A Short Study on the Success of the Gene Ontology

Michael Bada[1], Robert Stevens[1], Carole Goble[1], Yolanda Gil[2], Michael Ashburner[3], Judith A. Blake[4], J. Michael Cherry[5], Midori Harris[6], Suzanna Lewis[7]

[1] {bada|rstevens|carole}@cs.man.ac.uk; Department of Computer Science, University of Manchester, Kilburn Building, Oxford Road, Manchester M13 9PL, UK
[2] gil@isi.edu; Information Sciences Institute, University of Southern California, 4676 Admiralty Way, Suite 1001, Marina del Rey, CA 90292, USA
[3] ma11@gen.cam.ac.uk; Department of Genetics, Downing Street, Cambridge CB2 3EH, UK
[4] jblake@informatics.jax.org; The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609, USA
[5] cherry@stanford.edu; Department of Genetics, School of Medicine, Stanford University, Stanford CA 94305-5120, USA
[6] midori@ebi.ac.uk; EBI-Hinxton, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK
[7] suzi@fruitfly.org; Berkeley Drosophila Genome Project, Department of Molecular and Cell Biology, University of California-Berkeley, Berkeley, CA 94720-3200, USA

While most ontologies have been used only by the groups who created them and for their initially defined purposes, the Gene Ontology (GO), an evolving structured controlled vocabulary of nearly 16,000 terms in the domain of biological functionality, has been widely used for annotation of biological-database entries and in biomedical research. As a set of learned lessons offered to other ontology developers, we list and briefly discuss the characteristics of GO that we believe are most responsible for its success: community involvement; clear goals; limited scope; simple, intuitive structure; continuous evolution; active curation; and early use.

Keywords: Gene Ontology, ontology development, biological database, annotation

Introduction

As part of the evolution toward a Semantic Web, Web resources will have to be annotated with significant amounts of markup that can be reliably processed by computational agents. Many believe that this markup will largely consist of terms from ontologies, in which conceptualizations of domains have been specified in the form of hierarchically structured sets of uniquely named and defined terms and their interrelationships. To this end, a number of ontologies in a variety of domains have been constructed. However, most of them remain used only by the groups who created them and for their initially defined purposes.

An exception to this trend is the Gene Ontology (GO; http://www.geneontology.org) [2, 3], a structured controlled vocabulary in the domain of biological functionality. GO initially consisted of a few thousand terms describing the genetic workings of three organisms and was constructed for the express purpose of database interoperability; it has since grown to a terminology of nearly 16,000 terms and is becoming a *de facto* standard for describing functional aspects of biological entities in all types of organisms. Furthermore, in addition to (and because of) its wide use as a terminological source for database-entry annotation, GO has been used in a wide variety of biomedical research, including analyses of experimental data [1] and predictions of experimental

results [4]. In light of this success, we have sought to distill GO's noteworthy characteristics, and to draw conclusions from them, for consideration by other groups developing ontologies.

The Gene Ontology

It is clear that organisms across the spectrum of life, to varying degrees, possess large numbers of gene products with similar sequences and roles. Knowledge about a given gene product (*i.e.*, a biologically active molecule that is the deciphered end product of the code stored in a gene) can often be determined experimentally or inferred from its similarity to gene products in other organisms. Research into different biological systems uses different organisms that are chosen because they are amenable to advancing these investigations. For example, the rat is a good model for the study of human heart disease, and the fly is a good model to study cellular differentiation. For each of these model systems, there is a database employing curators who collect and store the body of biological knowledge for that organism. This enormous amount of data can potentially add insight to related molecules found in other organisms. A reliable wet-lab biological experiment performed in one organism can be used to deduce attributes of an analogous (or related) gene product in another organism, thereby reducing the need to reproduce experiments in each individual organism (which would be expensive, time-consuming, and, in many organisms, technically impossible). However, querying these heterogeneous, independent databases in order to draw these inferences is difficult: The different organism database projects may use different terms to refer to the same concept and the same terms to refer to different concepts. Furthermore, these terms are typically not formally linked with each other in any way. GO seeks to reveal these underlying biological functionalities by providing a structured controlled vocabulary that can be used to describe gene products, and shared between biological databases. This facilitates querying for gene products that share biologically meaningful attributes, whether from separate databases or within the same database.

