen to improvable, with editing and re-defining a necessary part of the house-keeping involved in keeping the ontology up-to-date with best current domain knowledge. Yet unlike Bedau, the Gene Ontology cannot handle the notion of values in understanding which concepts in the ontology are especially pertinent to solving specific biological problems. And nor can ontological realism as a basis for ontology construction accommodate Bigelow's suggestion of hypothetical functions, validated through experiment and the scientific discourse. The suggestion that there are entirely hypothetical functions for bio-molecules which, given contexts or a lack of evidence, may not exist yet still serve a valid purpose in testing functional explanations – and thereby deserve representation in an ontology – runs counter to Basic Formal Ontology.

I believe criteria for biological functions based on concept analysis in the domain could make very useful classifications systems. They would certainly be a closer representation of how different groups of biologists think about theories, and could incorporate the values of these sub-domains into controlled vocabularies. Most importantly, elements of conceptual analysis permit the representation of hypothetical entities in ontologies, this being an important way to build new functional terms into ontologies and bioinformatic analysis tools, a method entirely ignored by the Gene Ontology.

#### *4.1.2.4 GO functions as proper functions*

Millikan's theory of 'proper functions' is difficult to resolve against a constantly changing evolutionary background, and presents special problems as a means for objectively identifying functions in molecular biology. The ethos of Millikan's approach is closely allied with ontological realism. Functions for biological entities can be objectively identified as true, or proper, functions. They are instantiated in reality and are given authority by their etiology. In the case of molecular biology, sequences evolve over time to encode proteins with functions in the cell, and a proper function is the goal this sequence has evolved over time to achieve.

The proper function of FLT1 is therefore to bind growth factors and, through this binding, transmit signals to vasculogenesis pathways in the cell. The reason why FLT1 is expressed in cells is to perform this function, because if we look back through this history of a species, the growth of blood vessels is an important function necessary for the survival of that species. Yet Millikan's etiological account and her 'proper functions' ignores both hypothetical functions, which are useful to the work of the biologist testing theories, and does not deal with what I term 'potential functions', which are unrealised functions for molecules in the cell.

A thought experiment can illustrate this problem of 'potential functions' and the way in which Millikan's argument for functions cannot handle their consequences. Using genetic engineering it is relatively straightforward to introduce any gene sequences into a host genome, and ensure that this is reproduced in subsequent cell divisions. One might select a gene encoding a receptor for a molecule like dopamine which, given the host environment, would never be active.

If I insert a gene for a dopamine receptor into yeast cells, the function would never be realised because yeast do not express dopamine. The host cells now carry a new potential function, encoded in the gene sequence which has been engineered into their genome. At no point in the past and, potentially, at no point in the future will these cells perform the activity associated with the gene

product. The GO Molecular Function term ‘dopamine receptor activity’ (GO:0004952) is defined as “Combining with the neurotransmitter dopamine to initiate a change in cell activity”. The occurrent of that function may never occur in my mutant yeast cells, yet they express dopamine receptors on their surface and, with further manipulation, could be made to respond in some way to dopamine.

Do the mutant yeast carry the ‘dopamine receptor activity’ function? Should I now annotate GO:0004952 to a new database entry in my yeast database? The relationship has no reference to reality, yet commonsense suggests that the function is still there, latent and ready to become active, given the right conditions. The ‘proper function’ of this receptor in yeast cells is not to act as a dopamine receptor, because there is not etiological justification for this function. The potential function was added in the previous generation, and has never served any useful purpose in the life-cycle of the yeast species. Yet we arguably must assign some function to this dopamine receptor, given certain conditions, or according to how we define, in Cummins’ style, the containing system and attendant capacities.

In the same way the Basic Formal Ontology talks of ‘activities’, and the requirement for occurrents in reality to avoid the difficult problem of potential, unrealised biological functions, so too does Millikan’s approach and her theory of ‘proper functions’ as historically-grounded, ‘real’ purposes for traits in biology.

#### *4.1.2.5 Does the Gene Ontology treat functions as selected effects or causal roles?*

Neander’s approach to the role of functions in explanations in biology is largely ignored by the GO Consortium. The evolutionary history of functions or any effort to group functions based on their history is not included in the relationships encodes by the Gene Ontology. Early in the design of the vocabulary, a senior developer explicitly declared that the ontology was not intended to represent phylogeny, or evolutionary relatedness, something quite possible using sequence data to identify ancestral patterns common between species. Instead, the Gene Ontology has always been seen as entirely agnostic with respect to evolutionary origins or the idea that the same function in different species might be related by a history of selected effects.

The refusal of the Gene Ontology to admit any species information into the vocabulary, as demonstrated by the failure of so-called ‘sensu’ terms, successfully obliterates any phylogeny of functions. Functions are represented in the ontology files devoid of evolutionary context even though Neander’s idea of a function being distinguished as any effect selected for by evolution, is an idea grounded in one of the most important theories of the biosciences domain.

If the Gene Ontology therefore avoids the possibility that a function is given force primarily by the fact it is an effect selected for by evolution, does it accept Amundson and Lauder’s defence against Neander with their functions as causal roles?

Without going so far as to incorporate causal role effects into a taxonomy of function for molecular biology, the Gene Ontology actually ends up creating an inadvertent phylogeny of function in direct contradistinction to its original mission. The categorisation of structural molecules by the super-structure they participate in for example consequently groups gene products according to the evolutionary history of a species. The presence of eyes, muscles or bones for example imply a taxonomy for functions related to membership of the kingdom Animalia, which is nothing like the

causal role approach advocated by Amundson and Lauder which is designed to entirely ignore, and for good reason, phylogeny in categorising functions.

Similarly the reliance on ligands-bound in order to create a taxonomy of functions for receptor activities in the Gene Ontology fails to merge phylogenetic thinking into the resultant classification. The classification of steroid hormone receptor activities in the GO Molecular Function Ontology only makes sense by accepting that steroid hormones are synthesised in organisms with adrenal glands or gonads. These specialised organs in part define an evolutionary history as selected effects in a special group of species. The opportunity to define functions according to their causal role in theoretical systems, much like a Cummins analysis of functions, severely limits the power of the Gene Ontology as it relates to theory testing. When asked what functional role a gene product annotated to the steroid hormone receptor activity has, any explanation is indelibly tied in Gene Ontology-thinking to pre-existing theories of phylogeny and purpose for families of genes. The opportunity to drop selected effects as the bases for functions in biological theories is lost. We have come full circle instead to the uncomfortable fact of regress, where the presence of traits and their functions cannot be separated from explanations which either rely on 'Because it evolved that way' or 'Because that is its proper purpose'.

### 4.1.3 An argument against 'objective' biological ontologies which cannot define functions

The Gene Ontology, in its failure to adopt a consistent philosophical position for biological functions, creates a contradiction in its design. On the one hand the ontology claims to be a model of knowledge about occurrents in reality, rather than a model of mental concepts and hypothetical explanatory constructs. On the other hand, the Gene Ontology's treatment of functions, as explored in this chapter, falls somewhere between Millikan's notion of proper functions whereby nodes in the GO graph correspond to gene product functions in an idealised cell, and Bigelow and Pargetter's suggestion that functions act as potentials to perform certain cellular roles.

Each of these approaches to biological functions admits a strong element of relativity into their solutions. Millikan's solution demands contextual inflexibility, whereby the question of what gene products *ought* to do in a cell is determined by a consensus view in biology. The aberrant, mutative and pathological states that biologists commonly study are sidelined in favour of the explanatory power gained by conceiving an ideal model cell and its 'proper' function. Bigelow's thinking on functions, which is forward-looking and argues for the importance of potential functions that may be realised in specific environmental contexts, corresponds very closely with GO Consortium aims to create a tool for biologists which aids theoretical predictions. Yet this flexibility with regard to contexts and the requirement to test hypothetical functions after the manner of Cummins' dispositions and containing systems is not reflected in the GO organisational philosophy. Ontological realism does not admit this kind of theoretical reflexivity. By design it is context independent, because this meets a standard of objectivity and naturally complements machine-based reasoning according to the logical formalisms of Basic Formal Ontology.

What we are left with though is an uncomfortable contradiction, where GO by design is logically consistent and deliberately inflexible, but in practical terms and by implementation is crying out for a way to incorporate the plastic, context-dependent ways biologists normally handle biological functions in explanatory theories.

No better illustration of this contradiction can be found than in the somewhat obscure treatment of biological processes in the Gene Ontology. Biological processes are poorly defined in GO guidelines, are massively over-represented across the three sub-ontologies, and are not even recognised in the existing literature on biological functions and explanatory talk in biology.

The Gene Ontology claims that biological processes are a different class of entity in the reality of biological systems because, unlike functions which are atomic, processes involve more than one activity acting in concert to achieve a goal or activity in the cell.

Yet in the Molecular Function Ontology we have nested categories of functions, raising the question of how super-classes relate to their children? If the super-class function term 'binding' in the Molecular Function Ontology has multiple children, the activities of which contribute to the global function in living organisms of 'binding things', how is this distinguishable from a biological process we might call 'binding'?

Early in the design of the Gene Ontology, there was talk that the Molecular Function Ontology would in fact have no relationships or hierarchical structure, and instead would simply be a list. This design option was dropped over time, and has resulted in a situation whereby there are no clear criteria separating processes from functions and arguably any node in the Molecular Function ontology might be decomposed further into sub-functions, and large numbers of biological process term could be re-defined as molecular functions. We have seen this in the results from the term obsoletion analysis, where terms are dropped from the Molecular Function Ontology and get moved to a new node in the Biological Process Ontology.

For example, consider gene products annotated to the Molecular Function Ontology node 'DNA binding'. DNA binding is considered to be an atomic function in GO philosophy. However DNA binding can be decomposed into further sub-functions such as sequence recognition, maintenance of protein shape, stability of surface ionic charge, passage through cellular fluids, ability to be degraded by proteolysis – all of which are molecular-level functions essential to the higher goal of binding DNA. Is DNA binding actually therefore a biological process? My argument is that the Gene Ontology offers no satisfactory criteria for resolving this kind of question. Any criteria it may invoke are certainly not grounded in a framework for biological functions drawing on existing philosophical arguments. If anything, this question is determined by an arbitrary set of rules created by a small number of GO developers. These developers have demonstrably created a successful for tool for indexing gene products, but may have left their classification on very soft ground. Might not classifications for gene products constructed according to different theoretical models for biological functions be more powerful for growing new knowledge in the domain? This possibility is explored further in Chapter 4.2.

A strong teleological undercurrent runs through the design of many sections of the Gene Ontology, and again, this is entirely consistent with several existing approaches to functions in biology, especially Bigelow and Pargetter. The Biological Process Ontology is the representation of biological goals in the life of a cell, and annotations to these terms ascribe purpose to the gene products.

This is not necessarily problematic, but the Gene Ontology has never stated a position of teleological definitions for functions, and if it accepts this philosophical tenet, it must be prepared to defend this position. Large numbers of Biological Process Ontology terms are authored with standardised term

strings which are entirely goal-based, and cannot be constructed otherwise. Thus where a gene product is annotated to a GO term of the form, 'Binding Substance_X', in the absence of 'Substance_X' does said gene product still reserve this function if it can never, in reality, achieve its purpose?

Likewise for GO terms of the form, 'Substance_X receptor activity', should the binding to a receptor of 'Substance-X' fail to always initiate said activity, does a gene product still possess, in latent form, a potential to perform this activity? Like most chemical reactions, biological processes adopt equilibrium kinetics. If I don't have enough acetylcholine to trigger a nerve depolarisation event, my receptors clearly still possess a purpose, a purpose to become activated in the presence of a specific neurotransmitter.

The Gene Ontology remains silent on the explanatory role that functions and activities grounded in teleological talk and unrealised potentials may play in reasoning systems for bioinformatics. This is a purview ontological realism does not confront, and should unrealised functional potentials be an essential part of a classificatory system for gene products, there is a strong argument for extending this to unrealised *hypothetical* functions and activities which currently have no place in the Gene Ontology knowledge schema.

As the Gene Ontology states in its guidelines, a function should be unambiguous and should mean the same thing regardless of which species is being dealt with. The GO developers have grappled with the difficulty of achieving a pure, context-free classification for gene products. In their fleeting deployment of 'sensu' qualifiers for GO terms to clarify term usage in particular species to the incorporation of special terminology to describe host-symbiotic interactions between cells from different species, contextual dependency has crept into the ontology. Even in its current state and as discussed in the Chapter 3.2, there are many areas of the GO graphs where compound terminology indicates how, for example, processes occur in specific anatomical locations (such as cell differentiation in the heart). The GO Consortium does anticipate a point in the future where partner ontologies and cross-vocabulary linkages will amend many of these GO terms where a biological context, such as anatomical location, is written into a term definition. My conclusion in this section though, and grounded in Cummins' realization that there is a need to define containing systems if we are to adequately integrate biological functions into logically consistent explanations for phenomena, is that context cannot be scripted from classifications for gene products.

It may be that although there is a functional analogy between DNA binding in a healthy human cell, a cancerous human cell and a dividing bacterium, it could be necessary to resist the assumption that there is functional equivalence between these different contexts. We are drawn once more to the semantic role theoretical terminology plays in biological explanations. DNA binding might not mean the same thing to different biologists given a range of potential starting conditions, environmental constraints, species differences and over-arching biological goals acting as the object of study.

Rather than aiming at a universal classification for gene products, there exists the possibility, supported by the evidence in this thesis, that there is a place for context-dependent, system-centric vocabularies for describing gene products in biology, catering for research niches in the domain, and specifically designed to support machine-based reasoning on special topics and biological problems.

## 4.2 Alternative classification standards for biology

### 4.2.1 On improving classifications for molecular biology

#### *4.2.1.1 A case for ontological subjectivism?*

The results of the previous chapters support the argument that the Gene Ontology is not presently, nor can never in the future, be predicated on the principles of ontological realism. Subjectivity cannot be eliminated, even from the sciences, despite the efforts scientists may try to develop systems of thought that are established wholly on an objective sensibility. In the way molecular biologists conceptualise theories, in the manner in which the Gene Ontology has been created and changed over time, in the practice of using and reporting GO analyses in the literature: all these results indicate that biologists think in different, flexible, individual and creative ways about biological theories.

The progress of molecular biology in the last fifty years would seem to suggest that despite this feature of the biosciences, it has not impeded the growth of knowledge.

The Gene Ontology was created to solve a problem in molecular biology. Different specialities in the domain used different languages and practices for describing the functions of genes, which made cross-searching between different species databases impossible. The Gene Ontology was designed to solve this problem by creating a single knowledge representation that might be shared across all sub-domains in molecular biology. Yet the evidence in this thesis indicates that the authors of the Gene Ontology have solved the problem by the exercise of their institutional power in bringing together major species databases, imposing standards for a controlled vocabulary, and structuring this vocabulary according to rules that eliminate problematic concepts or ways of thinking about theoretical entities in the domain.

Basic Formal Ontology, the standard around which the Gene Ontology is designed, strongly resists the idea that ontologies represent concepts in the minds of biologists. Ontology terms represent entities in reality. However ontologies, or any other type of controlled vocabulary used for describing resources in molecular biology, perhaps ought to try and describe the subjective way biologists think and talk about theories in the domain. Why? Because this is how biologists think and work on biological problems. Different kinds of vocabularies which do try and represent mental concepts in molecular biology could act as a useful adjunct to ontologies constructed according to ontological realism. They do not need to replace existing ontologies. But if the primary goal of the molecular biology domain is the growth of knowledge, forcing scientists to adopt a single knowledge representation for the whole subject could risk limiting their intellectual freedom, the freedom to think about biological problems in creative, plastic, non-paradigmatic ways.

If we accept that alternative classifications for molecular biology might therefore be useful, what would these classifications, constructed according to what we might 'ontological subjectivism' look like?

#### *4.2.1.2 Pluralism for scientific classifications*

If we are to establish alternative principles for creating scientifically sound classifications that support knowledge discovery in molecular biology, it is important to accept that rather than having

*one* classification that serves the entire domain, it may be necessary to develop *many* classifications to serve different, specific purposes or people in the domain.

The philosophical tenets of ontological realism mean that an ontology aims to represent occurrents and continuants in reality. Each node in an ontology represents a function of process that may occur in the lifecycle of a cell, and each node may be related to other nodes – that is other functions and processes in reality – in a limited number of ways. Paths through an ontology back to the root node describe nested sets of kinds, such that the node 'positive regulation of cell growth' is a kind of 'cell growth', a category which itself is subsumed by the broader process of 'growth' in the Biological Process Ontology. Ontological realism restricts both:

I. What constitutes a valid node in the ontology; the kinds of functions and processes that exist in reality
II. What constitutes a valid relationship in the ontology; the subsumption of narrower categories of entities in reality by broader kinds of functions and processes

Ontological subjectivism would mean that there could be any number of ways of representing knowledge as understood in the minds of molecular biologists. The scientific method would place certain restrictions on what a classification of these theoretical entities might be since:

I. Biologists would have to agree on that the object of study might exist in reality, and was a valid entity to investigate empirically
II. Relationships between different types of functions and processes would be constrained by the logic of theoretical arguments

These restrictions will be described in more detail below, but the philosophical basis for creating scientific classifications in this manner has been explored in some detail previously by other researchers in the information sciences domain under the rubric of pluralism.

#### 4.2.1.2.1 What is pluralism?

The process of creating scientific classifications is bound by the tension between creating unified, stable and objective classifications on the one hand, and reflecting the multiple, fluid viewpoints and inherent subjectivity of human knowledge [281, 282] .

The first perspective rests on the assumption that it is possible to create a single classification of scientific knowledge that will be consistent across different scientific domains. This assumption derives from reductionism, such as the belief that chemical knowledge can be explained in terms of the physics of sub-atomic particles, or that biological processes can be explained in terms of the kinetics of chemical reactions. Scientific classifications are also assumed to be stable in that ultimately, once our knowledge of reality is improved beyond a certain point, classification will mirror reality, and will not need to be changed or altered further. In addition, scientific classifications are objective classifications because they do not incorporate human politics, social thinking or subjectivity; they are not relativistic and do not depend on what individuals believe, only on what reality shows to be true.

The second perspective on scientific classifications is that there is no single, unified view of reality which can be represented in a structured vocabulary. Different social groups believe different things, and a classification should represent these different beliefs. Knowledge is always changing as we

make new discoveries and integrate new ways of thinking into our knowledge of reality. Scientific classifications therefore must always change, stability is a myth and vocabularies must represent the dynamic nature of knowledge. Furthermore all knowledge, even scientific knowledge, is subject to the social behaviours of groups of working people. Even seemingly objective knowledge is bound by the social conventions of those who created that knowledge, even in scientific domains, and so a scientific classification will necessarily integrate those prejudices and politics into the lists of terms and relations of which it is composed.

Our competing theories of scientific classifications therefore fall under the precepts of objectivity on the one hand, and pluralism on the other.

Pluralism is the philosophical viewpoint that there exist multiple valid explanations for observable phenomena. It is not relativism, and pluralism is demonstrably consistent with many of the features of scientific objectivity. The consequence for classification is that rather than creating one classification, one can create many, and all are valid in a scientific sense.

##### 4.2.1.2.2 What advantages does pluralism offer for classifications?

The Gene Ontology is a good example of classification according to the first perspective above, that of objectivity.

The Gene Ontology aspires to unity across the different sub-domains of molecular biology, resolving the knowledge of different biological models in different species with knowledge of physiology, pathology and developmental biology.

Despite incorporating various mechanisms for updating the ontology with the latest biological knowledge, the Gene Ontology ultimately aspires towards stability. Edits, large scale amendments to the ontology, and term obsoletions are conducted with the aim of creating a more stable vocabulary in the long-term.

Ontology relations must meet the condition of the 'True Path Rule'. All relations from any term in a sub-ontology back to the ontology root must be true according to the current state of knowledge in molecular biology. Truths are beliefs validated by the scientific method and all annotations to terms are based on evidence and coded accordingly. The Gene Ontology is therefore objective; if one scientist believes differently, he must convince the scientific community, using empirical data and hypothesis testing, that an alternative perspective is true.

However, there are certain benefits for the domain were the Gene Ontology, or any other ontology in biology, founded on classification according to pluralism (see Table 30 below).

Table 30: Some examples of the potential benefits of pluralistic classifications

| BENEFIT | EXAMPLE |
|---|---|
| **Admission and description of tentative, unproven knowledge** | Providing explanations for controversial theories such as epigenetic mechanisms for Lamarckian inheritance |
| **Better representation of domain-specific knowledge** | Classifications tailored for plant or viral biologists which have previously proven difficult to merge into single, unified classifications |
| **Achievement of epistemic aims** | Scientific work is prioritised according to what scientists believe will be a fruitful direction |
| **Solution to the Gene Ontology problem of annotation inconsistency** | Annotators do not necessarily have to agree on term annotations because there are multiple, equally valid explanations for phenomena |
| **Potential technical benefits** | Better information retrieval, better data management, greater possibility for computer-generated hypotheses |

#### 4.2.1.2.3 How can a scientific classification be pluralistic and objective?

As mentioned before, pluralistic classifications are not necessarily unscientific classifications. I say this because pluralism can be viewed as a re-worded relativism. If reality is not the final arbiter of the categories and classes we believe to exist in the universe, then how does one decide on the kinds of things worthy of inclusion in a scientific classification system? If a scientific classification is not a mirror of the universe, then what does it reflect? In the case of scientific explanations, to take a pluralistic approach is not to devolve to relativism. If an explanation is a statement in which the *explanans* contains a law-like statement, along with accessory boundary claims and conditions, which provides an explanation for the *explanandum* (the facts to be explained), then pluralism states there can be more than one valid explanans for any explanandum.

In biology this view is acceptable as it becomes progressively more difficult to offer universal explanations for natural phenomena at any *level* of explanation [33]. For example, a geneticist offers one explanation for heart disease in the population according to various genetic markers, probabilities of inheritance, and environmental backgrounds. The likelihood that a person gets heart disease is a statistical probability. The physiologist or molecular biologist will offer a quite different explanation for heart disease in the population, based on the interaction of diet and fitness against a biochemical background in the individual. The explanation will be causal-mechanistic rather than statistical. Even though we have two different explanans for the same explanandum, it is uncontroversial to say that even though these explanations are quite different and cannot be resolved into a single explanation, they are not mutually incompatible.

The consequence is that it is plausible that different kinds of explanations, using different kinds of classifications, can offer valid law-like explanations for a range of phenomena in biology. Here is pluralism in action, a pluralism which is scientific to the extent that it can explain and make predictions without collapsing into relativism. There are thinkers who take the failure of reductionist thinking in the sciences, to which pluralism is one potential solution, to a logical extreme in which there is no unity to the sciences, and that all talk of essences, kinds, and explanations for phenomena is a myth [18]. I do not go so far as to accept there is a fundamental disunity in the

sciences. Rather, I believe there are different, equally valid ways of looking at and solving scientific problems. Different solutions may require different ways of classifying scientific entities.

One reason why the Gene Ontology is founded on ontological realism is that it offers a methodology by which a unified, stable and objective knowledge representation can be created. This after all was the objective of the GO project from the outside – to unify the disparate ways in which different sub-domains in biology described gene product functions. To accept pluralism in the form I have described above would be to admit that there is no single way to harmonise functional descriptions for gene products across biology.

### 4.2.2 Alternative ways to classify gene products

I will now describe three different methods to classify gene products in the molecular biology domain which incorporate pluralistic ideals:

- Natural language
- User-defined keywords
- Faceted classifications

#### *4.2.2.1 Natural language*

Molecular biologists use language to represent knowledge about the biology of cells. Biological language, its syntax and semantics, describes scientific theories: what theories there are, how theories relate to one another, how theories are accepted as true. The richness of biological language in molecular biology corresponds to the range and complexity of scientific discourse in that domain, of these different theories and perspectives on the machinations of the cell. The natural language of molecular biology in texts such as scholarly articles therefore represents a potential source of terminology to index gene products with molecular function.

Natural language processing in the biomedical sciences has been extensively researched in recent years [283-289]. The TREC Genomics tracks offer a good overview into the kinds of challenges presented by biological text retrieval [129, 131], from identifying gene names in biological texts, to newer tasks such as spotting relationships between genes and diseases, or finding interactions between different proteins.

A mature body of literature in bioinformatics now describes numerous strategies for handling biological language and using extracted meanings to process biological data.

Rather than designing an artificial language to describe to gene products, semantic processing takes biological texts and attempts to infer entities and functions from the language therein. An interesting example of using community-derived texts as the basis for bioinformatics tools for functional gene analysis is Gene Wiki [286]. Here, Wikipedia entries for genes are used to create functional categories which can be applied in term enrichment analysis to interpret lists of up- or down-regulated genes.

Natural language processing could provide a powerful adjunct to the Gene Ontology, specifically with respect to automatically indexing large numbers of genes with potential functions. However, no such project exists in the molecular biology domain, with ontologies currently *de rigeur* and a paucity of competition creating a strong barrier to automatic indexing based on natural language. The Gene Ontology was originally conceived as a solution to the vagaries of biological language; to

adopt semantic techniques drawn from natural language corpuses runs counter to the design philosophy of the GO Consortium.

However a pluralistic approach, indexing gene products with terms derived from natural language and a system of weighting or ranking for words and phrases is not necessarily so foreign to the molecular biology domain. If a gene product record in a species database is considered much as any other document indexed and ranked by a search engine, then one can conceive of a system for interrogating gene expression data much as a search algorithm ranks a document set for meaning. The output of a gene functional term analysis would no longer be lists of enriched GO terms, but lists of enriched terms and phrases drawn directly from the molecular biology literature, and scored for significance like any other information retrieval result.

Natural language approaches could even index gene products using different sets of domain-specific terminological sources which are not necessarily resolvable into a single vocabulary. Meanings for the same terms can be different in different domains, and thus under a pluralistic rubric, plant scientists could interrogate gene expression data using natural language search tools based on plant science literature, and yeast biologists could similarly interpret results using semantic tools indexing yeast genes with functional terms pulled from a yeast research discourse corpus.

#### *4.2.2.2 User-generated keywords*

Rather than taking the text of different discourses in the biosciences domain to create vocabularies that can index gene products, we might ask biologists to create their own keywords. This approach is known as 'social tagging', 'collaborative tagging' [290]or the creation of 'folksonomies' [291].

The existing literature on social tagging concentrates on how keywords can be used to index scholarly articles [290, 292-294]. Research has explored the reasons why users tag documents, such as to facilitate information retrieval, to aid navigation across a website or document set, to classify documents with minimal cost or to bring users together into a social network. Some work also exists comparing the quality of socially created tags with professionally created index terms, and the efficacy of these two systems in aiding information retrieval or the browsing experience on websites.

Data sources for research into user-generated keywords in the LIS domain have included the popular photo sharing site Flickr's tags [295], user-generated tags in library systems (and their comparison to traditional subject headings) [296] or the application of social tags to scientific papers on CiteuLike [297, 298]. However there is currently no research investigating the potential for social tagging of gene products in the molecular biology domain.

Major databases provided by the NCBI and EBI collate information on genes and gene products but do not offer facilities for users to contribute keywords or tags that to describe entries. The Gene Ontology discussed the importance of community-derived content in a mailing list thread on wikis in February 2006. However despite providing a wiki page for every Gene Ontology term where annotators and users may add usage notes for the term in question, these pages rarely have content added by GO users, are not used in GO analysis workflows, and are not actively promoted the GO Consortium.

Collaborative tagging has several important advantages over existing ontology approach to indexing gene products.

The scientific community decides the content of vocabularies. As indicated in the previous chapters, problems in deciding ontology content and the propensity of users to re-work and re-categorise ontology terminology are good indicators that there exists interest in alternative 'bottom-up' approaches.

Collaborative tags can be used as the basis for more sophisticated classifications [109, 185, 299-301]. Previous research into folksonomies suggests that the high variability of language used in tags can prove obstructive to efficient information retrieval. However tags can be used as the starting point for classifications that better reflect the thought processes of molecular biologists. The sheer scale of the GO project and the challenge of developing and implementing a large controlled vocabulary make the prospect of a cheap, quick, community-led development an attractive proposal.

With greater community involvement, the contents of scientific classifications are not determined by a small group of experts. Collaborative tagging could act as a way to leverage the collective knowledge of many biological experts. It can also act as a methodology for indexing gene products which permits the construction of multiple classifications to describe the same data object. This is a way to achieve pluralistic classifications which can explain multiple functions for the same gene product from diverse theoretical contexts, and can accommodate the feature of multiple explanans for any explanandum.

#### *4.2.2.3 Faceted classifications*

##### 4.2.2.3.1 What is a faceted approach?

Faceted classifications offer an opportunity to blend vocabulary control with the freedom to create context-rich, compound terminology to index gene products [190, 191]. The presence of pre-existing facets presently encoded into the Gene Ontology structure has been touched upon earlier in which the GO term 'cardiac cell differentiation' was decomposed in various elemental concepts. The proposal here is that this methodology could be extended throughout the GO vocabulary to devise a faceted classification for gene products.

Despite their advantages described above, fully faceted classifications are no longer widely used, due to their awkwardness in use, particularly for arranging printed materials [264, 302]. Recent changes to UDC highlight how a faceted approach can improve classification systems [303]. The advantages in the UDC example and its relevance to improving the Gene Ontology relate to making the vocabulary more coherent, removing compound terms, making the vocabulary smaller and reducing repetition.

Whilst arguably the Basic Formal Ontology structuring for biological ontologies are exceedingly coherent, the profligacy with which GO developers have expanded compound terms and the current size of the ontology make the possibility to reducing the size of the vocabulary an attractive prospect. This is especially the case when considering usability issues for the molecular biology community. A lighter classification for indexing gene products would prove a boon for users interpreting data, and could potentially encourage more biologists to involve themselves in the task of annotation. The complexity and size of the Gene Ontology is currently a major barrier to extending annotations to uncharacterised gene products and to new species. Stream-lining with facets could accelerate gene product annotation which is of primary importance if the GO vocabulary is to prove its utility in knowledge discovery by automated means.

The majority of terms in the Molecular Function and Biological Process Ontologies are compound terms. Vickery writes this is a common feature in special classifications, and therefore a scheme should provide a flexible means for combinations of different terms [190]. Such a faceted scheme also lets users retrieve entries on any individual term within a compound. Complex notational or fixed taxonomies of compound terms do not allow for this form of term decomposition and search.

The Gene Ontology, though it may provide several alternative paths through the ontology back to the root, cannot offer flexible compound term creation or facet search for users, other than via contrived free text methods. Ranking systems for free text searches of the ontology files provided by the GO Consortium are rudimentary, and a full information retrieval analysis of these search algorithms may be a productive area for improving ontology services. However a faceted classification would offer a powerful new way to construct searches of the ontology files, and potentially to analyse empirical data.

#### 4.2.2.3.2 How can we draw facets from the existing Gene Ontology structure?

Although a full re-structuring of the Gene Ontology into a faceted classification would be a considerable undertaking, a brief example a potential methodology can be suggested here. Ranganathan's PMEST formula can be used to re-use and sub-divide existing parts of the Gene Ontology, together with other controlled vocabularies, to sketch a facetted classification for gene products.

The PMEST formula corresponds to the facets Personality, Matter, Energy, Space and Time. In terms of Personality, which has been criticised as being overly vague in Ranganathan's scheme, I would adjudge this to characterise the biological discipline or sub-domain from which the other facets derive. The Gene Ontology has previously tried to incorporate Personality into GO terminology, by using 'sensu' qualifiers, and thus the main species or taxa a set of terms is aimed at describing would be appropriate for this facet. Examples would include specific species like yeast or fruit flies, together with broader facets such as Mammalian model organisms or prokaryotic cell biology.

Furthermore, the MeSH vocabulary also provides adequate descriptions for other disciplines in biology, and thus sub-divisions of MeSH [H01.158] such as anatomy, biochemistry and pharmacology could be used to communicate a sense of the Personality facet for particular terms.

The Matter facet describes the physical material under consideration. In the Gene Ontology, terms are intended to describe gene products, which in the majority are protein molecules encoded by gene cassettes. However, the Gene Ontology has previously struggled to capture the functions of other biologically active molecules encoded by the genome which are not necessarily proteins, such as short-sequence RNA strings which in the last decade have been found to regulate a host of biological processes in living cells.

Complexes of more than one protein, together with polypeptides bound to inorganic molecules are also difficult to represent in the current GO design. Many biochemical products not directly derived from genes play a central role in the functioning of cells. The simplest example is water, which composes the greatest proportion of our own bodies, yet is not a target for annotation by the Gene Ontology because it is not a gene product.

The Matter facet of the PMEST formula can therefore be extended for molecular biology to cover not just protein gene products but also RNA molecules, DNA strings, complexes of more than one protein, and also inorganic molecules including vitamins and ions. All of these matter-types arguably have a functional role in cells. Under the present Gene Ontology rubric, they play no part in functional explanations represented in controlled vocabularies, and this is a major omission.

Energy according to PMEST is any activity that occurs with respect to the subject, the subject being the organism, cell or molecule being represented. The Molecular Function Ontology is an ideal candidate source for activities in the biological sense, and by taking the most annotated high-level terms in this GO sub-ontology, one can derive a simple set of preliminary facets. Examples of these most popular functional activities include binding, catalysis, enzyme regulation, signal transduction, transcription factor activity, receptors, structural molecules, and transporters.

A large number of sub-divisions for Molecular Function Ontology terms are according to the target molecule type for an activity, for example a deeper node in the GO graph linking to 'transport activity' is '*oxygen* transport activity'. Under my proposal for a faceted classification, these compound terms would be easily described by combination with the Matter facet.

The Space facet according to PMEST is the location a subject for description occupies. The Cell Component Ontology is a detailed representation of potential places in different kinds of cells, from membranes and organelles to extracellular locations and structures specific to certain cell-types.

The Cell Component Ontology could therefore be simply re-authored as a set of detailed facets describing where activities are occurring in cells. The Gene Ontology has long considered the creation of cross-ontology linkages to enable this kind of content description, yet has never implemented the proposal because of the logical complexity of encoding these relationships into formal ontological descriptions.

The Space facet would also be extendable to other anatomy-specific vocabularies to capture the action of gene products in particular organs, such as kidneys in Mammals or leaves in plants.

As to the Time facet of the PMEST formula, this is dealt with only fleetingly by the Gene Ontology. This is again another serious omission, since biology is the study of life over time, be it through the existence of an organism from conception to death, or the life of species over an evolutionary history. My proposal is to represent a Time facet for molecular biology according to periods in the life of a cell, such as cell division or growth. Beyond the cellular level, a Time facet would also describe phases in the life of a whole organism, such as birth, growth, reproduction and death. Finally, additional Time facets could include tasks organisms spend time performing, such as feeding, excretion, communication and movement. Many of these tasks correspond to roles described in the Biological Process Ontology.

Table 31: Some examples of facets for describing gene product functions

| Personality | Matter | Energy | Space | Time |
|---|---|---|---|---|
| **physiology** | FLT1 protein | differentiation | cardiac tissue | gestation |
| **pathology** | FLT1 protein | cancer cell invasion | cancerous epithelial cells | adulthood |
| **plants** | light | growth | leaves | maturation |
| **cell biology** | ssRNA | transcription | nucleus | cell division |

### 4.2.3 Testing pluralistic classifications in the molecular biology domain

A large amount of the information I propose capturing with a faceted classification is encoded in the network of terms, annotations and entities described by the Gene Ontology project.

For example, annotations to gene products indirectly communicate the relationship of a Matter facet (a gene product) to an Energy facet (like 'binding'). Furthermore the Gene Ontology annotation files do indicate the database source of a gene product entry, such as a species database for *Drosophila melanogaster*, which under the PMEST formula largely corresponds to the Personality facet.

However in a GO term enrichment analysis, this contextual information is lost, and all annotations and gene product database sources are rendered equal. This is intentional by design, because all molecular functions in different species are assumed to be equivalent. Yet this assumption deletes a large quantity of rich, accessory data that biologists would normally incorporate into their theoretical work. It also means that biologists are restricted to analyses across the whole ontology, rather than restricting themselves to particular areas of interest in the GO graph.

The GO Consortium is aware of this, and rather than allowing users to delete arcs of the GO graphs they consider less important and creating bespoke ontologies specific to their information needs, there are offered 'GO Slims', or simplified GO graphs which are tailored to various disciplines.

The real power of a faceted classification would be in allowing users to flexibly interrogate compound terms constructed according to the PMEST formula, and to home in on sets of facets which are particularly relevant to specific biological problems. The structure of the faceted classification would also make it easy to index *any* kind of document with terms, including original research articles, empirical data or images. By relaxing the rules of ontological realism, the indexer is permitted to describe any object, not just gene products.

A faceted classification is also an objective classification. Definitions and scope notes for each term in a facet restrict their application, whilst the faceted structure itself offers considerable freedom for the user to create meaningful index terms that can capture a plurality of nuances.

How could a faceted classification be tested against the Gene Ontology? A straightforward challenge might be to index a set of gene products with terms from a faceted classification, and compare these to an existing Gene Ontology annotation set. User behaviour studies can be used to assess how relevant users find two different sets of index. Information retrieval tasks could appraise whether, given specific initial information queries, whether a faceted classification retrieve more relevant gene product results.

An interesting potential application for a faceted classification would be in inviting users from the molecular biology community to choose facets they feel are relevant to data objects. On the gene product entry page for a particular species database, users could select terms from boxes corresponding to each PMEST facet. Since there are no complex compound terms, and a faceted system might offer more flexible than the existing pre-coordinated Gene Ontology terminology, users could be more inclined to participate in efforts to index and describe databases entries. Such an approach would blend the formal structure of a controlled vocabulary with the freedom for users to index content with terms they feel are relevant.

The creation of index terms for database entities would possibly be simplified using a faceted scheme. Natural language processing tasks could be deployed to index large numbers of gene products immediately with useful facet terms, such as extracting all the gene names in cardiovascular text corpus and indexing the matching entities in a database with the Space facet ‘cardiovascular system’ or ‘cardiac cells’. Text processing is already well advanced in the molecular biology domain, and a faceted classification would be a way to make use of this expertise to vastly expand the indexing and description of gene products in species databases.

Finally, the Gene Ontology is used extensively to index uncharacterised gene products. These are usually gene sequences derived from genomic mining techniques which are likely to encode gene products in an organism, but for which there is currently no empirical evidence for function. Many of these sequences come from the genomes of new or poorly researched model organisms, for which there are not the resources to test, one by one, the biological activity of each putative gene sequence.

A faceted classification could be used in the same way as the Gene Ontology to propose functions for uncharacterised gene sequences. A faceted scheme could indicate sequences of interest not currently describable according to GO terminology, thus improving recall, such as genes known to be involved in pathological processes or genes that usually function in particular organs.

# 5 Summary and Conclusion

Classification has always been central to the growth of knowledge in biology, from the creation of taxonomies of species to the design of ontologies for indexing gene products in the post-genomic era.

Biologists have an information need to classify gene products according to functions, in order to make predictions about new nucleic acid sequences, and to explain why sets of genes are differentially regulated. Genomic data sources are now so large that without classifications like the Gene Ontology, it would impossible to integrate and analyse data across the burgeoning biosciences information infrastructure. The Gene Ontology has therefore served an important purpose in the domain, helping biologists leverage complex data sets to offer hypothetical functional explanations for changes in gene expression.

The classification of species has been guided in part by essentialism, and the need to explain why different kinds of species exist. In the 20$^{th}$ century, various critics questioned the principles guiding species classification, and even today there are unresolved issues regarding how taxonomists classify, and whether objective, universal classifications for species even exist.

In comparison, the classification of gene products has been largely guided by a commitment to realism, and the assumption that there is an underlying unity to molecular functions across different model organisms. This has led to the development of ontologies like the Gene Ontology, which are constructed according to special formalisms and tenets of ontological realism.

The original aims of this research were three-fold. Firstly, it aimed to understand the rules for creating ontologies unique to the molecular biology domain (Aim 1). These rules, explored in some detail throughout the thesis, have been informed by the overall aim of the Gene Ontology project itself, which was to create a universal classification for gene products. The vagaries of scientific language and conceptualisations for common functions in the domain created certain problems for the project. The broad commitment to ontological realism solved many of these problems for the developers, by constraining the kinds of classes and dependencies that could be described by the ontology.

Secondly, the thesis aimed to explore how the Gene Ontology developers created their controlled vocabulary (Aim 2). The developers created their classification system without reference to existing vocabulary standards, and grew the project over time amongst a small group of motivated editors and content creators. In trying to maintain high standards of scientific objectivity, certain design decisions proved controversial amongst the developers. Ontology contents we decided through negotiations and, where consensus proved difficult, by the exercise of authority by senior developers.

Thirdly, this thesis sought to investigate how the Gene Ontology is used in analysing gene data. Interestingly, although the Gene Ontology has developed in accord with the requirements of ontological realism, aspiring to high standards of objectivity and with careful consideration paid to obsoleting terms in an effort to make the ontology better, results suggested that biologists tend to re-work terminology in reporting their research, and cited Gene Ontology data quite poorly. Perhaps there may be a difference between what the Gene Ontology vocabulary provides, and what molecular biologists want to use classifications in their research.

This thesis presents LIS domain methods applied to the design history of the Gene Ontology. Guided by the overarching research approach of Hjørland's domain analysis, a mixed methodology was used to study the development of this special classification in the molecular domain, in order to ask whether ontological realism offers the best philosophical framework for classifying gene products.

Through a concept analysis of a single GO term, 'cardiac cell differentiation', it was found that ontological realism is closest to the epistemology for concepts known as rationalism (see Research Objective I). This GO term was presented according to other, rival epistemologies for concepts, and this illustrated how different theories for concepts can change how a GO term can be represented in a classification.

This result represents evidence that ontological realism is not the only way to represent scientific concepts in biology, and other methods for classifying gene products could be designed around alternative epistemologies. In particular, ontological pragmatism, whereby classifications for gene products are designed around specific research goals, may offer an alternative way of indexing gene products which is consistent with realism and can represent concepts not currently described by the Gene Ontology.

In addition, it was found that the Gene Ontology acts as a boundary object in the molecular biology domain, bridging competing conceptualisations for knowledge from different sub-disciplines. This realisation is interesting because it opens the Gene Ontology to further study using social research methods. This result serves to offer some resolution to the research question 1, which asked why ontological realism is a core tenet in the design of ontologies for the molecular biology domain. Vocabularies like the Gene Ontology serve to bridge conceptualisations for knowledge in various sub-domains in biology. Adopting ontological realism makes it easier to construct a boundary object like a classification to bring these disciplines together.

The analysis of vocabulary construction standards in the Gene Ontology found that the GO project made no reference to existing standards during its development (see Research Objective II). In designing the ontology, the developers still managed to incorporate many important features and rules from the NISO Z39.19 which were devised from first principles, without reference to LIS domain expertise.

However, the GO vocabulary exhibits poor synonym control, and has no explicit warrant to explain how it chooses preferred terms or incorporates new knowledge into the ontology. By committing to ontological realism, the Gene Ontology tries to circumvent certain semantic problems which controlled vocabularies are designed to solve. In fact, GO developers are inventing non-canonical term names to describe nodes in the GO graph, a practice which poses the risk of exacerbating the semantic problems in the domain that the ontology is intended to solve. This practice also renders the ontology less objective, and introduces strong element of subjectivism into the representation of knowledge in the ontology.

Furthermore, the Gene Ontology cannot represent hypothetical entities, a potentially useful pre-condition for creating classifications of gene products that support the growth of original knowledge. The design of ontologies which do represent tentative or hypothetical knowledge therefore may offer an opportunity to modify how gene products are presently classified.

A critical discourse analysis of GO mailing list texts proved to be a rich source of evidence for how the ontology developed over time, and results showed how mailing list or message board threads can be used to study electronic-mediated conversation (see Research Objective III).

Social relations did play a part in determining ontology content and senior GO developers did exercise their power and authority in order to determine ontology content, a process described by Susan Leigh Star as clearance. Clearance in the Gene Ontology implements arbitrary definitions and structures in the ontology, which manufactures the appearance of objectivity in the vocabulary, facilitating the growth and acceptance as a tool to support data integration in the domain.

However GO users encountered difficulties when trying achieve consensus on definitions for high-level concepts in the ontology relating to reproduction. Results suggest that different epistemologies for concepts and individual subjectivity play a central role in determining how different biologists adopt different understandings for the same scientific concept.

In answer to Research Question 2, which asked how the Gene Ontology developers have tried to make the ontology objective, the critical discourse analysis suggests that editors have created non-canonical names for functions and have negotiated ontology content at controversial nodes in order to end up with a classification that looks as objective as possible. The exclusion of hypothetical entities also renders the Gene Ontology more scientific. This is a design choice which makes the ontology better at retrieving information on gene products from different databases.

Term obsoletion in the Gene Ontology is indicative of the clearance of out-moded knowledge representations from the molecular biology domain (see Research Objective IV). The most important kinds of terms obsoleted from the ontology were gene product names, which are often used to denote molecular functions in the domain literature.

In obsoleting these kinds of terms, the Gene Ontology is removing a potentially useful way of describing gene functions that many biologists would commonly use. The short time-frame for obsoletions and ease with which the GO Consortium has cleared these representations is much like Latour's scientist working at 'fact production' [255]. Obsoleted terms may represent alternative ways to classify gene products, and suggest how other methods could be used to classify gene products in ways that are meaningful to biologists. This highlights a limitation of the Gene Ontology as queried in Research Question 2: because of the ontology rules devised over time, particular ways of thinking about gene products are necessarily excluded from the ontology, and this may serve to limit the effectiveness of the vocabulary as a data management tool. It is potentially rendering classes of gene products as 'invisible'.

In turning to the peer-reviewed literature which uses the Gene Ontology to analyses empirical data, compliance with GO data citation policy was found to be poor (see Research Objective V). Authors extensively re-worded GO terms and presented GO data in ways which served to potentially confuse readers. The majority of GO data analysis tools exhibited poor transparency, and the consequence is that Gene Ontology analysis results in the literature are difficult to trust and difficult to reproduce.

A solution to this finding is that journal editors and peer-reviewers might be encouraged to mandate authors to appropriately cite Gene Ontology data and GO terminology.

A broader conclusion to this thesis is that it is possible to practically apply Hjørland's domain analysis to study elements of the communication chain in the science domain. This suggests that LIS experts could make considerable contributions to the scientific information infrastructure, and with the expansion of e-science, the LIS domain would do well to address the implications of new information technologies using traditional information science research techniques.