GO was begun as an effort between the FlyBase project (which catalogs information about *Drosophila melanogaster*, a species of fruitfly; http://flybase.bio.indiana.edu/), the Mouse Genome Informatics project (which integrates information about the laboratory mouse; http://www.informatics.jax.org/), and the *Saccharomyces* Genome Database project (which concentrates on the budding yeast *Saccharomyces cerevisiae*; http://www.yeastgenome.org/). Prior to collaboration, each database had an existing controlled vocabulary for internal use in the annotation of gene products. For example, in FlyBase, the precursor to GO was a controlled vocabulary of approximately 1,000 terms detailing functions of *Drosophila* gene products, which were stored in a simple flat list and used to annotate several thousand gene products within FlyBase. In 1998, representatives from the three database groups established the GO Consortium to develop a common vocabulary that would describe biological functions shared by these organisms and, ideally, any function found in any organism.

GO chose the simplest data structure sufficient for the initial task—a directed acyclic graph; the structure of GO is described in more detail below. The terminology of GO is used to annotate gene products with respect to three attributes: the specific molecular functions that these products possess, the higher-level biological processes in which they participate, and the cellular components in which they can be found. GO currently

has been used to annotate more than one million gene products within the various participating databases. A fragment of GO can be seen in Figure 1.

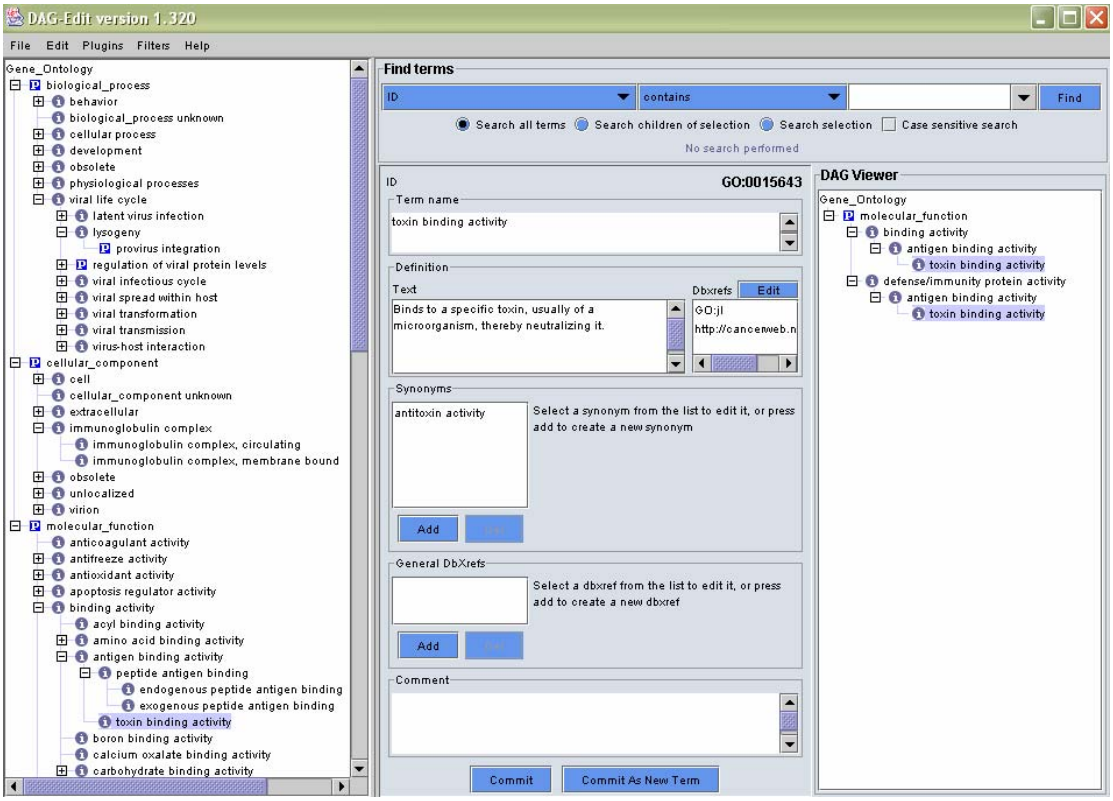

Figure 1. A screenshot of a portion of GO in DAG-Edit (http://www.geneontology.org/GO.Curators.intro.html#dagedit), a tool for editing controlled vocabularies that are represented as directed acyclic graphs. The left pane displays the GO hierarchies, from which "toxin binding activity" has been selected. The lower center and right panes show information for this term, including its natural-language definition, synonym, external-database reference, and all of its ancestor terms. A term related to its parent via an *is-a* relationship is shown with a circled "i" to the left of the term, and a term related to its parent via a *part-of* relationship is shown with a squared "p" to the left of the term.