In addition, this research has shown that surprisingly the Gene Ontology has no clear philosophical position on the role of functions in theoretical explanations, and does not provide criteria for distinguishing a molecular function from a biological process. This is surprising because the primary aim of the GO project is to provide a comprehensive controlled vocabulary of functions for gene products.

Cummins' approach of defining containing-systems and the functions of component parts might be a good basis for functions in the Gene Ontology. This conclusion points at ways in which the Gene Ontology could be improved, as posed in Research Question 4. Furthermore, the Gene Ontology itself could act as a source material for philosophers interested in understanding how molecular biologists use functional talk in their theories to explain biological phenomena. The Gene Ontology could be making the same mistakes in trying to create universal classifications of gene products as taxonomists did when trying to create universal classifications of species.

A clear definition for molecular functions, or the admittance of alternative epistemologies for concepts of functions, may make better classifications for gene products.

Ontological subjectivism is one way to solve design problems in the Gene Ontology, and pluralism could offer a better basis for innovative knowledge discovery in the molecular biology domain. Faceted classifications are one alternative to ontologies in biology, and their flexibility could garner stronger community involvement in representing a plurality of epistemologies for scientific concepts.

As it stands now, the Gene Ontology is largely a closed system, where the GO Consortium controls the overall shape and ontology content from a top-down perspective. This organisational structure has facilitated the ontology's development, but has contributed to the problems I highlighted in this research.

Open ontologies could be constructed according to different design principles. Open ontologies might well be better at supporting innovative knowledge discovery, since they could represent the diverse and creative ways working molecular biologists think about gene products and their functions. Open ontologies might also feed into current trends in open science and user-generated digital content.

There is potential to create more alternative classifications for gene products other than the Gene Ontology. Though the Gene Ontology serves an important function in helping biologists store and retrieve functional information about genes from different sources, there are difficulties in it trying to be a universal classification, suitable for all sub-domains in biology. Pluralism and open ontologies are ways that the growth of *new* knowledge in the molecular biology domain could be supported, rather than subscribing to a simplified and rarefied form of *existing* knowledge about the purpose of gene products.

## 5.1 Further work

There are other ways this work on the Gene Ontology might be extended.

The Gene Ontology Consortium has developed an extensive vocabulary for molecular functions in biology, without reference to existing philosophies of biological functions. A clear model for functions in biology which is consistent with how the Gene Ontology understands what the function of a gene product is could be an interesting contribution to the philosophy of science.

The discourse analysis methodology used in this thesis is extensible to other existing discourse data sources, such as the minutes of Gene Ontology meetings or electronic discussions conducted on the Sourceforge platform. A better understanding of how biologists talk about categories for gene products could feed into the creation of improved classifications for biological entities.

The very success of the Gene Ontology project, and the features of its design discussed in earlier sections, means that the vocabulary has become quite large and unwieldy. New users need to invest considerable time and effort to understand the ontology and how terms relate to annotations. A simplified, faceted form of the Gene Ontology suitable for quick, broad classifications of gene products could be a useful adjunct to the main ontology.

No research exists on how Gene Ontology annotators apply the terminology to gene product entries in species databases. Semi-structured interviews to would be a way to give a voice to the annotators, letting them describe in their own words the usage and application of Gene Ontology classes. Furthermore, this kind of further work may assist in tackling the problem of inter-indexer consistency. In fact, the Gene Ontology has never studied this challenge quantitatively, and therefore metrics on the problem of objectively classifying gene products, and the differences between different indexers, might prove most useful.

Finally, an observational study of how biologists use and interpret Gene Ontology classes in their empirical work could broaden the picture of how the Gene Ontology is used. Gene Ontology editors and annotators are ontology experts, well versed in the structure and complexity of their controlled vocabulary. The opinions and understandings of the non-expert ontology user though, who just wants to use the Gene Ontology as a tool to leverage their experimental data, could suggest improvements to the classification.

# 6 Appendices

## 6.1 Notes from semi-structured interviews with Gene Ontology developers

The same email containing a series of short questions was sent to twelve different Gene Ontology editors and annotators in the United Kingdom and United States. The email offered the opportunity to reply via electronic mail, or to be interviewed in-person. The questions asked were as follows:

1. How do you establish the guidelines for how to annotate and how to train new curators? Are there common guidelines between different groups?
2. How are annotation projects funded? What problems are associated with funding these projects?
3. How are annotation priorities and targets established by different groups?
4. How is annotation quality assessed?
5. What makes a good curator? What is the background of curators?
6. What challenges to curators experience when annotating?
7. What tools and resources do curators use to do their job?
8. Are there issues specific to annotation efforts to particular species / biological processes?

The response rate was poor, with only two individuals agreeing to a face-to-face interview. However, the notes from these two successful interviews are included in this appendix because of the context they provide to this thesis, and the anecdotal evidence they offer supporting persistent classificatory problems confronting the Gene Ontology, such as maintaining objectivity and applying GO terms universally across different species.

### 6.1.1 Notes of an interview with ED, an annotator with the European Bioinformatics Group, March 2010

- On the GO Consortium : ED described the Consortium as a real success, bringing together biologists from diverse fields under the banner of GO. Discussions in the Consortium sound complicated, potentially passionate, and very much like negotiations on how the ontology is going to develop, accommodate the interests and existing practices of interest groups and establish standards which everyone can agree upon.

Existing as a Consortium, ED alluded to the fact that the group has more power, more sway over practices in the biological community as a whole, and over interested parties accessory to the community. For example, the Consortium can get the ear of funding bodies or publishers in an effort to promote GO and establish best practices for annotation. Applications for funding also benefit from a consortium of many different biological areas - it is easier to get money, although since the scale of GO is very large (the entirety of biological knowledge) expectations on the reward for investment is similarly large. Is return proportionate to investment? Does GO represent good value for money, and how could this be measured?

Is the Consortium too large? ED did not think so, and argued that the benefits of collaboration outweigh the complexities of negotiating between many parties, or accommodating sub-domain requirements into the main ontologies.

- On communication within Consortium and interest groups : ED gets lots of emails, and mentioned arriving in the mornings to be greeted by the many messages sent by her American colleagues overnight. Is she overworked? She described the limitations in funding and necessity to be realistic about how GO can be developed. Naturally, sub-groups working on specialised ontology projects would be ideal, but without the money there was little point imagining what could be. Other communication channels mentioned included: Sourceforge (rated highly - ED liked this), various 'camps' and jamborees, face-to-face meetings (both at the EBI and overseas), Skype ("I have a long Skype list"), and other online submission things

- On community / expert involvement : Expert involvement cited by ED as really important, principally because so much can be done by picking the brains of top people in their field. Term development is quicker, and experts can direct ontology developers to appropriate references to support definitions and annotations. Experts have little time though, and contributing to GO is not likely to be a major priority. ED saw this as a shame, because good terms and annotations tie into the wider philosophy of open access to data and sharing the fruits of research. ED talked about the renal group as a good example of successful expert involvement. All that can be offered to experts though is travel expenses and lunch. I compared GO contribution to peer review - experts offer their services as peer reviewers for no remuneration, therefore GO is not setting a precedent.

Annotation suggestions at the point of submitting a paper to a journal (like keywording) would be a positive step, but publishers are a problem. ED said there had been some discussions (perhaps informal?) with publishers, but if there is no obvious economic benefit to adding another item to the long list of requirements for publication, then the journals were not interested. Are there citation advantages to well-annotated papers? How could publishers be convinced that GO terms are valuable, both to users and their economic model?

- On annotation : ED highlighted early on that annotators and the Consortium were concerned about annotation consistency. By this it was meant the application of annotation terms to database entries (genes, proteins, interaction) by different annotation groups in a manner which ensured that users could be confident that all annotations meant the same thing. For example, granularity meant that annotations were always to the most specific ontology term possible, and not to higher level, more general terms. I got the impression that annotation is a highly variable activity, and ED was expressing concerns that annotators were failing to capture the knowledge in papers.

Training was seen as way to improve consistency, and to create new annotators. The problem with the Consortium though is that with many groups all providing manual annotations, it is not easy to provide a single, over-arching training programme. There are guidelines for annotations, and a special group has been created to look at ways to improve consistency through revising protocols or creating automated methods to spot errors. There will be an annotation 'camp' in the summer, where annotators from different groups will come together to discuss issues. ED gave the impression this would be an animated gathering.

I asked ED how many annotators there were. EBI has a small team of 3-4 full-time annotators. Swiss-Prot has as many as 50 people contributing annotations. Other groups were thought to contribute another 40-50 individuals. It was not clear how many of these are full-time. The annotation camp will therefore feature about 100 people.

ED highlighted the importance of manual annotations. In her opinion, there is not enough money to manually annotate all model organism genomes. I asked her to clarify: electronic annotations are essential. There are too many genomes and too man entities to annotate manually, even though manual annotations are highly favoured over electronic equivalents. Manual annotations are always needed to train algorithms and tools, and so here is a secondary value, to software developers. ED emphasised that biology is dependent on electronic annotations, and biologists had to be aware of their limitations.

How much do annotations cost? ED said there were broad costings for annotations, and targets for manual annotation efforts based on the amount of funding received. I did not get the impression that the EBI group was strongly driven by statistics or performance measures. On the contrary, the 'art of annotating' was the sense I got from ED, and the difficulty the annotator has in grasping the complexities of unfamiliar areas of biology, of the structural nuances of GO, of trying to ensure annotation consistency when faced with, for example, style in scientific articles. By this, ED was referring to authors' attempts to 'push' certain conclusions drawn from their data and the annotators task of interpreting these conclusions and deciding whether the scientific method used, data observed and inferences made do indeed meet a GO annotator's standard for applying a GO term to a gene, protein etc.

- On the structure and development of GO : GO exists as three separate ontologies, but more effort is being put in to creating cross-references between the biological_process and molecular_function ontologies. The molecular_function ontology was, in ED's view, an ongoing problem. Almost all processes are clearly linked to molecular functions, but annotators did not consistently apply function terms once they had selected a process term. I asked if GO would be designed differently if they could start again. ED's response was pragmatic, in that this is the ontology we have, and efforts were being directed into changing and redesigning the ontology as it stands. There is too much economic and intellectual investment in GO and too many dependencies to other tools and practices for obvious problems in design to be radically re-worked.

GO development is constant and ED cited statistics of around 250 structural changes to the ontology per month. Hence, constant updates released to the community were very important. This update schedule is a little like the Library of Congress subject headings, and must resist criticisms that a changing classification is a difficult classification for experts to apply and users to understand.

- On the uses and abuses of GO : ED would like more communication with software developers. When asked whether this communication existed in any form the response was that some developers were plugged into GO annotation pipelines, updating their software with new releases, structural changes and so forth.

More of a problem is the diversity of statistical techniques used in the large number of different software tools. ED felt that some software and algorithms completely ignored important structural and design elements in GO, which lead to poor quality analysis results. Users are not necessarily aware of these design flaws. Lots of software is simply not maintained after release which, given the dynamic nature of GO, again creates problems and weaknesses for science based on these tools.

I asked ED about GO term summaries in papers. ED agreed that selectivity in reporting enriched GO terms could create bias. A lack of transparency in how the terms were created using the tool cited

was also a problem. Selection of terms high in the GO hierarchy served a purpose in summarising broad trends in data, but again could be uninformative ("Transcription factors are enriched".

ED felt that automated techniques using GO are only going to become more important given the speed with which genomic sequences for unusual species (eg, chipmunk) are being released. It is now cheap to sequence, but these sequences mean very little without annotation which, unless created using computer tools, can be very expensive. The reference genome project plays a vital role in supporting cross-species annotation efforts.

### 6.1.2 Notes of the interview with V and R from the Cardiovascular Annotation Group, March 2010

Why was the cardiovascular initiative established?

- Annotators V and R were working on HUGO nomenclatures
- Needed a job
- Funded annotators are very rare; V and R mentioned only the EBI team of 4 who are also annotating. This was evidently important to them; there are not many manual annotators, and yet manual annotations are valued.

How are you funded? What would you do with more money?

- Started November 2007
- Funded by the British Heart Foundation
- Application supported by seniors in the cardio group

How is the cardiovascular gene set selected and how is it prioritised?

- An expert committee suggested the original priority list
- Terms also pulled from other cardio gene lists
- Cardio-relevant GO terms with existing annotations also used to build up the list

Tell me what a typical work day is like.

- Work from home
- Read lots of papers
- Get to know a particular process eg, V is working on cardio development

What changes have been made to the way you annotate since starting?

- Moved from working gene by gene to focussing on a process and associated papers
- Taken on teaching responsibilities with Masters students (doing bioinformatics?)

What annotation tools do you and use and why?

- No specific tools mentioned
- R talked about a spreadsheet she uses to identify GO terms that are probably going to need to be added to particular genes

Are you involved in the development of new annotation tools?

- Not discussed

What annotation tools do you need which would be useful?

- No NLP or text mining tools mentioned as important
- Facility to perform good Pubmed searches was highlighted, with reference to preparing new grant applications

Why are manual annotations better than electronic annotations?

- Not discussed directly, although this was clearly an assumption in our conversation
- Text mining tools were seen to be relevant, but at some point in the future
- The average biologist does not know where annotations come from - they are 'just created'

Have you ever compared annotation consistency between yourselves?

- Alluded to differences between themselves
- Also discussed exercises with the Masters students in which they suggested many more annotations than V and R

What do you think about ontology terms?

- Not discussed directly
- Obvious need to create domain-specific terminology, hence they conducted a cardiovascular GO term workshop

Are annotations true?

- Did discuss confidence in annotations

There seems to be a strong holistic versus reductionist philosophy in operation with regard to ontology development. The reductionist is happy to create individual, logically consistent terms to address immediate needs because on their own, these new terms make sense, fit into the GO structure and meet the rules of GO. Yet V indicated that a holistic, systems-wide view is valued and does deliver results not achievable by the 'piecemeal' approach.

V indicated that during the workshop, a whole arm of cardiac conduction terms was created to meet the needs of the cardiovascular community. It struck me that this approach, together with the tissue-specific nature of many GO terms (they look like compounds of two or more other terms), does raise questions about the logic of GO. Why not link external ontologies to GO to create these types of terms, such as using a GO biological process term plus a link to an external cell type ontology? I remember reading that this had been suggested by the Consortium but was resisted because of technical limitations. Does this imply that GO is too big, and has too much inertia now, to make important changes? Is it a victim of its organisational structure?

Annotation is seen as a potential promotional tool for authors. Annotations to articles (linked via PMIDs) appear in databases and acts as channels for users who may not have expertise in the region of GO they find themselves. PMID and the linked article provide context and extra explanatory content to unfamiliar GO terms. If citations are important, annotations should be too since they are

another way of getting an author's work recognised. Annotators do not feel researchers realise or value this.

The move in annotation logic from gene-centric to process-centric was seen as a big step. Previously, genes were taken one by one from the cardiovascular priority gene list and annotated accordingly. This meant that GO terms could be selected from anywhere in the hierarchy and from any sub-domain in biology, and this created a significant intellectual hurdle for annotators to scale. I got the strong sense that annotators want to do justice to the job at hand, and feel they can only do this be really 'knowing' a subject. Therefore V was now working on developmental processes and R on pathways which meant they could really 'get into' the relevant areas of GO and push on with annotating more quickly, rather than spending long periods of time trying to get a grip on an unfamiliar path or process.

Community involvement felt very tokenary. Annotators admitted that scientists had very little interest in contributing to the ontology or annotations. There was a notable lack of interest even within the cardiovascular group the annotators were part of. Having said that, a researcher did ask in the presentation, 'Is my gene done yet?' Researchers do feel strong ownership over their gene, but this is not immediately linked to any motivation to help improve annotations, which is something annotators very much desire.

Comment from a colleague regarding GO analysis and the heterogeneity of GO term enrichment analysis results, depending on the web tool used. 'Pick the best' was the general consensus, which is far from objective but an interesting example of subjectivity and bias in systems biology. This mirrors differences in information retrieval algorithms and how to measure search efficiency. There is likely to be a GO analysis tool to meet whatever result the author wants, and this could be an interesting question to ask authors: why did you use that particular GO analysis tool?

## 6.2 Macro-level reading of the GO mailing list

Full text notes are included as a Word 2007 document file on the Data Supplement CD-ROM.

## 6.3 Discourse text 1: Missing term

| Speaker | Turn | Content |
|---|---|---|
| **BS** | 1: | 1. Hi all,<br>2. We have in our annotaion campaign done an annotation with the GO:id GO:0008729. However, now I can not find this go:id in either the MGI or the QuickGO browser. GO:ids before and after seems to still be in use - see below.<br>3. I thought that even terms that have been made obsolete should be possible to find - is that not true?<br>4. Yours<br>5. BS<br>6. 1 molecular_function term(s) matching query "GO:0008728": GTP pyrophosphokinase<br>7. 1 molecular_function term(s) matching query "GO:0008730": L(+)-tartrate dehydratase |
| **MH** | 1: | 1. Hi,<br>2. GO:0008729 was a secondary ID for GO:0003936 (hydrogen-transporting two-sector ATPase). It should still be there, as should another secondary ID, GO:0004006. I have no idea what happened, but I'll restore the secondary IDs.<br>3. MH |
| **BS** | 2: | 1. Hi MH,<br>2. Thanks for your attention.<br>3. Hope we can find out what happened to these guys and avoid it happen again.<br>4. How are the plans to implement the history function I have seen in the schema but not been implemented?<br>5. Yours<br>6. BS |
| **MH** | 2: | 1. Hi,<br>2. The history things will be implemented at the same time that the database becomes the official version of GO. The go-database list is the |

best place to send specific questions about it.

3. MH

### 6.3.1 Speakers

MH, Senior Developer

BS, Commerical Partner

### 6.3.2 Detailed notes

BS1.1, Informal greeting addressed to the whole mailing list

BS1.2, 'We have' uses the collective pronoun to suggest an inclusive organisation or team; 'our annotation campaign', military language, again with a personal pronoun; 'now I cannot find', switch from collective to individual 'I'; 'MGI or QuickGO browser' and 'GO:ids', technical names to describe different ontology viewing tools, example of jargon in scientific discourse serving to exclude non-expert users

BS1.3, An interesting sentence. 'I thought' indicates that BS is engaged in a personal mental process, and refers to 'terms that have been made obsolete'. This phrasing is present perfect, implying the process started in the past and is ongoing, yet the actor in the process is entirely omitted. In terms of transitivity, 'terms' are treated as a material undergoing a process, and this mirrors the ideology of GO which is that concepts must be instantiated in reality; they are material. The phrase 'possible to find' again emphasises this materiality; ontology terms are objects to be moved, lost, found

BS1.6-1.7, Data output from an ontology query, another example of technical language designed to objectify the process; the entire exchange can actually be considered as a single speech act, querying an error in the ontology which is likely due to a human mistake

MH1.2, MH offers a description of what should be there; terms are described in terms of place; MH favours using the GO identification numbers as shorthand for the term name itself; this is done frequently in discussions on the mailing list, and tends to objectify terms and strip away the context which a name normally lends to a sentence; 'I have no idea what happened...' implies MH is acting as a responsible party in the management of the ontology, and the fix is to restore the missing identity numbers to the ontology file

BS2.3, 'Hope we can' uses a personal, collective pronoun in working together to solve the problem, in keeping with the collaborative nature of the GO project. However, BS is mixing this collective senses, in that in BS1.2 he talks about 'we' in the sense of his team and their annotation efforts, versus this sense of 'we' which includes at least MH and BS, and by implication probably the whole GO project; 'happened to these guys' is yet another example of rewording the same concept for a GO term. The concept of 'a term' is prominent in GO ideology, and this overlexicalization serves to create a jargon for working with ontologies which serves to enforce the structure of power in GO (users who cannot adapt to the terminology are marginalized

BS2.4, Agency is removed here, in that plans are to be implemented, but the sentence does not indicate who should be doing the implementing; if this were to be paraphrased, it would more

naturally be a question directed at MH, a senior representative of the GO project, as to why they have not done any work on the new functionality

MH2.2, 'history things' is a casual rewording of 'history function' suggesting that it is perhaps as sensitive issue; GO must balance the requirements of users with the limited resources it has to hand; as sequencing, BS's issue takes precedence over the switch from flat files to a database version of GO; MH attempts to direct BS's query to another mailing list, which is either a speech intended to be helpful, or a speech act intended to finish the conversation on the main GO mailing list

### 6.3.3 Comments

- This is a short exchange, over two days in September 2001
- The text shows how a technical language around the ontology has developed and been adopted by users
- Errors in the ontology are technical problems to be fixed. The language of the ontology and the developers serves to limit queries and discussions to technical issues which are often solved by editing the ontology file or 'moving' terms

## 6.4 Discourse text 2: Ubiquitin removal

| Speaker | Turn | Content |
|---|---|---|
| **MH** | 1: | 8. Hi all,<br><br>9. I noticed that the function ontology has these:<br><br>10. %ubiquitin ; GO:0005551<br>11. %poly-ubiquitin ; GO:0005552<br>12. %ubiquitin-ribosomal protein fusion protein ; GO:0005553<br><br>13. I think we should get rid of them and add a function node that describes what ubiquitin and ubiquitin-related proteins do: they tag other proteins for degradation.<br><br>14. First, does everyone agree? Second, can anyone think of a more elegant way to word it than "tagging proteins for degradation"? |
| **JR** | 1: | 4. Speaking as a non-biologist, I've always thought that function terms should "sound like" functions, rather than specific products. Even though a specific product may correspond one-to-one with a specific function, (a) that's merely a reflection of the current state of knowledge - other functions may be discovered for the product, and other products may be discovered with that function; and (b) is there any value in saying (for example) that the function of ubiquitin is ubiquitin? I think this is Midori's question. If a user does not already know what ubiquitin does, this term doesn't help.<br><br>5. I know everyone's aware of this issue, but I figured, what the heck, why not harp on it some more, as long as Midori raised this example?<br><br>6. :)<br><br>7. I do think the function ontology has improved a lot in this respect, but there are many examples remaining.<br><br>8. JR<br><br>9. P.S. MH, how about "protein degradation tagging"? |
| **JB** | 1: | 7. Yes I think it a good idea/// |
| **MA** | 1: | 8. MH - go for it ! |
| **MH** | 2: | 9. OK, I did it. The function node "obsolete" now has children! |

### 6.4.1 Speakers

MH, Senior Developer

MA, Senior Developer

JB, Senior Developer

JR, Partner Database

### 6.4.2 Detailed notes

MH1.1, 'Hi all' informal greeting, familiar, addressing as one-to-many

MH1.2 Reference to the function ontology as a discrete entity, 'the function ontology' being a specialist term created by the GO project for its own purposes; also 'has these' assumes the reader will understand the object of the sentence which, as is often the case, are GO terms; transitivity here, in that it is the function ontology which 'as these' terms, rather than the authors themselves (who are reading these messages) being responsible for having added the terms in the first place

MH1.3-1.5, list of GO terms all related to ubiquitin; ubiquitin is a small protein which can mark proteins for degradation; its importance was recognised in 2004 when the Nobel Prize for Chemistry was awarded to the scientists who originally identified the ubiquitin pathway and its broad mechanism of action

MH1.6, Statement consists of a proposal (delete the terms), an indirect justification (the existing terms do not state what ubiquitin does), and suggestion for a solution. Proposal is personal ('I think) but appeals to the group ('I think we'); 'get rid of' is a distinctly unscientific phrase, reminiscent of something dirty or unpleasant; the justification assumes the reader knows why these terms ought to be deleted, and is a syntactic decision from alternative paraphrases such as '...because these terms do not say what ubiquitin does'

MH1.7, Call for comments appeals for unity 'Does everyone agree?' (rather than asking the negative paraphrase, 'Does anyone disagree?'); request for 'a more elegant way' contradicts the view held by some developers that the term string has no meaning, being only a string of words representing an entity in reality; the request for something 'more elegant' suggests an aesthetic importance to the term name, an aesthetic not normally referred to in either the ontology rules or literature

JR1.1, Attempts to articulate what this elegance might be; a long-standing confusion in biology is that gene product names ('ubiquitin') can be synonymous with a molecular function ('tagging proteins for degradation'); inverted commas around 'sound like' suggests the speaker, by quoting, is creating distance from the statement (that molecular functions ought not to share a name with a gene product, and as signs ought to be readily distinguishable); justification 'a' for deleting the terms and replacing with a new, better sounding term string, is hypothetical and 'b' is a rhetorical question; finally, JR re-phrases MH's original question about elegance to say something quite different, asking instead that molecular function term strings should fulfil an explanatory purpose for the user (this sentiment is much more in keeping with the idea that classifications are knowledge storage devices, serving to represent theories and ideas through term structures and relations

JR1.2-1.3, Self-deprecating apology for raising an issue which has been discussed extensively before; use of the emoticon expresses that the point is raised in a friendly manner

JR1.4, This is a criticism of the ontology which still has many long-standing examples of function term names which are the same as gene products

JR1.6, The suggested alternative is not a phrase that is ever used in the literature, and is a technical term invented by JR to satisfy the conditions of a good GO term

JB1.1, Agreement without comment

MA1.1, Enthusiastic encouragement

MH2.1, Confirmation that the nodes have been deleted. The comment that the 'obsolete node has children' refers to a work practice in the Gene Ontology where deleted terms are moved to the 'obsolete' node. Obsoletion is a technical term created for the purposes of the Gene Ontology, and has connotations of old age and redundancy, in keeping with the larger institutional aims of GO. MH's choice of phrase is an example of nominalization common in the Gene Ontology, whereby 'obsoletion' and its variant phrases allows all agency to be removed

### 6.4.3 Comments

- With minimal objection, the ontology developers have 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

## 6.5 Discourse text 3: Reproduction

| Speaker | Turn | Content |
|---|---|---|
| **MH** | 1: | 1. May I propose:<br><br>2. term: sexual reproduction<br>3. definition: Reproduction which is coupled with the combining of genes from two different individuals into new arrangements; usually involves fusion of cells of different mating types to form a zygonte of greater ploidy (e.g. two haploid cells fuse to form a diploid).<br>4. definition_reference: ISBN:0878932437<br>5. definition_reference: GO:mah<br><br>6. ...because I'd like to avoid saying 'male and female gametes' (it sounds very odd for protists and fungi, where the different mating types usually aren't called 'male' and 'female'). Also, I don't think it's necessary to repeat the broader definition of reproduction.<br><br>7. Along similar lines:<br><br>8. term: asexual reproduction<br>9. definition: Reproduction which is not coupled with the combination of genes from different individuals, and results in the formation of progeny that are genetically identical to the parent.<br>10. definition_reference: ISBN:0878932437<br>11. definition_reference: GO:mah |
| **DH** | 1: | 1. Does this definition of sexual reproduction fit for self-fertilizers like C. elegans? How about combining genes from two different haplotypes rather than individuals? |
| **MH** | 2: | 1. Good point; I'd neglected that. I don't mind changing it, bit let's see what the plant people say--haploid wouldn't do for polyploid plants, which is why I didn't specify haploid in the def.<br><br>2. hmm, maybe " ...two different cells, usually from different individual organisms..."? Any other suggestions? |
| **DH** | 2: | 1. Sorry, I meant haploid, not haplotype. It is still early in the morning. |
| **KC** | 1: | 2. Well, here's the full def of reproduction for the Oxford Dictionary:<br><br>3. (in biology) the production by an organism of new individuals that are more or less similar to itself. Reproductive strategies can be divided broadly into two categories: sexual reproduction, which requires the participation of male and female gametes and gives rise to genetic variation among the offspring; and asexual reproduction, which does not involve gametes and leads to the formation of offspring that are genetically identical to the parent. The latter takes various forms: examples include budding, fragmentation, parthenogenesis, spore formation, and vegetative reproduction.<br><br>4. rephrasing the sentence about sexual reproduction could give:<br><br>5. sexual reproduction - involves the participation of two types of gametes (e.g. male and female, a and alpha, etc.) and gives rise to genetic |

| | | | |
|---|---|---|---|
| | | | variation among the offspring |
| | | 6. | that stays nice and broad, would that do? |
| **DH** | 3: | 1. | I like this def. |
| **RL** | 1: | 1. | yea. |
| **LR** | 1: | 1. | Whoa--- here comes California chiming in. |
| | | 2. | Haploid would not be a good choice...there are plenty of polyploids that have sex. |
| | | 3. | For sexual reproduction: |
| | | 4. | Reproduction which is coupled with the combination of genes (genomes?) of two different individual that results in genetic variation among the offspring ; usually involving the fusion of two distinct types of gametes. |
| **KC** | 2: | 1. | Is bacterial conjugation considered to be sexual reproduction? |
| | | 2. | Anyway, I combined mine and LR's suggestions for sexual reproduction to get this: |
| | | 3. | Sexual reproduction - involves the combination of genes from two different individual and gives rise to genetic variation among the offspring; often involves the fusion of two distinct types of gametes (e.g. male and female, a and alpha, etc.) |
| | | 4. | and took the Oxford def to get this for asexual reproduction: |
| | | 5. | Asexual reproduction - does not involve gametes and leads to the formation of offspring that are genetically identical to the parent; examples include budding, fragmentation, parthenogenesis, spore formation, and vegetative reproduction |
| **LR** | 2: | 1. | Can we change does not involve gametes to 'does not require fusion of gametes'? |
| | | 2. | >From 'The Plant Kingdom' by Burns |
| | | 3. | asexual reproduction: |
| | | 4. | Any reproductive process that does not involve the fusion of cells or meiosis; offspring ordinarily are genetically identical to the parent. |
| | | 5. | >From Penguin Dictonary of Plant Sciences: |
| | | 6. | asexual reproduction: |
| | | 7. | The formation of new individuals from the parent without fusion of gametes. This may be achieved by budding, fission or fragmentation in the algae, fungi and lower plants or by spore formation, or vegetative reproduction in higher plants. Individuals so formed have a genetic constitution identical to that of the parent. |
| **KC** | 3: | 1. | I'm not keen on that wording, as it could suggest the use of gametes without fusion of them, and reading the 2 defs below does not suggest to me that any of the asexual reproductive methods of plants require |

| | | |
|---|---|---|
| | | gamete involvement of any kind. |
| **RL** | 2: | 1. i'm only interested in sexual reproduction. but two individuals are too many for C. elegans when they just care to SELF (the wormie term for self fertilization). hermaphrodism, at least among nematodes, is not rare.<br><br>2. DH's def seems to cover this uninteresting form of reproduction fine.<br><br>3. my 2 more c. |
| **LR** | 3: | 1. Im aghast at myself for the inclusion of 'individuals' - quite sloppy- I hang my head in shame. |
| **HD** | 1: | 2. Jumping in at a late stage, at DH's insistence?.<br><br>3. Lets look a a few defs from a different source that might have some useful concepts:<br><br>4. Dictionary of Genetics, 5th ed.<br><br>5. Sexual reproduction: reproduction involving fusion of haploid gamete nuclei, which result from meiosis (Note: meiosis will need to be a child of this; will gamete-creation?).<br><br>6. Asexual reproduction: reproduction without sexual processes; vegetative propagation<br><br>7. Interesting that there was no definition of "reproduction" by itself.<br><br>8. But consider:<br><br>9. Sex: (as a process)<br>10. Any process that recombines in a single organism genes derived from more than a single source. So, in prokaryotes, this may<br>11. involve genetic recombination between two cells (conjugation) or between a cell and an episome (like phage). Eukaryotic cells always involves two organisms/cells and *leads to alternating generation of haploid and diploid cells *<br><br>12. Eudora's chile peppers all over the place here.. |
| **KC** | 4: | 1. it's getting trickier the more I think about it<br><br>2. but, for sexual repro..., we can't specify haploid/diploid, as we need to be able to include diploid/tetraploid, etc. alternation, and partial genome recombination, e.g. bacterial conjugation, which can produce merodiploids, and doesn't involve "gametes" at all.<br><br>3. ? for the worm people, is 'selfing' really sexual reproduction? since all the genes come from one organism?<br><br>4. It seems like the only distinguishing feature is coming down to<br><br>5. sexual reproduction - production of progeny with genetic recombination |

| | | |
|---|---|---|
| | | resulting in genetic variation among the offspring<br><br>6. asexual reproduction - production of progeny without genetic recombination resulting in individuals that are genetically identical to the parent<br><br>7. We can throw in examples for each to clarify. |
| **ES** | 1: | 1. ? for the worm people, is 'selfing' really sexual reproduction? since all<br>2. the genes come from one organism?<br><br>3. It's definitely "sexual" in the sense that the progeny aren't just clones of the parent; meiosis occurs, so that a parent with any genetic heterogeneity at all will produce genetically variable progeny. Moreover, there is a natural (low) rate at which hermaphrodites (with two X chromosomes) generate male progeny (with one X chromosome) even through self-fertilization. In other words, self-fertilization isn't equivalent to parthenogenesis.<br><br>4. Parenthetically, there *are* parthenogenic nematode species.<br>5. See:<br><br>6. http://www.nyu.edu/projects/fitch/fresearch/fevolution/fother/Sexes.gif<br><br>7. for the phylogeny of sex determination in C. elegans and related species. |
| **KC** | 5: | 1. maybe just back to the one I proposed from the Oxford def then, it only specifies two types of gametes... |
| **MH** | 3: | 1. Is bacterial conjugation considered to be sexual reproduction?<br><br>2. No; bacterial conjugation is sex without reproduction. You start with two cells and end up with two cells.<br><br>3. >From Gilbert, Developmental Biology (ISBN:0878932437), p. 32:<br><br>4. "It should be noted that sex and reproduction are two distinct and separable processes. Reproduction involves the creation of new individuals. Sex involves the combining of genes from two different individuals into new arrangements."<br><br>5. That's where I got the inspiration for the defs I proposed; then it started to fray as I tried to take plants and then self-fertilizers into account (the quote from the book doesn't quite fit the latter, unless one stretches the interpretation of "two different individuals"). |
| **ES** | 2: | 1. No; bacterial conjugation is sex without reproduction. You start with two cells and end up with two cells.<br><br>2. In the mating of haploid yeast cells into a diploid, you go from two haploid cells to one diploid cell. And the yeast can stay diploid for a long time after that; it can be maintained as a diploid strain. Should we conclude that yeast mating isn't sexual reproduction either? I thought the whole point of making the definition of reproduction, sexual |

| | | | |
|---|---|---|---|
| | | | reproduction, etc. really broad was that it needed to include organisms such as yeast. |
| **MA** | 1: | 1. | I would like to review all of these emails next week before coming to a "consensus". |
| **JB** | 1: | 1. | I wonder if even with 'consensus' we might mull this over and only implement following the meeting next month. in the meantime, maybe a short 'proposal' can be written for the group regarding this update. There are important distinctions here at the highest level of the ontology. Email may not do it for this one...eh? |
| **MH** | 4: | 1. | But the need to avoid "male" and "female" holds even if we regard Saccharomyces mating as sex without reproduction--for example, Chlamydomonas reproduces sexually (2 cells in, 4 cells out). |
| | | 2. | "Male" and "female" are usually applied to heterogamous species, i.e. those that produce gametes of different sizes (male = small gametes; female = big gametes). |
| **ES** | 4: | 1. | But the need to avoid "male" and "female" holds even if we regard Saccharomyces mating as sex without reproduction--for example, Chlamydomonas reproduces sexually (2 cells in, 4 cells out). |
| | | 2. | I think I did see that point made in last week's discussion, which is why my most recent set of proposed definitions (on Thursday, 12/13) doesn't explicitly use "male" or "female". |
| | | 3. | I still don't quite understand whether yeast mating is or is not something that we want to include in reproduction. If it is, then, to the best of my understanding, we can't use the 1 cell -> 2 cells criterion because it does not in fact apply all that well to S. cerevisiae. |
| | | 4. | It seems to me that: I initially tried, back in July, to define "reproduction" in a way that basically encompassed metazoa; my proposed definition was considered unacceptable because it didn't encompass S. cerevisiae; I finally figured out how we might expand it to yeast on Thursday, after a lot of people made helpful comments; and now ... what? We do want reproduction to include yeast? Or we don't? I can work to either set of requirements but do need to know what the requirements are. |
| **MA** | 2: | 1. | I promised to return to this debate ! By far the best dictionary of genetics is |
| | | 2. | Rieger, MAis & Green: Glossary of Genetics and Cytogenetics. |
| | | 3. | Gustav Fischer Verlag, Jena, 1976 |
| | | 4. | (at least for classical terms). |
| | | 5. | Here are the definitions they propose: |
| | | 6. | term: reproduction |
| | | 7. | goid: |
| | | 8. | definition: The production of an organism by one like itself. |
| | | 9. | definition_reference: ISBN:0387520546 |

10. term: sexual reproduction
11. goid:
12. definition: The regular alternation, in the life cycle of haplontic, diplontic and diplohaplontic organisms, of meiosis and fertilization which provides for the production of off-spring. In diplontic organisms there is a life cycle in which the products of meiosis behave directly as gametes, fusing to form a zygote from which the (normally) diploid adult organism will develop. In diplohaplontic organisms a haploid phase (gametophyte) exists in the life cycle between meiosis and fertilization (e.g. higher plants, many algae & fungi); the products of meiosis are spores that develop as haploid individuals from which haploid gametes develop to form a diploid zygote; diplohaplontic organisms show an alternation of haploid and diploid generations. In haplontic organisms meiosis occurs in the zygote, giving rise to four haploid cells (e.g. many algae & protozoa), only the zygote is diploid and this may form a resistant spore tiding organisms over hard times.
13. definition_reference: ISBN:0387520546

14. term: asexual reproduction
15. goid:
16. definition: The development of a new individual from either a single cell or from a group of cells in the absence of any sexual process.
17. definition_reference: ISBN:0387520546

18. Note that the defn of term: sexual reproduction excludes bacterial conjugation, ie:

19. term: mating (sensu Bacteria)
20. goid: GO:0009291
21. definition: The process of unidirectional (polarized) transfer of genetic information involving direct cellular contact contact between a donor and recipient cell; the contact is followed by the formation of a cellular bridge that physically connect the cells; some or all of the chromosome(s) of one cell ("male") is then transferred into the other cell ("female"); mating (sensu Bacteria) occurs between cells of different mating types.
22. definition_reference: ISBN:0387520546

23. term: mating
24. goid: GO:0007618
25. definition: The pairwise union of unisexual individuals for the purpose of sexual reproduction, ultimately resulting in the formation of zygotes.
26. definition_reference: ISBN:0387520546

27. Comments ?

### 6.5.1 Speakers

MH, Senior Developer

DH, Partner Database

KC, Curator

RL, Curator

LR, Partner Database

HD, Curator

ES, Curator

MA, Senior Developer and one of the original founders of the Gene Ontology

JB, Senior Developer and one of the original founders of the Gene Ontology

### 6.5.2 Detailed notes

Definition count: IIIII – IIIII – IIIII – IIIII - III

MH1.1-5, MH makes a proposal for a definition for the term 'sexual reproduction'. Uses the 'I' form of the verbal request to indicate that this is a personal suggestion. References for the definition are ISBN for a standard textbook in developmental biology, and MH as an ontology editor. We assume that the definition suggested here is an amalgamation of these two sources, although MH does not explain how the dictionary definition has been amended.

MH1.6, The reason for the change is again phrased as a personal request, 'I'd like to avoid saying...' as MH is aware that personal concepts for sexual reproduction are likely to conflict. Transitivity here moves the requirement for a new definition to the species classes 'protists and fungi' rather than the scientists speaking about these species classes. This is realist doctrine in speech; entities in reality demand a better definition for sexual reproduction, rather than people talking about science.

MH1.7-11, A second alternative definition is offered

DH1.1, The first problem: an expert in worm biology queries whether this definition is appropriate. The concept 'self-fertilization' has no equivalent in the lexicon for eukaryotic or protest/fungal biology. The interests of this other research domain are not met, and again the phrasing omits the agency here (a group of biologists with conflicting interests). The proposal is to make the definition more complex. More complex syntactical structures are common to the exercise of power in scientific language, to a more elaborate logic which makes it harder to critique meanings.

MH2.1, MH accepts DH's variation, but widens the debate to include 'the plant people'. Here, slang is used to refer to other research domains in molecular biology, and this is rarely seen in 'official' scientific literature. The mention of haploid versus diploid plants is another level of complexity (see comment in DH1.1).

MH2.2, A modal adjustment 'hmm, maybe' shows that MH is deferring to the expertise of other contributors; this is a negotiation.

DH2.1, Another shift in modality, DH accepts a simple error confusing 'haploid' and 'haplotype'. DH looks for sympathy by reference to the early working hours, keeping the conversation familiar. Thus far, the conversation is between only DH and MH; other readers on the mailing list will soon identify an opportunity to take a turn at speaking.

KC1.1-4, Implicature, as a new speaker offers the Oxford dictionary definition. This suggestion is that no definition thus far has met the required standard. This definition is the longest thus far, and an edited alternative for sexual reproduction is offered. Note the 'could give' in KC1.3, a modal adjustment giving deference to other contributors. Note that there is no discussion behind what makes a good definition, or common standards for definitions in controlled vocabularies. Speakers are rephrasing the definition which Fowler identifies as a lexical process – underlexicalization for variants on sexual reproduction in different species are forcing circumlocutions from each speaker.

KC1.5, Syntactic phrasing obscures agency again, with the new definition ('that' in the sentence) presented without reference to the original author, vested interests, or aims of a good definition. 'would that do' is suggestive of a good enough solution, a compromise for the job.

DH3.1, Simple agreement with the last definition. Note the abbreviation of definition to 'def', an example over overlexicalization with respect to the work of developing the ontology.

RL1.1, Voices agreement with a single 'yea'. More message board readers are becoming interested in the discussion.

LR1.1, Good example of the casual language common to the GO mailing list. Again, this user does not refer to themselves in the first person, here announcing that 'California' is making a contribution to the discussion. Speakers are distancing themselves from their comments, to give an impression of objectivity. This confers authority on their statements.

LR1.2, Sequencing here makes the previously suggested definitions originate from a passive speaker. 'Haploids would not be a good choice' is in reference to the changes suggested by DH, and very quickly the discussion is elevated to a scientific discourse, despite the fact that it is still individual's thinking about the concept of 'sexual reproduction' and trying to accommodate different perspectives. It is inappropriate within the social confines of the ontology development discussion to personalise conceptualizations, and it a norm is to abstract and assume the passive voice (as is found in the scientific literature).

LR1.4, Another definition is suggested; this is the sixth re-write thus far

KC2.1, A technical question to open the message, with no solution being proposed. In part, this is an invitation to the rest of the mailing list to offer an answer, but it also serves as a challenge, a further layer of complexity to the problem. This is an example of implicature; KC is not senior enough to demand that the rest of the mailing list provide answers, but is permitted to throw ideas out to see if anyone bites.

KC2.2-2.3, Another definition, with credit attributed to LR for the new edit. The new definition is described as a 'suggestion' since the discussion is supposed to be open, without authority figures exerting power to decide the conversation. The tentative nature of each proposal is at odds with the philosophy of ontological realism, whereby instances in reality can be represented by nodes in an ontology. It is clear that in fact, the speakers in this message thread are well aware that there is no clear, universal definition for 'sexual reproduction' and certain standards of civility in the negotiation need to be observed, hence the 'suggestion'.

LR2.1, Chooses the first-person pronoun ‘we’ to refer to the next change proposed, highlighting the collective nature of decision making, despite the fact that thus far, several different highly individual conceptualisations of ‘sexual reproduction’ have been shared. The discussion has moved to a definition of asexual reproduction since it is closely related to its sexual counterpart.

LR2.2-7, LR quotes two different sources for a definition of asexual reproduction. Dictionary sources do have value to the developers working on definitions, although individuals frequently contest the validity of these definitions. Speakers are appealing to external authorities to support their arguments, again an effort at objectivity. Personal ,individual takes on these definitions are kept at a distance

KC3.1, ‘I’m not keen’ shows that KC is presenting from a personal point of view. This is a shift in modality commonly seen on the GO mailing list, where speakers show uncertainty or deference to the authority of the group by citing suggestions as personal. Choice of the phrase ‘that wording’ is a deletion; it was LR suggesting the new change and KC chooses not to say ‘your wording’. KC is trying to argue for the removal of any reference to gametes, and this is effort is motivated by a logic of essentialism – what are the necessary and sufficient conditions to unambiguously identify an instance of ‘asexual reproduction’ in reality? Implicature in all these messages is that these conditions do exist, and there must be a definition that satisfies this requirement (based in ontological realism)

RL2.1, Previously only commented with a ‘yea’ to a definition for sexual reproduction, RL now expands on the worm perspective where sexual reproduction occurs without a partner through a process known as ‘selfing’. The introduction ‘i’m only interested’ accepts that RL does not have the power to author the definition and that the following comment is a contribution to a wider debate. Again, a species is given a voice ‘...too many for C. elegans’. Obviously, worms don’t have opinions on scientific definitions; RL is speaking about what the C. elegans research community needs in order to annotate gene products to any node defined as ‘sexual reproduction’. Implicature in the sentence ‘...is not rare’ suggests that any definition has, by rights, to accommodate situations less common to eukaryotic systems, like hermaphrodism.

RL2.2, Sarcasm is deployed in ‘...this uninteresting form’. There are clear tensions at play between the different contributors. RL’s comment exemplified an opinion that less popular model systems are poorly represented within the Gene Ontology. Objectively, all model organisms ought to have an equal voice yet this discussion shows how this is an ideal in the social milieu of ontology development

RL2.3, Final comment is a highly abbreviated form of the colloquialism ‘My two cents’. This is subordination to authority, the recognition that alternative forms of view are going to be seen as personal, and less valid.

LR3.1, A flippant comment, but revealing as LR’s ‘mock’ shame does reflect how the mailing list is still a professional forum, and speakers are revealing the limits of their knowledge and making mistakes. Another representation of power in this discourse, as the authors of errors are expected to apologise to the group for failings, even if these failings are overtly accepted by the rest of the GO group as quite normal.

HD1.1, The phrase ‘Jumping in at a late stage’ highlights that the discussion thus far is well-developed, and that there may be an inertia on the part of new speakers to join this message thread. The meaning here is suggestive of deference again to the expertise already thus displayed; the speaker is risking professional standing by possibly making a poor or erroneous contribution.

HD1.2, The slang term ‘defs’ is part of the overlexicalization in ontology development, with definitions being an important issue, and source of jargon. Reference to ‘useful concepts’ is a good example of the usage of the word ‘concept’ without explaining what it means. In the context of the Gene Ontology, mental concepts are not instantiable, and all conceptual talk is ideally with reference to instances in reality. To ‘have some useful concepts’ is part of the ontology developers’ approach to traditional knowledge sources, where anything ‘useful’ can be absorbed into GO. What the criteria for inclusion, for utility in the sense HD uses it here, is never made clear. This opacity is a source of the GO Consortium’s power.