GO was started as an unfunded project, although all of the founding groups had grant funding for their own databases. AstraZeneca (a major pharmaceutical company) gave a pump-priming grant early in 1999 and, in 2000, the project was awarded a grant by the US National Institutes of Health; this has recently been renewed for a further three years. In addition, there is funding to the Consortium, or its members, for the implementation of GO from the UK Medical Research Council and from the European Commission.

Characteristics of GO

Here we present and briefly discuss the primary characteristics of GO that we believe to be responsible for its widespread use.

1. Community Involvement

One of the factors that account for GO's success is that it originated from within the biological community rather than being created and subsequently imposed by external knowledge engineers. Terms were created by those who had expertise in the domain, thus avoiding the huge effort that would have been required for a computer scientist to

learn and organize large amounts of biological functional information. This also led to general acceptance of the terminology and its organization within the community. This is not to say that there have been no disagreements among biologists over the conceptualization, and there is of course a protocol for arriving at a consensus when there is such a disagreement. However, a model of a domain is more likely to conform to the shared view of a community if the modelers are within or at least consult to a large degree with members of that community.

The Consortium emphasized that GO was not a dictated standard. Rather, it was envisioned that groups would join the project because they understood its value and believed it would be in their interest to commit to the ontology. Participation through submission of terminology for new areas and challenging current representation of functionality was encouraged.

2. Clear Goals

The GO project had a clear goal from the start: to provide a common vocabulary for describing genes products, in terms of the three aforementioned attributes, for the primary purpose of consistently annotating entries in biological databases. (The Consortium also noted that some databases would initially annotate at the level of genes rather than gene products, as many lacked separate database objects to represent the different products that a given gene may encode.) The Consortium also called for tools to query and edit the terminology and to aid users in assigning terms to gene products. They made it clear that annotation with GO terms is not a complete methodology to unify biological databases, although it is necessary a part of this process. Furthermore, the team made a deliberate decision that GO terms and relationships should not attempt to reflect the structure of gene products or indicate whether they are related to one another in an evolutionary sense.

3. Limited Scope

It was decided to limit the scope of GO (at least initially) by defining only those terms needed to describe relatively specific molecular functions, more general biological processes, and cellular components. Obviously, many other areas might have been included, but it was felt that these three categories can be applied to all gene products across all organisms and thus made a good place to start.

4. Simple, Intuitive Structure

Since GO is represented as a directed acyclic graph, it is more complex than a simple hierarchy in that a term can have more than one parent. However, the structure is still relatively simple: Each node in the graph is a natural-language term with a corresponding unique seven-digit numeric identifier and, for the majority of terms, a natural-language definition, while each edge is either an *is-a* or *part-of* relationship. GO is subdivided into three hierarchies corresponding to the three biological attributes within its scope, and each of these hierarchies is constructed from a mixture of taxonomy and partonomy (*i.e.*, each term can be related to each of its parents via *is-a* or *part-of*). This dichotomy has been criticized by many from the computer-science community, as most agree that the hierarchies of formal ontologies should be built exclusively from *is-a* links between terms. GO's taxonomic/partonomic hierarchical structure is not intrinsic to the domain, as the same knowledge could be identically represented as a strict *is-a* hierarchy with additional *part-of* relationships between the

terms. However, the GO Consortium felt that their directed acyclic graph was a representation encoding the desired knowledge that was also relatively clear to the intended audience, and indeed, most biologically oriented users find the hierarchical organization intuitive. In addition, it was agreed that the three hierarchies would be independent of each other (*i.e.*, that there be no links between terms from separate hierarchies) so that a given attribute of a given biological entity could be assigned a term without the risk of implying additional (possibly erroneous) information about the entity. Additional complexity was to be considered in the future.