HD1.3-5, Two new definitions taken from another dictionary. The developers are searching for a combination of authoritative source to support their common (or uncommon) understanding of sexual/asexual reproduction. This is the Bowker/Star process of ‘deletion’ of old ideas and their replacement with fresh concepts through vocabulary development.

HD1.6, This comment is founded on the GO philosophy that nodes have parents, therefore sexual reproduction ought to be a child of a more general process ‘reproduction’ even if this is not normally recognised in the domain. Passive voice here hides the fact that it is the developers and the Gene Ontology project which finds this ‘interesting’. The project is meant to be representing common knowledge in molecular biology, not authoring new knowledge, hence HD avoids stating that as a group , they think in this odd way about terminology.

HD1.8-10, Another definition for sexual reproduction, now abbreviated to ‘sex’. HD emphasises ‘alternating...’ because this is original. This re-authoring of standard definitions is the process of authoring a new understanding of molecular biology. Here, the discussion is really about a very fundamental concept in biology, sex. The passive voice, the presentation of formal definitions, removes agency from this process, most likely because it is controversial for the Gene Ontology developers to admit they are trying to re-write the fundamentals of biology, which they see as messy and inaccurate. They are solving a problem in biology, and for the Gene Ontology to exercise any power, it must be authored by the facts of reality, and not by a social group of designers creating what looks like truth and, by the rules of formal logic, is true.

HD1.11, Possibly a reference to the mail client.

KC4.1, This statement obscures agency, whereby KC is presumably describing referring to peers as finding the definition tricky. KC frames the problem as a personal one, in which each speaker is grappling with what reproduction could mean. The phrasing is transitive because in a sense, the domain of biology has never adequately defined sexual reproduction, perhaps because there is no universal definition. Continued obfuscation of agency or framing of the challenge as one personal to each speaker on the mailing list insulates the wider scientific discourse from criticism

KC4.2, A further expansion of the definition's complexity, with different genome numbers all potentially corresponding to a concept of sexual reproduction. Another abbreviation for 'sexual reproduction' is used

KC4.3, KC refers to 'the worm people'. As a form of address, this is casual to the point of denigrating to the researchers working on C. elegans. Although the address is flippant, there is a difference in power, with the object of worm research being strange and unconventional. The question 'is selfing really reproduction?' is a strong disputation in the context of the mailing list discussion. No previous comments or suggestions have raised a direct challenge in the form KC has spoken.

KC4.4, Effort at closing the discussion, after the preceding attempt to marginalise 'the worm people'.

KC4.5-7, Two more definitions. Comment to 'throw in' examples is a vocabulary choice with the expressive value of getting work done, of sorting out the definitions quickly. KC is trying to resolve the dispute.

ES1.1-7, ES, a new speaker, rejects KC's proposal that the worm process of 'selfing' does not count as sexual reproduction. ES offers a complex rebuttal, which in terms of sentence length is unusual for the mailing list, includes a reference to another source supporting the argument and drastically undermines KC's effort at simplifying the definitions. The implicature here in the response is that it is not acceptable to ES and for the worm people in general to ignore their conceptualisations of reproduction. Consider also that this is only one species causing a problem here; across all species there are likely to be many more unusual sexual processes. These species have no voice on the mailing list, and the worm people are challenging the broader power structure in the Gene Ontology, which aims towards representing the most heavily researched model organisms – mouse, human, yeast.

KC5.1, KC responds cautiously, with the deference in the 'maybe', a return to an authoritative source in the Oxford dictionary and a technical justification for the change. Since KC has taken a turn quickly before any other speakers can make a comment, it seems as though ES's message has had an effect.

MH3.2, Flat out rejection of the suggestion that bacterial conjugation constitutes sexual reproduction. Note that the normal civility common to the list, and the normal modality of hesitancy when making a proposal to the list, is lost.

MH3.3-4, MH quotes, with ISBN and page number, the precise source distinguishing sex from reproduction. The discussion is now becoming a technical one, with evidence brought in to back up MH's claim. This is unusual on the GO mailing list, with most terms being defined without disputation either by individual GO editors or by quoting directly from existing dictionaries.

MH3.5, MH explains the source of 'inspiration' for the definitions (sic) proposed. Interesting choice of 'inspiration' since MH is a senior ontology developer, one of the oldest active participants on the mailing list. Implication is that MH's work is, for want of a better word, divine in origin, an indication of a power MH has over ontology content. Choice of 'it started to fray' then removes the personal aspect; MH is suddenly distanced from the work, the definitions become problems again, to be discussed objectively. MH leaves an enormous amount of potential discussion unspoken. The final comment 'unless one stretches...' is a big question, of how the ontology might approach distinguishing individuals, but MH implies that mailing list participants ought to understand the

complexities of this question. Another exercise of power, of making basic requirements on knowledge for participation in the debate.

ES2.1-2, ES is an active mailing list contributor, but is relatively junior in terms of the Gene Ontology project. He replies to MH's contention that conjugating bacteria are having sex, but not reproducing. ES's technical point is that two different yeast can combine two genomic copies (have sex in MH's reasoning) and then divide over and over again, maintaining a diploid state. ES asks 'should we conclude' a modal shift masking a rhetorical question aimed at challenging MH. ES believes the yeast example presents major problems for classifying sexual reproduction. MH's thinking is binary, after the manner of all good ontologists, whereas ES is asking if a broader, less specific definition might allow processes which look 'a bit like' sexual reproduction to be included. When ES says 'I thought the whole point', this is a personal question, almost directed at himself, asking what the very purpose of the Gene Ontology is. When ES says 'to include organisms such as yeast' this expresses how some species can be marginalised by definitions. ES is questioning MH's and the GO project's power to determine what terms mean.

MA1.1, MA, one of the original architects of the Gene Ontology project, arrives to the message thread. MA immediately cuts the discussion, asking to review the emails in the future, forcing the mailing list, which is diachronos conversation, to pause and wait. This is a strong, forceful action on a mailing list, equivalent to asking everyone to stop talking in a room whilst a leader thinks. MA's usage of quotations marks around the word 'consensus' is an ironic touch, and revealing. Even the top ontology developers realise that the ontology, its structure, the definitions, are a compromise between varying points of view. MA is saying that consensus is achievable in only the loosest sense; contentious areas as sexual reproduction has become are settled by arbitrary agreement, and usually by someone like MA.

JB1.1, MA's comment is echoed by a counterpart of equal standing on the project, JB. This is two of the most senior Gene Ontology authors attending this mailing list discussion, and these speakers highlight that this is considered an important issue. JB agrees with MA's usage of consensus, and proposes a delay until the next face-to-face GO meeting. Important issues are carried over to these GO meetings, which creates a problem for user participation as continued involvement in the debate involves travelling to this meeting. This is a simple exercise of power by JB, in order to take control, and JB places further requirements by asking for a proposal (or written report) summarising the debate. Clearly, the mailing list discussion is being sidelined in favour of more powerful institutional instruments - a meeting, a written report – and JB even says 'Email may not do...'. There is a strong disconnect here, between speakers on the mailing list trying to determine ontology content, only for a senior developer to entirely undermine the value of this electronic communication. Phrasing of 'There are important distinctions...' is a transitive statement; the discussion has been about different individual conceptualizations of what sexual reproduction can be, but JB abstracts the debate and moves it to 'the highest level of the ontology', making it a physical place.

[ES and JB have a separate exchange about delaying the discussion to the next GO meeting, and this is analysed separately under the section 'Less than 2 months. please?']

MH4.1, MH continues the discussion regardless of MA's request, most likely because of the senior status held. The choice of 'need to avoid...' is a nominalization which further insulates decision-making from criticism because it is no longer an individual asserting a change. The requirement here

is founded on the nature of the definition as it now stands, in which sex is to be rendered genderless to accommodate species in which notions of gender have no meaning.

MH4.2, Another technical point, where genders have been re-worded to mean small or large gametes. Note that MH uses the equals symbols, which emphasises the scientific nature of this argument – lay persons would most likely find it odd that these biologists do not recognise traditional male or female identities as Gene Ontology-compliant.

ES4.1-2, ES defends position, taking MH's previous comment as a personal criticism – this is indicated by the sequencing here, Es starting the sentence with 'I think...'. ES continues to personalise the discussion by referring to 'my most recent of proposed definitions...'. This is often an unsuccessful argumentation strategy on the GO mailing list, with passive verbs and nominalizations common to speakers who get the changes they want. Choice of the adverb 'explicitly' carries the implicature that even though a final definition may not use male/female, a common understanding may accept that gender is still important. ES does not have the authority to make this implicature, unlike MA when using the ironic 'consensus'.

ES4.3-4, ES uses the phrase 'quite understand' which is a modality communicating uncertainty. ES is in difficult territory, trying to negotiate a preferred definition. Solution offered is to not include yeast in a definition, and ES's syntax has become increasingly complex. The final response is full of qualifiers and modalities. ES finally asks for a statement of requirements to work to, and this captures the essence of the Gene Ontology project and power as revealed in discourse. There are no fixed requirements, and GO standards are amended as needed. This need though is identified and satisfied by senior developers. ES, who lacks seniority in the social world of the GO mailing list, has discovered that proposals can be rebutted without any clear basis. Reasons are *a posteriori* to decisions, such as whether or not to include males and females in definition, or whether to accommodate yeast biology.

MA2.1-4, MA returns to the mailing list, and offers an opinion. MA claims that 'by far the best dictionary is...'. This deletes agency – which would be in this case MA stating his personal opinion – and offers no modal adjustment, uncertainty or qualification. The only concession MA makes is in MA2.4, where it is advised that this dictionary is the best for 'classical terms'. The suggestion is almost aesthetic, and confers to a norm in molecular biology which is the drawing of distinction between 'classical' and 'modern' genetics. The definitions are presented as old answers to an active and current discussion on the GO mailing list, and MA is exercising authority to resolve these disputes.

MA2.5-26, MA quotes verbatim the definitions taken from this dictionary, but displays the information in the form of 'draft' GO definitions, with term names and blank spaces left for anticipated GO id reference numbers. MA and MH are the only list speakers to present proposals in this form, and this stylistic convention is very much a kind of speech act. Presentation according to GO format communicates professionalism and compliance with GO Consortium conventions. Furthermore, MA's suggested definitions are by the far the most elaborate in terms of syntactic complexity and language used – they would look, to a lay-person, like the most 'scientific' definitions seen thus far.

MA2.19, MA resolves the question over bacterial conjugation by reference to two pre-existing GO terms for ‘mating’, one which is a general mating terms which does not specify male of female involvement, and a second term which is a now defunct ‘Sensu’ term, marking a ‘mating’ definition to be used in specific reference to bacterial mating (‘Sensu bacteria’). MA suggests in MA2.18 that a definition for sexual reproduction need not include bacterial conjugation, for it would be covered under ‘mating (Sensu bacteria)’. Modality for this statement is assertive ‘Note that...’, employs a lexical variant of definition (‘defns’) to communicate that MA is adept at this type of ontology work and in MA2.27, MA simply asks ‘Comments?’, seeming to be confident in his proposal. The message, which is one of the longest in the thread, is professional, complex, assured and conforms to GO standards for GO definition authorship.

## 6.6 Discourse text 4: less than 2 months. please?

| Speaker | Turn | Content |
|---|---|---|
| **ES** | 1: | 15. On Fri, 14 Dec 2001, JB wrote:<br><br>16. > I wonder if even with 'consensus' we might mull this over and only implement following the meeting next month.<br><br>17. This is early December; the GO meeting's in early February. So the GO meeting's in two months (not one). Must we wait that long?<br><br>18. (In fact, I thought MA said he had planned to go over this next week.)<br><br>19. > Email may not do it for this one...eh?<br><br>20. Gosh, I hope that's not true. I had really hoped that we would use electronic communications to make things happen.<br><br>21. Our actual face-to-face meetings are necessarily infrequent, and not everybody can attend them all the time (I will not be able to attend the Tuscon meeting, for instance).<br><br>22. If it's really so crucial that I physically be in the room with MH and MA to resolve this, I don't see the point in waiting two whole months, then going to Tuscon, and then vainly trying to compete for their time and interest in a room of thirty people meeting in Tuscon for two seven-hour days. Instead, I think I'll make an appointment to meet them myself in Hinxton, UK, and just fly myself out there when I have some hope of getting their undivided attention.<br><br>23. Should I have to do that? Because I can, and I will if I have to. Is *that* what's required?<br><br>24. Or can we just use e-mail?<br><br>25. --ES |
| **JB** | 1: | 10. ES,<br><br>11. Just a suggestion...it is an important high level change, the discussion moves forward, so maybe email will suffice. If people via email are ok with suggested structure and definitions, then it is settled. You sound frustrated, but we do try to work by consensus, and you, MH and MA do not a consensus make. reproduction affects us all...:)<br><br>12. Hopefully we can think this through before Feb.<br><br>13. JB<br><br>14. On Fri, 14 Dec 2001, ES wrote: |

15. > > No; bacterial conjugation is sex without reproduction. You start with two cells and end up with two cells.

16. > > In the mating of haploid yeast cells into a diploid, you go from two haploid cells to one diploid cell.

17. exactly, 2 -> 1

18. above MH was talking about 2 -> 2, where there is an exchange of DNA, but

19. > And the yeast can stay diploid for a long time after that; it can be maintained as a diploid strain. Should we conclude that yeast mating isn't sexual reproduction either? I thought the whole point of making the definition of reproduction, sexual reproduction, etc. really broad > was that it needed to include organisms such as yeast.

20. > --ES

---

### 6.6.1 Speakers

ES, Curator

JB, Senior Developer

Note: The discussion thread here comes under the title 'Less than two months. please?', and is a secondary thread originating from a longer discussion about definitions for reproduction. The '>' symbol indicates a message line quoted from a previous message; messages have been formatted to facilitate reading.

### 6.6.2 Detailed notes

ES1.1-1.2, ES starts by quoting a line from another discussion on the GO mailing list; without the context – an extended discussion over a term definition – the comment is difficult to understand, but this kind of deletion is common on the mailing list; the speaker in the comment is hesitant ('we might'),a modality suggesting deference to the rest of the GO group; 'consensus' is inverted commas to lend an ironic twist since consensus has proven difficult to achieve; transitivity in the phrasing 'only implement' removes the agency in who is implementing what – the GO Consortium is the entity with the power to make decisions and edit the ontology, not the individual users

ES1.3, Corrects an error using a complex structure, to be polite to the original speaker and in deference to the power reading the message, since it is unusual to criticise procedures for discussing the ontology comments; 'must we wait' is an imperative modality for a question, contesting that the GO authority is unreasonable in dealing with this term change, and appealing to a personal collective in opposition to that authority – the user that needs the terms to get work done

ES1.4, 'I thought MA said' is another uncertain modality, as ES questions whether the information is correct; 'go over this' shows that although agency is clear (MA said he would look at the problem), the object of the sentence, even in the context of the previous statements is still not clear. The

ongoing vagueness is suggestive of a deletion, where the social group of scientists discussing the ontology changes know what the problems are, but continue to suppress or ignore precisely what the basis for the controversy is. The problem is instead rendered as an opaque, bureaucratic issue which is in the hands of the GO power brokers to resolve

ES1.5-1.6, The email discussion lists offer a relatively open forum for ontology users to express how they want the ontology to develop. However when changes are deemed major, either in the sense that they will alter the structure of the ontology or the rules governing its construction, decisions are pushed on to the agenda of a GO meeting. 'Email may not do it' is transitive in the sense that the agent responsible for the process of emailing is not stated, and the user quoted has sequenced this statement to remove the identities of any persons who may be doing the emailing. This successfully abstracts the problem, which is no longer about real people talking about what they understand by the term 'reproduction', and instead creates a nebulous issue which only the authorities present at a GO meeting may resolve. ES is frustrated, and it is unusual in academic mailing lists to express oneself like this

ES1.7, Contrast between the collective and individual pronouns ('Our [...] meetings' versus 'I will not be able to attend')

ES1.8, Very little transitivity here, as ES expresses in personal terms feelings about how he feels about the dispute resolution process. Ellipsis continues, with the nature of the problem not stated and the focus kept on the dissatisfying bureaucracy in the GO Consortium. As a speech act, ES is trying to get the senior developers to accept his solution that email should be used to resolve problems. Implicature is at play, as it is a common understanding on the GO mailing list that difficult decisions are reserved for GO meetings or expert panels

ES1.9, Rhetorical statement, designed to reinforce the speech act in ES1.8; really not clear who ES believes he is addressing, and it is only by not directly naming anyone that he achieves the distance necessary to express himself this forcefully. As modality, ES choice of phrasing - 'have to', 'will' and 'required' - clearly expresses what is perceived by the organisation to be desirable, though this does not correspond with what ES feels

ES1.10, The 'we' pronoun identifies ES with the users, 'just use' is yet another rewording of what ES wants, and can be identified as the fourth attempt to say 'I want us to use email to resolve this problem'

JB1.2, JB is a senior GO developer, therefore her responding to ES's problem is a sign that the issue is taken seriously. It is unusual for users to voice strong complaints about the organisation of the project. In the first sentence of JB1.2, no reference is made to the actual issue ES has raised. JB is trying to be diplomatic and assuage ES, yet through deletion we are unclear as to whether the technical question is the issue here ("How do we define sexual reproduction?") or ES's problem with the lengthy discussion period and difficulty at getting opinion's heard. "If people via email are ok..." obscures the agency behind decision making since in large part the discussion referred to has revolved around a small number of senior GO developers, rather than any wider community implied by the choice of the word 'people'. Sequencing here in the final sentence by starting with 'You sound frustrated...' suggests that ES is the primary problem, and the collective noun 'we' implies a power agency in agreement. The reference to requiring consensus is contradictory to ES's experience on

the mailing list, where for some ontology changes, only one or two people are involved. There is implicature here; sexual reproduction as an issue is different.

JB1.3, 'Hopefully' is a modality suggestive of an aspiration, rather than a definite proposal. Again, agency is suggested as 'we', whereas ES's issue is precisely that collective decisions are frustratingly difficult to arrive at, take a long time and require excessive effort by some participants to be heard. Again, deletion of what is being 'thought through'.

### 6.6.3 Comments

- The fact that JB does feel it is necessary to postpone any decision on a restructuring of reproductive terms implies that not all terms are equal; some are considered more important than others and in this example the senior editor feels that an email discussion is an inferior communication method to sitting down for a meeting. This assumption disenfranchises all those users who cannot attend GO meetings

## 6.7 Bibliometrics

Bibliometric methodologies include techniques to analyze publication data and citation relationships [304-308]. The information science domain has a long-standing history of publications dealing with the structure of various publications corpuses, and specifically there exists much work applying bibliometric techniques to the biosciences domain, for examples see [309-314].

Several sections of this thesis refer to a corpus of Gene Ontology research constructed by searching popular scientific literature databases, collating metadata on Gene Ontology publications, and processing the publication data for patterns using several bibliometric techniques. This corpus was therefore a subset of the broader molecular biology literature, which itself subtends a range of other domains, including genetics, clinical medicine and bioinformatics.

This technique was used to demonstrate that Gene Ontology applications and research has grown since the project's inception, and that the ontology is actively used and referred to in the molecular biology domain literature.

### 6.7.1 What can bibliometrics tell us about ontologies in biology?

A research question in this thesis: how can the Gene Ontology be improved? This question rests on the assumption that Gene Ontology is used in the first place, and that by making it better, it might be used more. Gene Ontology usage is therefore quantified, to an extent, in the subsequent section, using evidence from the published literature which indicates how ontologies are an increasingly popular tool in the molecular biology domain for the analysis of gene expression data.

A simple bibliometric approach was adopted to analyse a sample of published articles from the biosciences domain. Since the Gene Ontology was only initiated in 1996, and prior to this there were really no applications for ontologies in the sciences, this sample covers the period from 1996 to 2010. Bibliometric techniques does not 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 validity of using the Gene Ontology as a research subject questioning how best to design classification for biology is supported by using bibliometrics to show that:

- Publication rates on ontology topics in the biosciences domain are increasing year-on-year
- Biologists are citing ontology papers, since this indicates that ontologies matter in the domain
- Ontology papers are published in a wide range of biosciences journal, from various sub-domains, indicating that they are not simply a specialist tool

### 6.7.2 The dataset

Searches were restricted to Ovid MEDLINE and ISI Web of Knowledge since these provide good coverage of published articles in peer-reviewed journals. Published articles offer a measure of quality; they have been read and reviewed by peer experts, and edited and published according to the standards of scientific journals. ISI Web of Knowledge does not provide such strong coverage for conference proceedings, nor for the mass of grey literature available in the sciences.

The bibliometrics dataset presented below therefore offers a quantitative snapshot of Gene Ontology research in the biosciences, albeit a snapshot of arguably the most important literature source in the sciences, published articles. The search strategy for constructing the dataset is described in Table 32.

**Table 32: Search strategy to construct Gene Ontology papers corpus**

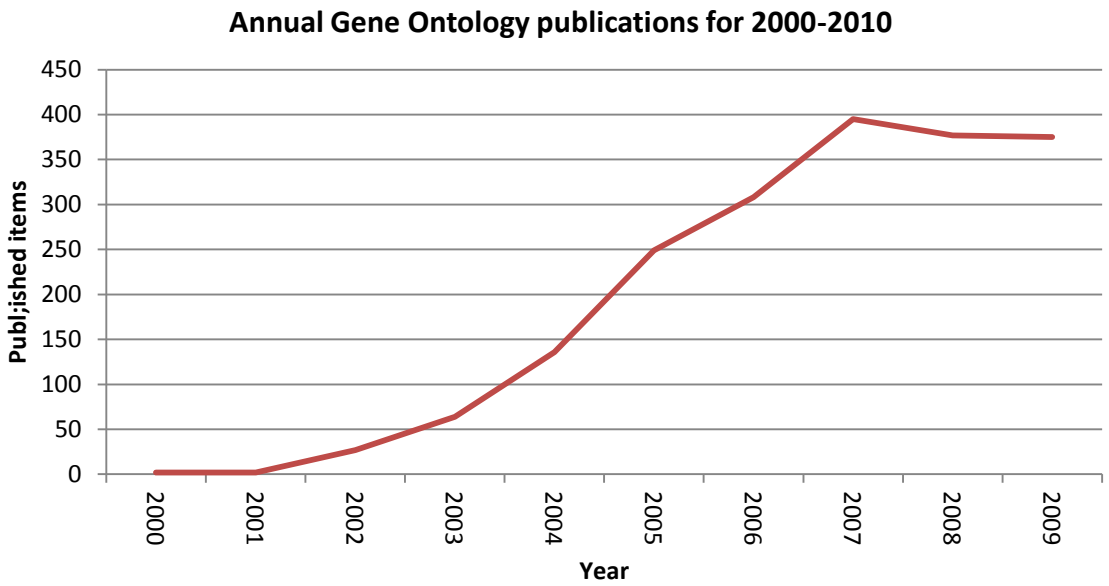
| Search | Method |
|---|---|
| **Ovid MEDLINE** | • Searched Ovid MEDLINE(R) <1996 to January Week 3 2010><br>• Used (gene adj ontology).mp. [mp=title, original title, abstract, name of substance word, subject heading word, unique identifier]<br>• Retrieved 2101 records<br>• Imported ui, author, title, journal, year into an Access database |
| **Web of Science** | • Searched WoS for Timespan=1995-2010 Databases=SCI-EXPANDED, SSCI, A&HCI, CPCI-S<br>• Used Topic (gene AND ontology)<br>• Retrieved 3472 records<br>• Imported all to Access database |
| **Cross-match** | • Searched for each title in the Web of Science results with detail from the MEDLINE records, and copied each corresponding PMID<br>• Resulted in 1935 matched records with citation information in Web of Science results |

### 6.7.3 Total number of articles and annual publication rates

The total number of published articles is an indicator of research productivity in a domain, and as of 2009 a total of 1935 articles had been published relating to the Gene Ontology. The first articles were published in 2000, indicating that the research using GO is likely to still be in its infancy. The top publication year thus far, based on the dataset, was 2008 with 395 Gene ontology articles. Preliminary data based on Web of Science citation reports for September 2011 indicate that with newly indexed literature, 2010 is now the top year for published GO articles (results not shown). Therefore the research field demonstrates annual growth (shown in Figure 14).

The Gene Ontology 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 key paper by Ashburner *et al.*, which were to be followed over the successive years by a growth in the publication rate of articles relating to the Gene Ontology. Publications rates have stabilised in 2008/2009 and each year several hundred papers are published which use the Gene Ontology.

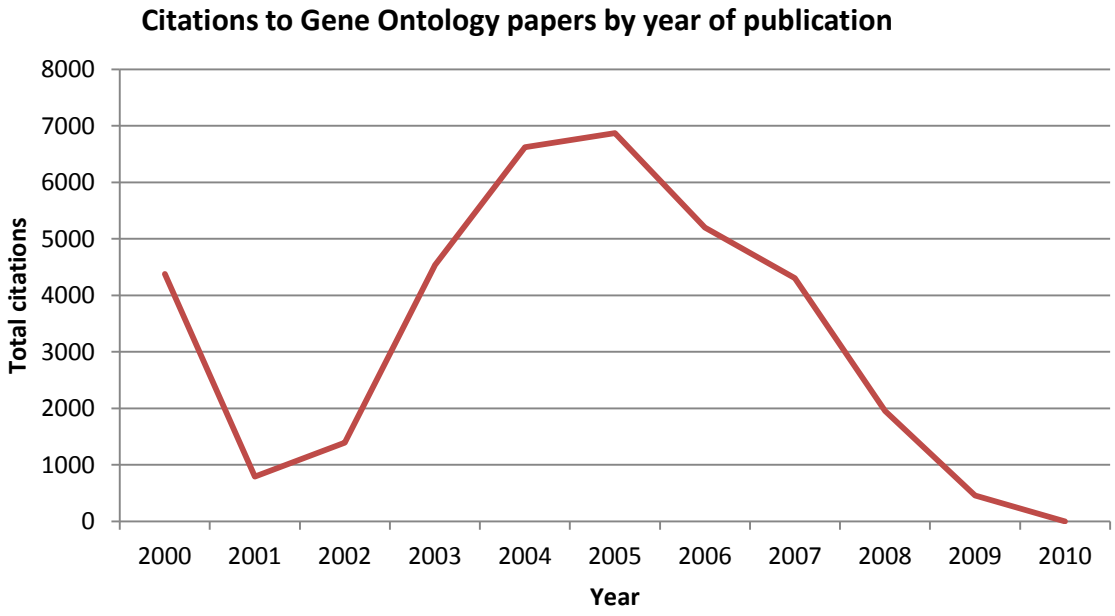
Figure 14: Annual Gene Ontology publications for 2000-2010




### 6.7.4 Citation rates

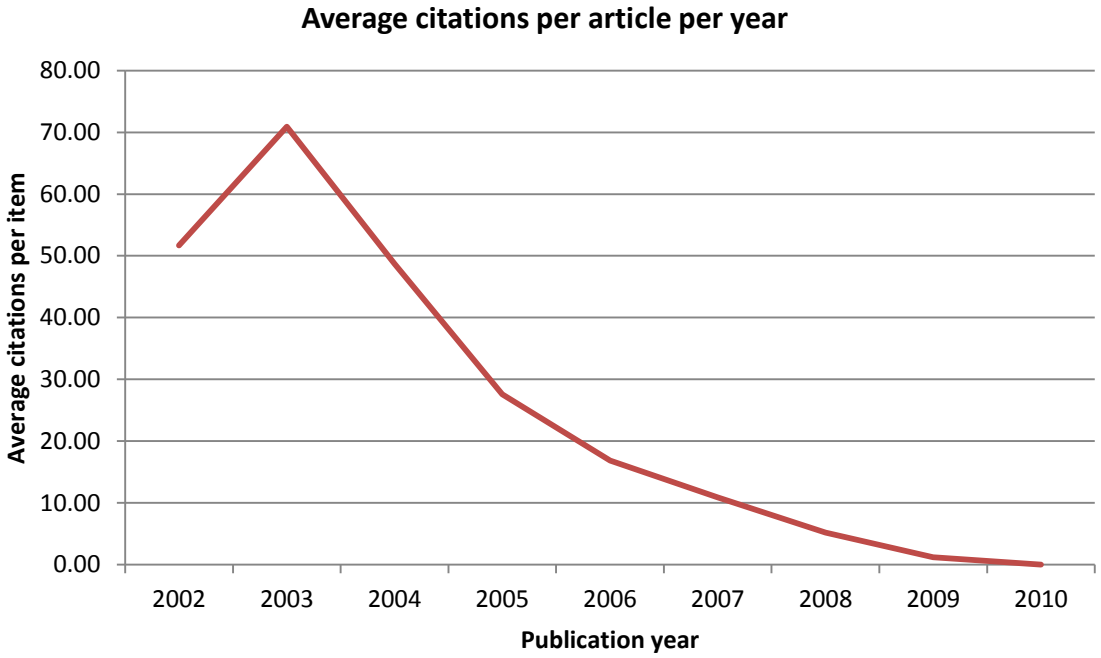
Total citation rates indicate that papers published in 2005 account for the largest number of citations in the dataset. A strong peak for citations to papers published in the year 2000 is explained by the presence of the canonical paper 'Gene Ontology: tool for the unification of biology' by Ashburner *et al*. which the GO Consortium requires authors to cite in all instances where the Gene Ontology is referenced. Citation rates show an annual decline in citations, as older papers have accumulated more citations (see Figure 15).

Figure 15: Citations to Gene Ontology papers



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 16). 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 16: Average citations to Gene Ontology articles



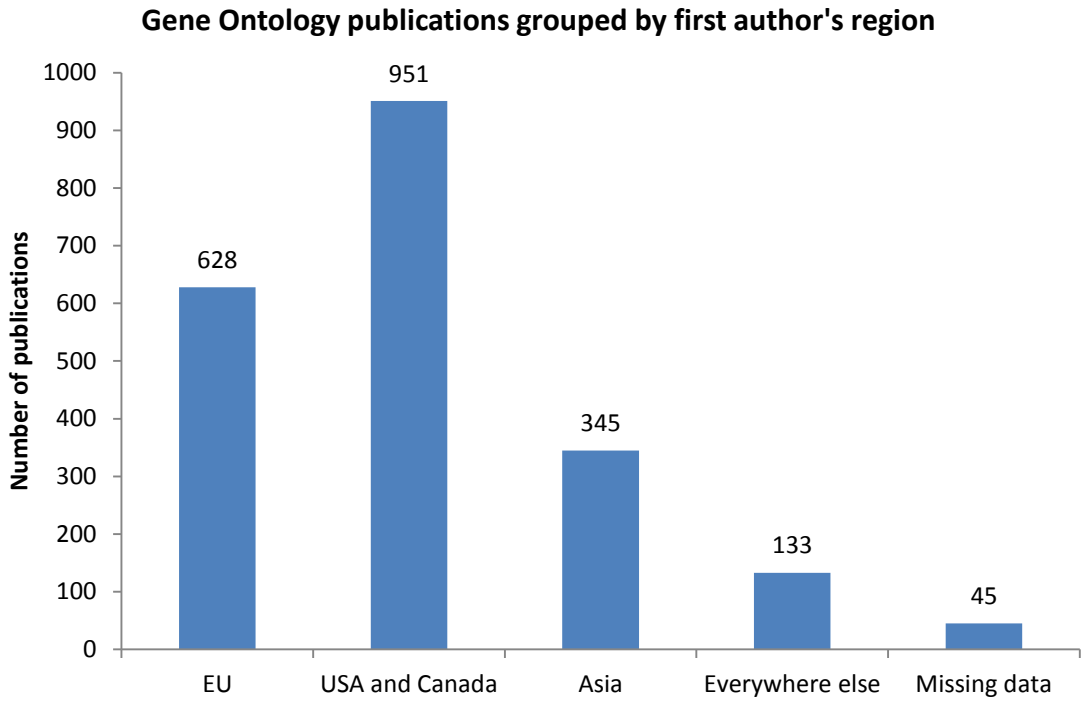

### 6.7.5 Gene Ontology research by country

The United States is the top country for Gene Ontology research publications worldwide, with 882 articles (42% of all articles, see Table 33). The United Kingdom follows with 158 articles (7.5% of all articles), and China in third with 136 articles (6.5% of all articles). After adjusting publication rates for population size, Korea and Canada show disproportionately high Gene Ontology publication rates suggesting a high level of interest in this domain for these nations.

Table 33: Gene Ontology research by country of first author

| Country | Published articles | Articles per million population |
|---|---|---|
| **USA** | 882 | 2.8 |
| **UK** | 158 | 2.5 |
| **China** | 136 | 0.1 |
| **Germany** | 115 | 1.4 |
| **Japan** | 82 | 0.6 |
| **Canada** | 69 | 2.0 |
| **Spain** | 54 | 1.2 |
| **Korea** | 55 | 2.3 |
| **France** | 53 | 0.8 |
| **Italy** | 53 | 0.9 |

Figure 17: Gene Ontology research output by geographic region




### 6.7.6 Journal analysis

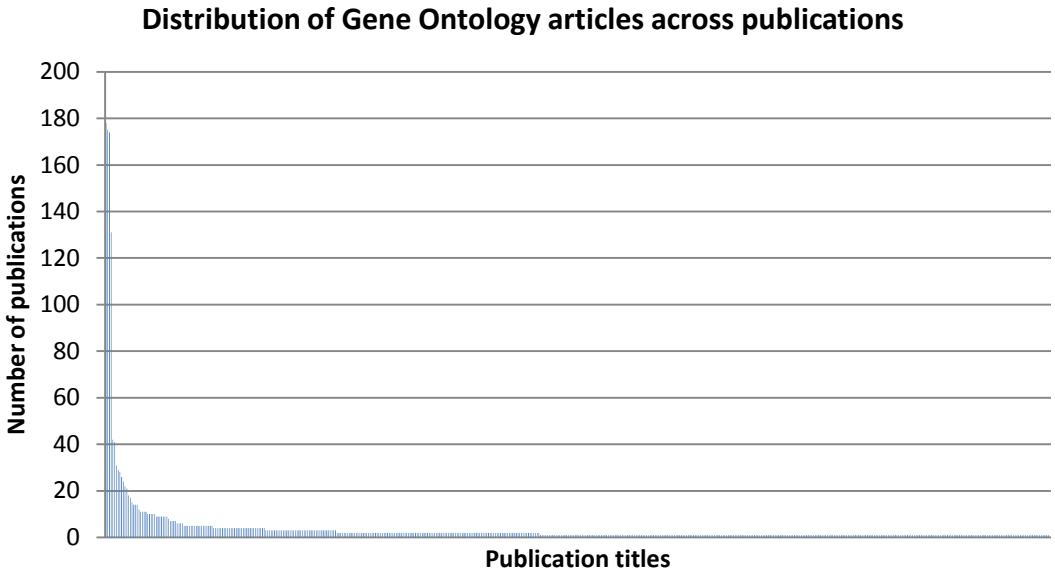
Most Gene Ontology research is published in one of four journals: BMS Bioinformatics, Nucleic Acids Research, Bioinformatics or BMC Genomics (see Table 34 below). These four journals account for 31.4% of all the Gene Ontology articles published before 2010.

Table 34: Top Gene Ontology research journals

| Journal | 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 on Biocomputing** | 29 |
| **Genome Research** | 28 |
| **J. of Proteome Research** | 26 |

Gene ontology 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 18 below).

Figure 18: Pareto distribution for Gene Ontology research across all journals




Journals dedicated to the study and application of ontologies to the management of information first appeared in 2006 with the 'Journal of Applied Ontology', and this illustrates the increasing interest in, and importance of, ontologies generally.

### 6.7.7 GO applications in published papers

The bibliometric data above tells us very little about what authors are writing about when they reference the Gene Ontology. In order to look in more detail at the subject of Gene Ontology papers, the dataset was filtered for the MeSH heading:

*Like "*Oligonucleotide Array Sequence Analysis*"*

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 (see Table 35).

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 Gene Ontology categories. An analysis paper was considered to use the Gene Ontology to categorise and interpret empirical data, usually in what is known as a 'term enrichment analysis' (are specific GO terms enriched in a list of gene products?). A model paper uses the Gene Ontology to create computational models of biological systems. Software papers reported computer applications, normally available to the bioinformatics community, which incorporate Gene Ontology files into the software (these are normally analytical tools for biologists to interpret datasets).

Table 35: Results of GO paper categorisation by type

| Category of GO paper | Results |
|---|---|
| **Analysis** | 522 |
| **Model** | 114 |
| **Software** | 66 |
| **Database** | 18 |

# 8 Glossary

*Please note that the definitions given here are designed to aid the non-expert reader in understanding the context in which I use these terms in my thesis, and are in no sense officially mandated dictionary definitions. Many of the more official biological definitions can be found by searching the Gene Ontology itself at [http://amigo.geneontology.org/](http://amigo.geneontology.org/)*

Base-pairs :- a chemical linkage between two nucleotides on opposing DNA or RNA

Determinism :- philosophical principle that events are necessitated by antecedent conditions and the Laws of Nature (e.g. with light, water, nutrition and an appropriate climate, my tomato plant will grow taller)

DNA :- the biological code (composed of four different chemical 'letters') in which genes are written and from which the structures and processes of life happen

E-science (or e-Research) :- the use of distributed, high-throughput and often collaborative computer systems to process scientific data

Gene :- an element usually made of a long stretch of DNA which encodes a protein or other biological product which does something in a cell (e.g. the CDH1 gene in humans encodes a protein called E-cadherin which helps cells stick to one another)

Gene Ontology :- a controlled vocabulary created by the Gene Ontology Consortium and released in various formats in order to describe biological processes, functions and cell components

Gene Ontology Consortium :- the organisation which created and maintains the Gene Ontology

Genetic code :- a term often used to describe the regularities such as genes and regulatory elements common between species and stored in the millions of chemical residues in DNA

GO paper:- an article mentioning the Gene Ontology, either in full or abbreviated form, in the title, abstract or indexing

Messenger RNA (mRNA) :- when the DNA code is read by the cell machinery, it is converted into an intermediary chemical message called mRNA

Nucleotide :- the chemical units of DNA or RNA; one of either adenine, thymine, guanine or cytosine in DNA

Open Biomedical Ontologies Foundry (or OBO Foundry) :- an online community which has created and shared many ontologies created for the bioscience domain according to specific 'open' principles

Ontology :- a structured representation of the concepts used by a particular domain of users; often used in the informatics and biosciences domain to include what information science would distinguish as controlled vocabularies, classifications and taxonomies

Ontological realism :- the principle that ontologies are representations of reality

Protein :- the chemical structures, composed of amino acids and derived from genes, which do the work of creating life; examples of proteins include hair and hemaglobin

Scientific realism :- the philosophical doctrine that there exists a single objective reality which scientists can observe and infer scientific knowledge from

Sourceforge :- software repository used by the GO Consortium for sharing ontology files

Sourceforge tracker :- a managed online forum system on Sourceforge allowing developers to receive, prioritise and discuss comments from users

Species :- different classes of organisms (e.g. dogs and cacti)

Reductionism :- the principle in science that complex phenomena can be completely explained in terms of simpler 'levels' of explanation (e.g. the movement of my hand can be explained in terms of, or reduced to, explanations at progressively simpler levels such as physiology, cell biology and ultimately the expression and regulation of genes in the cells of my body)

RNA :- the chemical structure into which the genetic code is initially

Transcription :- the conversion of DNA into RNA

Translation :- the conversion of RNA into proteins

# 9 Bibliography


1. The Library of Congress. *Library of Congress Classification Outline*. 2011 [last accessed 04 January 2011]; Available from: http://www.loc.gov/catdir/cpso/lcco/
2. NCBI. *NCBI Taxonomy Browser*. 2011 [last accessed 23 May 2011]; Available from: http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/
3. Finn, R.D., et al., *The Pfam protein families database.* Nucleic Acids Res, 2010. **38**(Database issue): p. D211-22.
4. International Union of Biochemistry and Molecular Biology. Nomenclature Committee. and E.C. Webb, *Enzyme nomenclature 1992 : recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes*. 1992, San Diego: Published for the International Union of Biochemistry and Molecular Biology by Academic Press. xiii, 862 p.
5. Ashburner, M., et al., *Gene Ontology: tool for the unification of biology.* Nature Genetics, 2000. **25**(1): p. 25-29.
6. The Human Phenotype Ontology. *Human Phenotype Ontology Website*. 2010 [last accessed 23 May 2011]; Available from: http://www.human-phenotype-ontology.org/index.php/hpo_browse.html
7. Davidson, D., et al., *The mouse atlas and graphical gene-expression database.* Semin Cell Dev Biol, 1997. **8**(5): p. 509-17.
8. US National Library of Medicine. *Medical Subject Headings*. 2010 [last accessed 04 February 2011]; Available from: http://www.nlm.nih.gov/mesh/
9. Crick, F., *Central dogma of molecular biology.* Nature, 1970. **227**(5258): p. 561-3.
10. Sustar, P., *Crick's notion of genetic information and the 'Central Dogma' of molecular biology.* British Journal of the Philosophy of Science, 2007. **58**(1): p. 13-24.
11. Coleman, W., *Biology in the nineteenth century : problems of form, function, and transformation*. History of Science. 1977, Cambridge ; New York: Cambridge University Press. vii, 187 p.
12. Aristotle, *Vol. 5, De partibus animalium*. The Works of Aristotle, ed. W.D. Ross. 1928-1952.
13. Lennox, J.G., *Aristotle's philosophy of biology : studies in the origins of life science*. Cambridge studies in philosophy and biology. 2001, Cambridge: Cambridge University Press. xxiii, 321 p.
14. Smith, J.E.H., *The problem of animal generation in early modern philosophy*. Cambridge studies in philosophy and biology. 2006, Cambridge: Cambridge University Press. xiii, 456 p.
15. Gotthelf, A. and J.G. Lennox, *Philosophical issues in Aristotle's biology*. 1987, Cambridge Cambridgeshire ; New York: Cambridge University Press. xiii, 462 p.
16. Balme, D.M. and A. Gotthelf, *Aristotle on nature and living things : philosophical and historical studies : presented to David M. Balme on his seventieth birthday*. 1985, Pittsburgh, Pa.: Mathesis Publications. xxix, 410 p.
17. Falcon, A., *Aristotle and the science of nature : unity without uniformity*. 2005, Cambridge, UK ; New York: Cambridge University Press. xvii, 139 p.
18. Dupré, J., *The disorder of things: metaphysical foundations of the disunity of science*. 1993: Harvard University Press.
19. Dupré, J., *Wilkerson on Natural Kinds.* Philosophy, 1989. **64**(248): p. 248-251.
20. Winsor, M.P., *Non-essentialist methods in pre-Darwinian taxonomy.* Biology and Philosophy, 2003. **18**(3): p. 387-400.
21. Winsor, M.P., *The creation of the essentialism story: an exercise in metahistory.* History and Philosophy of the Life Sciences, 2006. **28**(2): p. 149-74.
22. Mayr, E., *Biological classification: towards a synthesis of opposing methodologies.* Science, 1981. **214**: p. 510-516.

23. Mayr, E., *The growth of biological thought : diversity, evolution, and inheritance*. 1982, Cambridge, Mass.: Belknap Press. ix, 974 p.
24. Mayr, E., *One long argument : Charles Darwin and the genesis of modern evolutionary thought*. Questions of science. 1991, Cambridge, Mass.: Harvard University Press. xiv, 195 p.
25. Müller-Wille, S., *Collection and collation: theory and practice of Linnaean botany.* Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 2007. **38**(3): p. 541-562.
26. Creath, R. and J. Maienschein, *Biology and epistemology*. Cambridge Studies in Philosophy and Biology. 2000, Cambridge: Cambridge University Press. xviii, 295 p.
27. Smith, P., *Realism and the progress of science*. Cambridge studies in philosophy. 1981, Cambridge Cambridgeshire ; New York: Cambridge University Press. 135 p.
28. Ereshefsky, M., *Darwin's solution to the species problem.* Synthese.
29. Ereshefsky, M., *The poverty of the Linnaean hierarchy : a philosophical study of biological taxonomy*. 2001, Cambridge ; New York, N.Y.: Cambridge University Press. x, 316 p.
30. Ereshefsky, M., *The poverty of Linnaean hierarchy : a philosophical study of biological taxonomy*. 2000, Cambridge, U.K. ; New York: Cambridge University Press.
31. Ereshefsky, M., S. Sarkar, and A. Plutynski, *Systematics and taxonomy*, in *A companion to the philosophy of biology*. 2008, Blackwell Publishing. p. 99-118.
32. Ayala, F.J. and T. Dobzhansky, *Studies in the philosophy of biology : reduction and related problems: [proceedings of the Conference on Problems of Reduction in Biology, held at Bellagio, Italy, 9-16th September 1972]*. 1974, London: Macmillan.
33. Brigandt, I., *Explanation in Biology: Reduction, Pluralism, and Explanatory Aims.* Science & Education, 2012(Preprint): p. 1-23.
34. Hull, D.L., *Reduction and Genetics.* Journal of Medicine and Philosophy, 1981. **6**(2): p. 125-144.
35. Sarkar, S., *Genetics and reductionism*. 1998: Cambridge University Press.
36. Smith, B., *Mereotopology: A theory of parts and boundaries.* Data & Knowledge Engineering, 1996. **20**(3): p. 287-303.
37. Hull, D.L., *The effect of essentialism on taxonomy: two hundred years of stasis.* The British Journal for the Philosophy of Science, 1965. **XVI**(61): p. 1-18.
38. Hull, D.L. and M. Ruse, *The philosophy of biology*. Oxford readings in philosophy. 1998, Oxford: Oxford University Press. ix, 772 p.
39. Pliny the Elder, *Natural history: a selection*, ed. J.F. Healy. 1991.
40. Appel, T.A., *Cuvier-Geoffroy debate : French biology in the decades before Darwin*. Monographs on the history and philosophy of biology. 1987, New York ; Oxford: Oxford University Press. 305 p., [16] p. of plates.
41. Padian, K., *Charles Darwin's views of classification in theory and practice.* Systematic Biology, 1999. **48**(2): p. 352-364.
42. de Queiroz, K., *Systematics and the Darwinian Revolution.* Philosophy of Science, 1988. **55**(2): p. 238-259.
43. Stamos, D.N., *Darwin and the nature of species*. SUNY series in philosophy and biology. 2007, New York ; Bristol: State University of New York, University Presses Marketing. xix, 273 p.
44. Morange, M., *A history of molecular biology*. 1998, Cambridge, Mass.: Harvard University Press. 336 p.
45. Haldane, J.B.S., *What is life?* 1949, London,: L. Drummond. x, 261 p.
46. Sterelny, K., *The "Genetic Program" program : a commentary on Maynard Smith on information in biology.* Philosophy of Science, 2000. **67**(2): p. 195-201.
47. Stegmann, U.E., *The arbitrariness of the genetic code.* Biology and Philosophy, 2004. **19**(2): p. 205-222.
48. Schneider, T., *Evolution of biological information.* Nucleic Acids Research, 2000. **28**(14): p. 2794-2799.

49. Sarkar, S., *Information in genetics and developmental biology : comments on Maynard Smith.* Philosophy of Science, 2000. **67**(2): p. 208-213.
50. Uniprot Consortium. *Protein nomenclature publication list*. 2011 [last accessed 25 May 2011]; Available from: http://www.expasy.ch/cgi-bin/lists?nomlist.txt
51. Wellcome Trust Sanger Institute. *Pfam: Family: Pkinase (PF00069)*. 2011 [last accessed 22 February 2011]; Available from: http://pfam.sanger.ac.uk/family?acc=PF00069
52. Bateman, A., et al., *The Pfam protein families database.* Nucleic Acids Res, 2000. **28**(1): p. 263-6.
53. Overbeek, R., et al., *WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction.* Nucleic Acids Res, 2000. **28**(1): p. 123-5.
54. Mewes, H.W., et al., *MIPS: a database for genomes and protein sequences.* Nucleic Acids Research, 2002. **30**(1): p. 31-4.
55. Stevens, R., et al., *TAMBIS: transparent access to multiple bioinformatics information sources.* Bioinformatics, 2000. **16**(2): p. 184-5.
56. Harmar, A.J., et al., *IUPHAR-DB: the IUPHAR database of G protein-coupled receptors and ion channels.* Nucleic Acids Research, 2009. **37**(Database issue): p. D680-5.
57. Tatusov, R., et al., *The COG database: an updated version includes eukaryotes.* BMC Bioinformatics, 2003. **4**(1): p. 41.
58. Saier, M.H., Jr., et al., *The Transporter Classification Database: recent advances.* Nucleic Acids Research, 2009. **37**(Database issue): p. D274-8.
59. Riley, M., *Functions of the gene products of Escherichia coli.* Microbiology Reviews, 1993. **57**(4): p. 862-952.
60. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
61. Cimino, J.J. and X. Zhu, *The practical impact of ontologies on biomedical informatics.* Yearbook of Medical Informatics, 2006. **45 Suppl 1**: p. 124-135.
62. Consortium, G., *The Gene Ontology (GO) database and informatics resource.* Nucleic Acids Research, 2004. **32**(suppl_1): p. D258-261.
63. Ashburner, M., et al., *Creating the gene ontology resource: Design and implementation.* Genome Research, 2001. **11**(8): p. 1425-1433.
64. Kohler, J., et al., *Quality control for terms and definitions in ontologies and taxonomies.* BMC Bioinformatics, 2006. **7**.
65. Smith, B., J. Kohler, and A. Kumar, *On the application of formal principles to life science data: a case study in the Gene Ontology*, in *Data Integration in the Life Sciences*. 2004. p. 79-94.
66. Ceusters, W., et al. *Mistakes in medical ontologies: Where do they come from and how can they be detected?* in *Workshop on Ontologies in Medicine*. 2003. Rome, ITALY.
67. Kumar, A. and B. Smith. *The Universal Medical Language System and the Gene Ontology: Some critical reflections*. in *26th Annual German Conference on Artificial Intelligence*. 2003. Hamburg, Germany.
68. Ogren, P.V., et al., *The compositional structure of Gene Ontology terms.* Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 2004: p. 214-225.
69. Gene Ontology Consortium. *Extended GO Ontology Relations*. 2010 [last accessed 22 July 2010]; Available from: http://www.geneontology.org/GO.ontology-ext.relations.shtml
70. Gruber, T., *A translation approach to portable ontology specifications.* Knowledge Acquisition, 1993. **5**(2): p. 199-220.
71. Yu, A., *Methods in biomedical ontology.* Journal of Biomedical Informatics, 2006. **39**(3): p. 252-266.
72. Smith, B. and W. Ceusters, *Towards industrial strength philosophy: how analytical ontology can help medical informatics.* Interdisciplinary Science Reviews, 2003. **28**(2): p. 106-111.
73. Guarino, N. *Formal Ontology and Information Systems*. in *Proceedings of FOIS'98*. 1998. Trento, Italy: IOS Press.

74. Smith, B., *Beyond concepts: ontology as reality representation*, in *Proceedings of the Third International Conference (FOIS 2004)*. 2004, IOS Press: Amsterdam. p. 73-84.
75. Smith, B., *From concepts to clinical reality: An essay on the benchmarking of biomedical terminologies.* Journal of Biomedical Informatics, 2006. **39**(3): p. 288-298.
76. Smith, B. and A. Kumar, *Controlled vocabularies in bioinformatics: a case study in the gene ontology.* Drug Discovery Today: BIOSILICO, 2004. **2**(6): p. 246-252.
77. Smith, B., *Formal ontology, common sense and cognitive science.* International Journal of Human-Computer Studies, 1995. **43**(5-6): p. 641-667.
78. Smith, B., et al., *Relations in biomedical ontologies.* Genome biology, 2005. **6**(5).
79. Smith, B. and C. Rosse. *The role of foundational relations in the alignment of biomedical ontologies*. in *11th World Congress on Medical Informatics*. 2004. San Francisco, CA.
80. Smith, B. and A.C. Varzi, *Fiat and bona fide boundaries.* Philosophy and Phenomenological Research, 2000. **60**(2): p. 401-420.
81. Quine, W.V., *From a logical point of view : 9 logico-philosophical essays*. 2d ed. 1980, Cambridge, Mass.: Harvard University Press. xii, 184 p.
82. Witte, R., T. Kappler, and C.J.O. Baker, *Ontology design for biomedical text mining*, in *Semantic web : revolutionizing knowledge discovery in the life sciences*, C.J.O.C.H.-H. Baker, Editor. 2007, Springer. p. 281-313.
83. Ingwersen, P. and K. Järvelin, *The turn : integration of information seeking and retrieval in context*. 2005: Springer.
84. Cronin, B., *The sociological turn in information science.* Journal of Information Science, 2008. **34**(4): p. 465-475.
85. OBO Foundry. *The Open Biological and Biomedical Ontologies Foundry*. 2011 [last accessed 19 January 2011]; Available from: http://www.obofoundry.org/
86. OBO Foundry. *OBO Foundry Principles*. 2006 [last accessed 13 May 2011]; Available from: http://www.obofoundry.org/crit.shtml
87. OBO Foundry. *OBO Foundry Principles*. 2011 [last accessed 13 May 2011]; Available from: http://www.obofoundry.org/wiki/index.php/OBO_Foundry_Principles
88. Dumontier, M. and R. Hoehndorf, *Realism for scientific ontologies.* Frontiers in Artificial Intelligence and Applications, 2010. **209**: p. 387-399.
89. Merrill, G.H., *Ontological realism: Methodology or misdirection?* Applied Ontology, 2010. **5**(2): p. 79-108.
90. Lord, P. and R. Stevens, *Adding a Little Reality to Building Ontologies for Biology.* PLoS ONE, 2010. **5**(9): p. e12258.
91. Smith, B. and W. Ceusters, *Ontological realism: A methodology for coordinated evolution of scientific ontologies.* Applied Ontology, 2010. **5**(3): p. 139-188.
92. Hill, D., et al., *Gene Ontology annotations: what they mean and where they come from.* BMC Bioinformatics, 2008. **9**(5).
93. Rhee, S.Y., et al., *Use and misuse of the gene ontology annotations.* Nature Reviews Genetics, 2008. **9**(7): p. 509-15.
94. Barrell, D., et al., *The GOA database in 2009-an integrated Gene Ontology Annotation resource.* Nucleic Acids Research, 2009. **37**(Database issue): p. D396-403.
95. Rogers, M. and A. Ben-Hur, *The use of gene ontology evidence codes in preventing classifier assessment bias.* Bioinformatics, 2009. **25**(9): p. 1173-1177.
96. Gene Ontology Consortium. *Current annotations*. 2004 [last accessed 28 April 2012]; Available from: http://www.geneontology.org/GO.downloads.annotations.shtml
97. Al-Shahrour, F., R. Diaz-Uriarte, and J. Dopazo, *FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes.* Bioinformatics, 2004. **20**(4): p. 578-580.
98. Bada, M. and L. Hunter, *Using the Gene Ontology to annotate biomedical journal articles.* Nature Precedings, 2009.

99. Blake, J. and C. Bult, *Beyond the data deluge: data integration and bio-ontologies.* Journal of Biomedical Informatics, 2006. **39**(3): p. 314-320.
100. Bodenreider, O., *Biomedical ontologies in action: role in knowledge management, data integration and decision support.* Yearbook of Medical Informatics, 2008: p. 67-79.
101. Doniger, S.W., et al., *MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data.* Genome biology, 2003. **4**(1).
102. Harel, A., et al., *GIFtS: annotation landscape analysis with GeneCards.* BMC Bioinformatics, 2009. **10**(1): p. 348.
103. Jenssen, T.K., et al., *A literature network of human genes for high-throughput analysis of gene expression.* Nature Genetics, 2001. **28**(1): p. 21-+.
104. King, O., et al., *Predicting gene function from patterns of annotation.* Genome Research, 2003. **13**: p. 896-904.
105. Seki, K. and J. Mostafa, *Gene ontology annotation as text categorization: An empirical study.* Information Processing and Management, 2008. **44**(5): p. 1754-1770.
106. Smith, B., et al., *The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.* Nature Biotechnology, 2007. **25**(11): p. 1251-1255.
107. Ruttenberg, A., et al., *Advancing translational research with the Semantic Web.* BMC Bioinformatics, 2007. **8**(Suppl 3): p. S2.
108. Coveney, P.V. and M.P. Atkinson, *Crossing boundaries: computational science, e-Science and global e-Infrastructure.* Philosophical Transactions of the Royal Society A: Mathematics, Physics and Engineering Sciences 2009. **367**(1897): p. 2425-2427
109. Dotsika, F., *Uniting formal and informal descriptive power: Reconciling ontologies with folksonomies.* International Journal of Information Management, 2009. **29**(5): p. 407-415.
110. Cui, H., *Competency evaluation of plant character ontologies against domain literature.* Journal of the American Society for Information Science and Technology, 2010. **61**(6): p. 1144-1165.
111. Schober, D., et al., *Survey-based naming conventions for use in OBO Foundry ontology development.* BMC Bioinformatics, 2009. **10**(1): p. 125.
112. Ceusters, W., *Applying evolutionary terminology auditing to the Gene Ontology.* Journal of Biomedical Informatics, 2009. **42**(3): p. 518-529.
113. Cimino, J.J., et al., *The caBIG terminology review process.* Journal of Biomedical Informatics, 2009. **42**(3): p. 571-80.
114. Robinson, L. and M. Karamuftuoglu, *The nature of information science: changing models.* Information Research, 2010. **15**: p. 5-5.
115. Robinson, L., *Information science : communication chain and domain analysis.* Journal of Documentation, 2009. **65**(4): p. 578-591.
116. Hjorland, B. and H. Albrechtsen, *Toward a new horizon in information science: domain analysis.* Journal of the American Society for Information Science, 1995. **46**(6): p. 400-425.
117. Arencibia-Jorge, R., R.L. Vega-Almeida, and Y. Martí-Lahera, *Domain Analysis for the Construction of a Conceptual Structure: A Case Study.* LIBRES: Library & Information Science Research Electronic Journal, 2007. **17**(2): p. 1-26.
118. Jeong, S. and H.-G. Kim, *Intellectual structure of biomedical informatics reflected in scholarly events.* Scientometrics, 2010. **85**(2): p. 541-551.
119. Abbott, R., *Subjectivity as a concern for information science: a Popperian perspective.* Journal of Information Science, 2004. **30**(2): p. 95-106.
120. Frohmann, B., *The role of the scientific paper in scientific information systems*, in *History of information science : proceedings of the 1998 conference on the history and heritage of science information systems*, H.T.B. Bowden Me and R.V. Williams, Editors. 1999, Information Today. p. 63-73.
121. Frohmann, B., *The power of images: a discourse analysis of the cognitive viewpoint.* Journal of Documentation, 1993. **48**(4).

122. Williams, R., *The epistemology of knowledge and the knowledge process cycle: beyond the "objectivist" vs "interpretivist".* Journal of Knowledge Management, 2008. **12**(4): p. 72-85.
123. Ménard, E., S. Mas, and I. Alberts, *Faceted classification for museum artefacts: A methodology to support web site development of large cultural organizations.* Aslib Proceedings, 2010. **62**(4/5): p. 523-532.
124. Chaomei, C., R.J. Paul, and B. O'Keefe, *Fitting the Jigsaw of Citation: Information Visualization in Domain Analysis.* Journal of the American Society for Information Science & Technology, 2001. **52**(4): p. 315-330.
125. Hartel, J., *Managing documents at home for serious leisure: a case study of the hobby of gourmet cooking.* 2010. **66**(6): p. 847-874.
126. Zins, C. and D. Guttmann, *Domain Analysis of Social Work: An Example of an Integrated Methodological Approach.* Knowledge Organization, 2003. **30**(3/4): p. 196-212.
127. Hjorland, B., *Domain analysis in information science: eleven approaches - traditional as well as innovative.* Journal of Documentation, 2002. **58**(4): p. 422-462.
128. Aitchison, J., A. Gilchrist, and D. Bawden, *Thesaurus construction and use : a practical manual*. 4th ed. / Jean Aitchison, Alan Gilchrist, David Bawden. ed. 2000, Chicago: Fitzroy Dearden ; London : Aslib. xiv, 218 p.
129. Hersh, W. and E. Voorhees, *TREC genomics special issue overview.* Information Retrieval, 2009. **12**(1): p. 1-15.
130. Hersh, W., et al. *TREC 2005 genomics track overview*. in *The Fourteeth Text Retrieval Conference (TREC 2005)*. 2005. Gaithersburg, MD.
131. Cohen, A.M. and W.R. Hersh, *The TREC 2004 genomics track categorization task: classifying full text biomedical documents.* Journal of Biomedical Discovery and Collaboration, 2006. **1**: p. 4.
132. McCain, K.W., *The structure of biotechnology research-and-development.* Scientometrics, 1995. **32**(2): p. 153-175.
133. McCain, K.W. and K. Turner, *Citation concetxt analysis and aging patterns of journal articles in molecular genetics.* Scientometrics, 1989. **17**(1-2): p. 127-163.
134. Bornmann, L. and H.D. Daniel, *Selection of research fellowship recipients by committee peer review. Reliability, fairness and predictive validity of Board of Trustees' decisions.* Scientometrics, 2005. **63**(2): p. 297-320.
135. Kuhn, T.S., *The structure of scientific revolutions*. 3rd ed. 1996: University of Chicago Press. 212.
136. Hjorland, B., *Concept theory.* Journal of the American Society for Information Science and Technology, 2009. **60**(8): p. 1519-1536.
137. Cornet, R. and A. Abu-Hanna, *Two DL-based methods for auditing medical terminological systems.* AMIA Annual Symposium Proceedings, 2005: p. 166-70.
138. Cimino, J.J., *Auditing the Unified Medical Language System with semantic methods.* Journal of the American Medical Informatics Association, 1998. **5**(1): p. 41-51.
139. Spackman, K.A., *Rates of change in a large clinical terminology: three years experience with SNOMED Clinical Terms.* AMIA Annual Symposium Proceedings, 2005: p. 714-8.
140. Bodenreider, O., et al., *Investigating subsumption in SNOMED CT: An exploration into large description logic-based biomedical terminologies.* Artificial Intelligence in Medicine, 2007. **39**(3): p. 183-195.
141. Ceusters, W., et al., *Ontology-based error detection in SNOMED-CT.* Studies Health Technology Informatics, 2004. **107**(Pt 1): p. 482-6.
142. Jiang, G. and C.G. Chute, *Auditing the Semantic Completeness of SNOMED CT Using Formal Concept Analysis.* Journal of the American Medical Informatics Association, 2009. **16**(1): p. 89-102.
143. Jiang, G., et al., *Context-based ontology building support in clinical domains using formal concept analysis.* International Journal of Medical Informatics, 2003. **71**(1): p. 71-78.

144. Kwakkel, J.H. and S.W. Cunningham, *Managing polysemy and synonymy in science mapping using the mixtures of factor analyzers model.* Journal of the American Society for Information Science & Technology, 2009. **60**(10): p. 2064-2078.
145. Weller, T., *Information history: its importance, relevance and future.* Aslib Proceedings, 2007. **59**(4): p. 437-448.
146. Weller, T., *Preserving Knowledge Through Popular Victorian Periodicals: An Examination of The Penny Magazine and the Illustrated London News, 1842-1843.* Library History, 2008. **24**(3): p. 200-207.
147. Weller, T.D., *Information in nineteenth century England : exploring contemporary socio-cultural perceptions and understandings*, in *Department of Information Science*. 2007, City University London: London.
148. Black, A., D. Muddiman, and H. Plant, *The early information society : information management in Britain before the computer*. 2007, Aldershot, England ; Burlington, VT: Ashgate. xiii, 288 p.
149. Eisenstein, E.L., *The printing press as an agent of change : communications and cultural transformations in early modern Europe*. 1979, Cambridge Eng. ; New York: Cambridge University Press.
150. Hey, A.J.G. and A.E. Trefethen, *e-Science and its implications.* Philosophical Transactions of the Royal Society, 2003. **361**(1809): p. 1809-1825.
151. Hey, A.J.G. and A.E. Trefethen, *The data deluge: an e-Science perspective*, F. Berman, G.C. Fox, and A.J.G. Hey, Editors. 2003, Wiley and Sons. p. 809-824.
152. Spink, A. and J. Currier, *Towards an evolutionary perspective for human information behavior: An exploratory study.* Journal of Documentation, 2006. **62**(2): p. 171-193.
153. Chen, H., Y. Wang, and Z. Wu, *Introduction to semantic e-Science in biomedicine.* BMC Bioinformatics, 2007. **8**(Suppl 3).
154. Fry, J., R. Schroeder, and M. den Besten, *Open science in e-science: contingency or policy.* Journal of Documentation, 2009. **65**(1): p. 6-32.
155. Hine, C.M., *New infrastructures for knowledge production: understanding E-science*, ed. C.M. Hine. 2006: Information Science Publishing.
156. Hine, C.M., *Computerization movements and scientific disciplines : the reflexive potential of new technologies*, in *New infrastructures for knowledge production : understanding E-science*, C.M. Hine, Editor. 2006, Information Science Publishing. p. 26-47.
157. Joint, N., *Data preservation, the new science and the practitioner librarian.* Library Review, 2007. **56**(6): p. 451-456.
158. Kidd, J., et al., *Mapping and sequencing of structural variation from eight human genomes.* Nature, 2008. **453**(7191): p. 56-64.
159. Mullins, J., *Bringing Librarianship to E-Science.* College and Research Libraries, 2009. **70**(3): p. 212-214.
160. Pitt-Francis, J., A. Garny, and D. Gavaghan, *Enabling computer models of the heart for high-performance computers and the grid.* Philosophical transactions. Series A, Mathematical, physical, and engineering sciences 2006. **364**(1843): p. 1501-16.
161. Rubin, D.L., et al., *National Center for Biomedical Ontology: Advancing biomedicine through structured organization of scientific knowledge.* Omics-a Journal of Integrative Biology, 2006. **10**(2): p. 185-198.
162. Woolgar, S. and C. Coopmans, *Virtual witnessing in a virtual age: a prospectus for social studies of E-science*, in *New infrastructures for knowledge production : understanding E-science*, C.M. Hine, Editor. 2006, Information Science Publishing. p. 1-25.
163. Wouters, P. and A. Beaulieu, *Imagining E-science beyond computation*, in *New infrastructures for knowledge production : understanding E-science*, C.M. Hine, Editor. 2006, Information Science Publishing. p. 48-70.

164. Bacon, F., P. Urbach, and J. Gibson, *The novum organum ; with, The great instauration*. Paul Carus student editions v. 3. 1993, Chicago: Open Court. 334p.
165. Chubin, D., *Open science and closed science : tradeoffs in a democracy.* Science, Technology, & Human Values, 1985. **10**(2): p. 73-81.
166. Murray-Rust, P., *Open Data in Science.* Serials Review, 2008. **34**(1): p. 52-64.
167. Fairclough, N., *Critical discourse analysis as a method in social scientific research*, in *Methods of Critical Discourse Analysis*, R. Wodak and M. Meyer, Editors. 2001, Sage: London. p. 121-138.
168. Gruber, H., *Analyzing communication in the new media*, in *Qualitative discourse analysis in the social sciences*, R. Wodak and M. Krzyzanowski, Editors. 2008, Palgrave Macmillan.
169. Gene Ontology. *Gene Ontology mailing lists*. 2004 [last accessed 25 April 2012]; Available from: http://fafner.stanford.edu/pipermail/go/
170. Chiluwa, I., *The discourse of digital deceptions and '419' emails.* Discourse Studies, 2009. **11**(6): p. 635-660.
171. Park, J.-R., *Interpersonal and Affective Communication in Synchronous Online Discourse.* The Library Quarterly, 2007. **77**(2): p. 133-155.
172. Herring, S.C., *Linguistic and critical research on computer-mediated communication: Some ethical and scholarly considerations.* The Information Society, 1996. **12**(2): p. 153-168.
173. Fairclough, N., *Language and power*. Language in social life series. 1989, London ; New York: Longman. xii, 259 p.
174. Fowler, R., *Power*, in *Handbook of Discourse Analysis*, T.A. Van Dijk, Editor. 1985, Academic Press: London.
175. Neuendorf, K.A., *The content analysis guidebook*. 2002, Thousand Oaks, Calif.: Sage Publications. xviii, 301 p.
176. Jin, Q., *Is FAST the Right Direction for a New System of Subject Cataloging and Metadata?* Cataloging & Classification Quarterly, 2008. **45**(3): p. 91-110.
177. Spiteri, L.F., *Incorporating Facets into Social Tagging Applications: An Analysis of Current Trends.* Cataloging & Classification Quarterly, 2010. **48**(1): p. 94-109.
178. Leroy, G., et al., *A balanced approach to health information evaluation: A vocabulary-based naive Bayes classifier and readability formulas.* Journal of the American Society for Information Science & Technology, 2008. **59**(9): p. 1409-1419.
179. Deokattey, S., A. Neelameghan, and V. Kumar, *A Method for Developing a Domain Ontology: A Case Study for a Multidisciplinary Subject.* Knowledge Organization, 2010. **37**(3): p. 173-184.
180. Nisonger, T.E., *Use of the Checklist Method for Content Evaluation of Full-text Databases: An Investigation of Two Databases Based on Citations from Two Journals.* Library Resources & Technical Services, 2008. **52**(1): p. 4-17.
181. Gene Ontology Consortium. *Gene Ontology CVS Repository*. 2004 [last accessed 28 April 2012]; Available from: http://cvsweb.geneontology.org/cgi-bin/cvsweb.cgi/go/
182. Gene Ontology Consortium. *Gene Ontology FTP Archive*. 1999 [last accessed 28 April 2012]; Available from: ftp://ftp.geneontology.org/pub/go/
183. Gene Ontology Consortium. *Wayback Machine internet archives for www.geneontology.org*. 2000 [last accessed 28 April 2012]; Available from: http://wayback.archive.org/web/*/http://www.geneontology.org
184. Gene Ontology Consortium. *Sourceforge development trackers for the Gene Ontology*. 2002 [last accessed 28 April 2012]; Available from: http://sourceforge.net/tracker/?group_id=36855
185. Chen, W.H., et al., *Generating ontologies with basic level concepts from folksonomies*, in *Proceedings of ICCS 2010 - International Conference on Computational Science*. 2010. p. 573-581.

186. Smith, B. *Basic concepts of formal ontology*. in *1st Conference on Formal Ontology in Information Systems (FOIS '98) / 6th International Conference on Principles of Knowledge Representation and Reasoning (KR '98)*. 1998. Trent, Italy.
187. Smith, B. *Formal ontology, common sense and cognitive science*. in *International Workshop on Formal Ontology in Conceptual Analysis and Knowledge Representation*. 1993. Padua, Italy.
188. Goodwin, W., *Structure, function, and protein taxonomy.* Biology and Philosophy, 2011. **26**(4): p. 533-545.
189. Gene Ontology Cardiovascular Group. *Minutes Heart Development workshop*. 2009 22 September 2009 [last accessed 03 January 2012]; Available from: http://wiki.geneontology.org/index.php/Minutes_Heart_Development_workshop
190. Vickery, B.C., *Faceted classification : a guide to construction and use of special schemes*. 1960: Aslib. 70p.,23cm.
191. La Barre, K., *Facet analysis.* Annual Review of Information Science and Technology, 2010. **44**(1): p. 243-284.
192. National Center for Biomedical Ontology. *Welcome to the NCBO Bioportal*. 2011 [last accessed 03 January 2012]; BioPortal provides access to commonly used biomedical ontologies and to tools for working with them.]. Available from: http://bioportal.bioontology.org/
193. Berne, R.M. and M.N. Levy, *Cardiovascular physiology*. 8th ed. The Mosby Physiology Monograph Series. 2001, St. Louis, MO: Mosby. xiv, 312 p.
194. Levick, J.R., *An introduction to cardiovascular physiology*. 3rd ed. 2000, London: Arnold. ix, 433 p.
195. Ozaki, Y., et al., *Effect of Direct Renin Inhibitor, Aliskiren, on Peripheral Blood Monocyte Subsets and Myocardial Salvage in Patients With Primary Acute Myocardial Infarction.* Circulation Journal, 2012.
196. Arsalan, M., et al., *Distribution of cardiac stem cells in the human heart.* ISRN Cardiol, 2012. **2012**: p. 483407.
197. van den Ameele, J., et al., *Eomesodermin induces Mesp1 expression and cardiac differentiation from embryonic stem cells in the absence of Activin.* EMBO Reports, 2012. **13**(4): p. 355-362.
198. Laflamme, M.A. and C.E. Murry, *Regenerating the heart.* Nature Biotechnology, 2005. **23**(7): p. 845-856.
199. Boheler, K.R., et al., *Differentiation of pluripotent embryonic stem cells into cardiomyocytes.* Circulation Research, 2002. **91**(3): p. 189-201.
200. Chien, K.R., et al., *Regulation of cardiac gene-expression during myocardial growth and hypertrophy - molecular studies of an adaptive physiological-response.* FASEB Journal, 1991. **5**(15): p. 3037-3046.
201. Kehat, I., et al., *Human embryonic stem cells can differentiate into myocytes with structural and functional properties of cardiomyocytes.* Journal of Clinical Investigation, 2001. **108**(3): p. 407-414.
202. Smart, N., et al., *De novo cardiomyocytes from within the activated adult heart after injury.* Nature, 2011. **advance online publication**.
203. Ayala, F.J., *Teleological explanations*, in *Philosophy of biology*, M. Ruse, Editor. 1998, Prometheus Books. p. 187-197.
204. Ayala, F.J., *Teleological explanations in evolutionary biology.* Philosophy of Science, 1970. **37**(1): p. 1-15.
205. Bedau, M., *Where's the good in teleology?* Philosophy and Phenomenological Research, 1992. **52**(4): p. 781-806.
206. Brandon, R., *Biological teleology: Questions and explanations.* Studies in History and Philosophy of Science Part A, 1981. **12**(2): p. 91-105.

207. Nagel, E., *Teleology revisited and other essays in the philosophy and history of science*. The John Dewey essays in philosophy. 1979, New York: Columbia University Press. viii, 352 p.
208. Perlman, M., *The modern philosophical resurrection of teleology*, in *Philosophy of biology : an anthology*, A. Rosenberg and R. Arp, Editors. 2009, Wiley-Blackwell: Chichester, U.K. ; Malden, MA. p. 149-163.
209. Shea, N., *Representation in the genome and in other inheritance systems.* Biology and Philosophy, 2007. **22**(3): p. 313-331.
210. Chintapalli, V.R., J. Wang, and J.A.T. Dow, *Using FlyAtlas to identify better Drosophila melanogaster models of human disease.* Nature Genetics, 2007. **39**(6): p. 715-720.
211. Lewis, B.P., et al., *Prediction of mammalian microRNA targets.* Cell, 2003. **115**(7): p. 787-798.
212. Takeda, K., T. Kaisho, and S. Akira, *Toll-like receptors.* Annual Review of Immunology, 2003. **21**: p. 335-376.
213. Hannon, G.J., *RNA interference.* Nature, 2002. **418**(6894): p. 244-251.
214. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nature Genetics, 2000. **25**(1): p. 25-9.
215. Star, S.L. and J.R. Griesemer, *Institutional Ecology, `Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39.* Social Studies of Science, 1989. **19**(3): p. 387-420.
216. Star, S.L., *This is Not a Boundary Object: Reflections on the Origin of a Concept.* Science, Technology, & Human Values, 2010. **35**(5): p. 601-617.
217. Gene Ontology Consortium. *Gene Ontology Documentation*. 2011 [last accessed 09 May 2011]; Available from: http://www.geneontology.org/GO.contents.doc.shtml
218. Mungall, C. *OBO Flat File Format 1.4 Syntax and Semantics [WORKING DRAFT]*. 2011 [last accessed 2012]; Available from: http://oboformat.googlecode.com/svn/branches/2011-11-29/doc/obo-syntax.html
219. Smith, B., *Basic concepts of formal ontology*, in *Formal ontology in information systems*, N. Guarino, Editor. 1998, IOS Press. p. 19-28.
220. Soldatova, L.N. and R.D. King, *Are the current ontologies in biology good ontologies?* Nature Biotechnology, 2005. **23**(9): p. 1095-1098.
221. Simon, J., et al. *Formal ontology for natural language processing and the integration of biomedical databases*. in *Symposium on Electronic Health Record Healthcare Registers and Telemedicine*. 2004. Prague, CZECH REPUBLIC.
222. Smith, B., et al., *Relations in biomedical ontologies.* Genome Biology, 2005. **6**(5): p. R46.
223. Ceusters, W., B. Smith, and L. Goldberg. *A terminological and ontological analysis of the NCI thesaurus*. in *49th Annual Conference of the German-Society-for-Medical-Informatics-Biometry-and-Epidemiology*. 2004. Innsbruck, AUSTRIA.
224. National Information Standards Organization (U.S.) and American National Standards Institute., *Guidelines for the construction, format, and management of monolingual controlled vocabularies : an American national standard*. National information standards series,. 2006, Bethesda, Md.: National Information Standards Organization.
225. Gene Ontology Consortium. *List of terms*. 2007 [last accessed 03 November 2011]; "This is the wikified version of go/scratch/sensutermlist.txt"]. Available from: http://wiki.geneontology.org/index.php/List_of_terms
226. Bard, J., S.Y. Rhee, and M. Ashburner, *An ontology for cell types.* Genome Biology, 2005. **6**(2): p. R21.
227. Gene Ontology Consortium. *Cell Ontology Progress Report 2010*. 2010 20 December 2010 [last accessed November 2011]; Work on the Cell Ontology is funded by an ARRA Competitive Revision to the GO Consortium Grant. This work commenced on September 30, 2009. The project is based at MGI, with sub-contract work at LBNL. ]. Available from: http://wiki.geneontology.org/index.php/Cell_Ontology_Progress_Report_2010

228. Diehl, A.D., et al., *Hematopoietic cell types: prototype for a revised cell ontology.* Journal of Biomedical Informatics, 2011. **44**(1): p. 75-9.
229. Smith, A.D., *Oxford dictionary of biochemistry and molecular biology*. Rev. ed. ed. 2000, Oxford: Oxford University Press. xi,738p.
230. Barrell, D., et al., *The GOA database in 2009--an integrated Gene Ontology Annotation resource.* Nucleic Acids Res, 2009. **37**(Database issue): p. D396-403.
231. Gene Ontology Consortium. *Mapping of GO terms to UniProt Knowledgebase keywords.* 2011 [last accessed 03 March 2011]; This mapping is generated by the UniProtKB and UniProtKB-GOA teams.]. Available from: http://www.geneontology.org/external2go/spkw2go
232. Gene Ontology Consortium. *Ontology Content Documentation*. 2010 [last accessed 01 December 2010]; Available from: http://geneontology.org/GO.contents.ont-cont.shtml
233. Budd, J., *Discourse Analysis and the Study of Communication in LIS.* Library Trends, 2006. **55**(1): p. 65-82.
234. Fairclough, N., *Technologisation of discourse*, in *Texts and practices : readings in critical discourse analysis*, C.R. Caldas-Coulthard and M. Coulthard, Editors. 1996, Routledge: London ; New York. p. 294.
235. Wetherell, M., et al., *Discourse theory and practice : a reader*. 2001, London ; Thousand Oaks, Calif.: SAGE. ix, 406 p.
236. Joos, M., *The five clocks*. 1967: Harcourt, Brace & World.
237. Van Dijk, T.A., *Discourse, power and access*, in *Texts and practices : readings in critical discourse analysis*, C.R. Caldas-Coulthard and M. Coulthard, Editors. 1996, Routledge: London ; New York. p. xii, 294 p.
238. Haider, J. and D. Bawden, *Pairing information with poverty: traces of development discourse in LIS.* New Library World, 2006. **107**(9): p. 371-385.
239. Haider, J. and B. David, *Conceptions of "information poverty" in LIS: a discourse analysis.* Journal of Documentation, 2007. **63**(4): p. 534-557.
240. Talja, S., *Analyzing Qualitative Interview Data: The Discourse Analytic Method.* Library & Information Science Research, 1999. **21**(4): p. 459-477.
241. Sundin, O., *Janitors of knowledge: constructing knowledge in the everyday life of Wikipedia editors.* Journal of Documentation, 2011. **67**(5): p. 840-862.
242. Matei, S.A. and C. Dobrescu, *Wikipedia's 'Neutral Point of View': settling conflict through ambiguity.* Information Society, 2011. **27**(1): p. 40-51.
243. Luyt, B. and D. Tan, *Improving Wikipedia's credibility: references and citations in a sample of history articles.* Journal of the American Society for Information Science & Technology, 2010. **61**(4): p. 715-722.
244. Foucault, M. and C. Gordon, *Power/knowledge : selected interviews and other writings, 1972-1977*. 1980, Brighton, Sussex: Harvester Press. xi, 270 p.
245. Cook, G., E. Pieri, and P.T. Robbins, *'The Scientists Think and the Public Feels': Expert Perceptions of the Discourse of GM Food.* Discourse & Society, 2004. **15**(4): p. 433-449.
246. Bowker, G.C. and S.L. Star, *Sorting things out: classification and its consequences*. 1999: MIT Press.
247. Gene Ontology Consortium. *Curator Guide: Obsoletion*. 2011 [last accessed 25 January 2012]; Available from: http://wiki.geneontology.org/index.php/Curator_Guide:_Obsoletion
248. Schober, D., et al., *Survey-based naming conventions for use in OBO Foundry ontology development.* BMC Bioinformatics, 2009. **10**.
249. Reeves, G.A., et al., *The Protein Feature Ontology: a tool for the unification of protein feature annotations.* Bioinformatics, 2008. **24**(23): p. 2767-72.
250. Natale, D.A., et al. *Framework for a Protein Ontology*. in *1st International Workshop on Text Mining in Bioinformatics*. 2006. Arlington, VA.

251. Eilbeck, K., et al., *The Sequence Ontology: a tool for the unification of genome annotations.* Genome Biology, 2005. **6**(5): p. R44.
252. GO Consortium. *October 2004 consortium meeting minutes in plain text*. 2004 [last accessed 01 January 2012]; Available from: http://www.geneontology.org/minutes/20041015_Chicago.txt
253. Ceusters, W. and W. Ceusters, *Applying evolutionary terminology auditing to the Gene Ontology.* Journal of Biomedical Informatics, 2009. **42**(3): p. 518-29.
254. Gene Ontology Consortium. *Proteases*. 2008 [last accessed 09 January 2012]; This item grew out of work on adding terms to the function ontology for enzyme activities, based on EC entries that don't have corresponding GO terms. ]. Available from: http://wiki.geneontology.org/index.php/Proteases
255. Latour, B., *Science in action: how to follow scientists and engineers through society*. 1987: Harvard University Press.
256. U.S. National Library of Medicine. *XML MeSH Data Elements*. 2011 [last accessed 03 February 2012]; Available from: http://www.nlm.nih.gov/mesh/xml_data_elements.html
257. Library of Congress, *Library of Congress to cancel LCGFT Character- and Franchise-Based Terms for moving images*, in *Cataloguing and Policy Support Office Bulletins*. 2011, Library of Congress.
258. Knowlton, S.A., *Three Decades Since Prejudices and Antipathies: A Study of Changes in the Library of Congress Subject Headings.* Cataloging & Classification Quarterly, 2005. **40**(2): p. 123-145.
259. Popper, K.R., *Conjectures and refutations : the growth of scientific knowledge*. 5th ed. 1989, London ; New York: Routledge. xiii, 431 p.
260. Gene Ontology Consortium. *GO Redistribution and Citation Policy*. c2011 [last accessed 06 February 2012]; Available from: http://www.geneontology.org/GO.cite.shtml
261. Stapleton, A.R. and V.T. Chan, *Subtoxic chlorpyrifos treatment resulted in differential expression of genes implicated in neurological functions and development.* Archives of Toxiocology, 2009. **83**(4): p. 319-33.
262. The Arabidopsis Information Resource (TAIR). *TAIR - Home Page*. 2012 [last accessed 07 February 2012]; Available from: http://www.arabidopsis.org/index.jsp
263. Dennis, G., Jr., et al., *DAVID: Database for Annotation, Visualization, and Integrated Discovery.* Genome Biology, 2003. **4**(5): p. P3.
264. Bawden, D. and L. Robinson, *Introduction to information science*. 2012: Facet Publishing.
265. de Keyser, P., *Indexing: from thesauri to the Semantic Web* 2012, Oxford: Chandos Publishing.
266. Monreal, C.S. and I. Gil-Leiva, *Evaluation of controlled vocabularies by inter-indexer consistency.* Information Research, 2011. **16**(4).
267. Gao, Q., *An empirical study of tagging for personal information organization: performance, workload, memory, and consistency.* International Journal of Human-Computer Interaction, 2011. **27**(9): p. 821-863.
268. Hughes, A.V. and P. Rafferty, *Inter-indexer consistency in graphic materials indexing at the National Library of Wales Inter-indexer consistency in graphic materials indexing at the National Library of Wales.* Journal of Documentation, 2011. **67**(1): p. 9-32.
269. Knautz, K. and W.G. Stock, *Collective indexing of emotions in videos.* Journal of Documentation, 2011. **67**(6): p. 975-994.
270. Wright, L., *Functions.* Philosophical Review, 1973. **82**(2): p. 139-168.
271. Rudwick, M.J.S., *The inference of function from structure in fossils.* British Journal for the Philosophy of Science, 1964. **15**(57): p. 27-40.
272. Cummins, R., *Functional analysis.* The Journal of Philosophy, 1975. **72**(20): p. 741-765.
273. Bigelow, J. and R. Pargetter, *Functions.* Journal of Philosophy, 1987. **84**(4): p. 181-196.
274. Millikan, R., *In defense of proper functions.* Philosophy of Science, 1989. **56**(2): p. 288-302.

275. Schindler, S., *Model, Theory, and Evidence in the Discovery of the DNA Structure.* Br J Philos Sci, 2008. **59**(4): p. 619-658.
276. Neander, K., *Functions as selected effects: the conceptual analyst's defense.* Philosophy of Science, 1991. **58**(2): p. 168-184.
277. Amundson, R. and G.V. Lauder, *Function without purpose.* Biology and Philosophy, 1994. **9**(4).
278. Enc, B. and F. Adams, *Functions and goal directedness.* Philosophy of Science, 1992. **59**(4): p. 635-654.
279. Gene Ontology Consortium. *Molecular Function Ontology Guidelines*. 2011 [last accessed 02 February 2012]; Available from: http://www.geneontology.org/GO.function.guidelines.shtml
280. Graeven, U., et al., *Melanoma-associated expression of vascular endothelial growth factor and its receptors FLT-1 and KDR.* Journal of Cancer Research and Clinical Oncology, 1999. **125**(11): p. 621-9.
281. Mai, J.-E., *Classification in Context: Relativity, Reality, and Representation.* Knowledge Organization, 2004. **31**(1): p. 39-48.
282. Mai, J.-E., *The Modernity of Classification.* Journal of Documentation, 2011. **67**(4): p. 7-7.
283. Andrade, M.A. and P. Bork, *Automated extraction of information in molecular biology.* Febs Letters, 2000. **476**(1-2): p. 12-17.
284. Andronis, C., et al., *Literature mining, ontologies and information visualization for drug repurposing.* Briefings in Bioinformatics, 2011. **12**(4): p. 357-368.
285. Faro, A., D. Giordano, and C. Spampinato, *Combining literature text mining with microarray data: advances for system biology modeling.* Briefings in Bioinformatics, 2012. **13**(1): p. 61-82.
286. Good, B.M., et al., *Mining the Gene Wiki for functional genomic knowledge.* BMC Genomics, 2011. **12**.
287. Hirschman, L., et al., *Accomplishments and challenges in literature data mining for biology.* Bioinformatics, 2002. **18**(12): p. 1553-1561.
288. Hirschman, L., W.S. Hayes, and A. Valencia, *Knowledge acquisition from the biomedical literature*, in *Semantic web : revolutionizing knowledge discovery in the life sciences*, C.J.O.C.H.-H. Baker, Editor. 2007, Springer. p. 53-81.
289. Zweigenbaum, P., et al., *Frontiers of biomedical text mining: current progress.* Briefings in Bioinformatics, 2007. **8**(5): p. 358-375.
290. Golder, S.A. and B.A. Huberman, *Usage patterns of collaborative tagging systems.* Journal of Information Science, 2006. **32**(2): p. 198-208.
291. Noruzi, A., *Folksonomies: (Un) Controlled Vocabulary?* Knowledge Organization, 2006. **33**(4): p. 199-203.
292. Woolwine, D., et al., *Folksonomies, Social Tagging and Scholarly Articles.* Canadian Journal of Information and Library Science, 2011. **35**(1): p. 77-92.
293. Morrison, P.J., *Why Are They Tagging, and Why Do We Want Them To?* Bulletin of the American Society for Information Science & Technology, 2007. **34**(1): p. 12-15.
294. Johncocks, B., *Web 2.0 and users' expectations of indexes.* Indexer, 2008. **26**(1): p. 18-24.
295. Rorissa, A., *A comparative study of Flickr tags and index terms in a general image collection.* Journal of the American Society for Information Science & Technology, 2010. **61**(11): p. 2230-2242.
296. Thomas, M., D.M. Caudle, and C.M. Schmitz, *To tag or not to tag?* Library Hi Tech, 2009. **27**(3): p. 411-434.
297. Makani, J. and L. Spiteri, *The Dynamics of Collaborative Tagging: An Analysis of Tag Vocabulary Application in Knowledge Representation, Discovery and Retrieval.* Journal of Information & Knowledge Management, 2010. **9**(2): p. 93-103.

298. Kipp, M.E.I., *Tagging of Biomedical Articles on CiteULike: A Comparison of User, Author and Professional Indexing.* Knowledge Organization, 2011. **38**(3): p. 245-261.
299. Weller, K. *Folksonomies and ontologies: two new players in indexing and knowledge representation*. in *Online Information 2007*. 2007. London.
300. Freddo, A.R. and C.A. Tacla, *INTEGRATING SOCIAL WEB WITH SEMANTIC WEB Ontology Learning and Ontology Evolution from Folksonomies*. Keod 2009: Proceedings of the International Conference on Knowledge Engineering and Ontology Development, ed. J.L.G. Dietz. 2009. 247-253.
301. Rey-Lopez, M., et al., *T-Learning 2.0: A Personalised Hybrid Approach Based on Ontologies and Folksonomies*, in *Computational Intelligence for Technology Enhanced Learning*, F. Xhafa, et al., Editors. 2010. p. 125-142.
302. Broughton, V., *Essential classification*. 2004, New York: Neal-Schuman. 324 p.
303. Broughton, V., *Concepts and Terms in the Faceted Classification: the Case of UDC.* Knowledge Organization, 2010. **37**(4): p. 270-279.
304. Borgman, C.L., *Scholarly communication and bibliometrics*. 1990, Newbury Park ; London: Sage Publications. 363 p.
305. Nicholas, D. and M. Ritchie, *Literature and bibliometrics*. 1978, London: Bingley [etc.]. 183p.
306. Garfield, E., *Citation indexing : its theory and application in science, technology, and humanities*. 1979, New York ; Chichester: Wiley. xxi,274p.
307. Garfield, E., *Citation analysis as a tool in journal evaluation.* Science, 1972. **178**(4060): p. 471-479.
308. Garfield, E., *The History and Meaning of the Journal Impact Factor.* JAMA: Journal of the American Medical Association, 2006. **295**(1): p. 90-93.
309. Frandsen, T.F., *Attracted to open access journals: a bibliometric author analysis in the field of biology.* Journal of Documentation, 2009. **65**(1): p. 58-82.
310. Costas, R., et al., *Scaling rules in the science system: Influence of field-specific citation characteristics on the impact of individual researchers.* Journal of the American Society for Information Science & Technology, 2009. **60**(4): p. 740-753.
311. Gupta, B.M., *Growth and obsolescence of literature in theoretical population genetics.* Scientometrics, 1998. **42**(3): p. 335-347.
312. Ma, N. and J.C. Guan, *An exploratory study on collaboration profiles of Chinese publications in Molecular Biology.* Scientometrics, 2005. **65**(3): p. 343-355.
313. Malo, S. and A. Geuna, *Science-technology linkages in an emerging research platform: The case of combinatorial chemistry and biology.* Scientometrics, 2000. **47**(2): p. 303-321.
314. Rey-Rocha, J., B. Garzon-Garcia, and J. Martin-Sempere, *Scientists' performance and consolidation of research teams in Biology and Biomedicine at the Spanish Council for Scientific Research.* Scientometrics, 2006. **69**(2): p. 183-212.