5. Continuous Evolution

GO is constantly evolving through the addition of new terms and refinement of the existing ontology. Taking this dynamic approach to content was a conscious decision that served dual purposes. First, there was an imperative to provide a practical resource of immediate utility. In addition, biological research (or any research for that matter) is by definition a domain that is continuously expanding, and GO must be able to continuously adapt to refined biological knowledge. Thus, it was unrealistic to wait until GO was "finished"; rather, it approaches "completeness" by being put to use: Deficiencies are noted, and GO is edited to address as many of these deficiencies as possible. Without this inherent iterative evolution, developing the rich, complex model of functional biology in GO would be impossible. The Consortium uses a central CVS repository to manage this ongoing process.

6. Active Curation

While community involvement is essential for refining particular subdomains within GO, it was also realized that there was an equally compelling need to retain a central team of dedicated curators who would provide cohesion and monitor the overall integrity of GO. There are currently four full-time editors who work in the Consortium's editorial office and nearly forty more researchers situated at the sites of the participating databases who also have permission to write to the CVS repository. GO team members alert each other (as well as users) of changes, be they restructurings of subtrees of the ontology or simpler additions of terms, via e-mail and at a site on Sourceforge (http://geneontology.sourceforge.net/). They also meet face to face three or four times a year.

Curators must obviously make changes carefully, especially when altering relationships between child and parent terms, as these could affect the direct and implied semantics of annotations that have been made using the terms. The natural-language term itself can be edited relatively easily, although any change in the term's conceptual content (*i.e.*, a change in its definition) requires a new corresponding unique identifier. Terms may be reclassified as obsolete but are never actually deleted; this is done so that existing annotations do not reference null values during the time lag between the term being made obsolete and the reannotation of the entity. In order to distinguish the lexically identical terms that separate organism communities have historically used with different connotations, the terms are denoted as being "in the sense of" (using the qualifier "*sensu*") to indicate the appropriate context for understanding the term.

7. Early Use

Initial success was documented in the Consortium's paper on the earliest GO implementation (in 2000), where it was stated that the terminology had already been

used to describe over 15,000 gene products in the original three member databases [2]. Specific examples of gene-product annotations with GO terms and explanations of their semantics were also included so that this was not an abstract exercise. Applications of GO in wider contexts were given, including the use of annotations to analyze gene-expression data (the results from an important type of biological experiment). Additional momentum was gained when two more major biological databases (one focusing on the model plant *Arabidopsis thaliana* and the other on the nematode worm *Caenorhabditis elegans*) joined in 2000, confirming the expectation that other organism databases would commit to GO [3]. This acceptance of GO by these major community databases granted it a certain early legitimacy.

Conclusions

We have endeavored to compile a set of lessons learned from the experience of the GO Consortium for the benefit of other groups constructing ontologies. Among those that stand out is the importance of community involvement: The ontology must have the approval of the community for which it is being built, and for that to happen, a representative subset should participate in, if not lead, the initial conceptualization. The tasks for which the ontology is suitable (and those for which it is not suitable) should be clearly defined. At a finer granularity, this should include a set of its competency questions (which for GO essentially amounts to being able to retrieve the taxonomic and partonomic relatives of the terms). Concrete examples of use should be shown, and proof of benefits from use will encourage others in the community to commit to the ontology.

The scope of an ontology of course does not have to be narrow, but it is important to keep in mind that development becomes increasingly labor- and time-intensive as richness and complexity grow. The community therefore must also be provided with intuitive software tools for ontology construction and editing and for browsing and querying its use in disparate databases. A community outside the realm of formal computer science is more likely to accept a resource with a relatively simple structure; if a complex ontology is constructed, the accompanying tools should be able to hide some of this complexity from members of the community attempting to work with it. Whatever the degree of complexity with which an ontology is initially built, it should be able to evolve to satisfy the requirements of those interested in using it. At the simpler end, this evolution can consist mostly of relatively inexpensive additions of concepts to the ontology, as is the case with GO. Such an effort should be coordinated by a team of dedicated curators.

Using these methods, the development of GO has been sustainable thus far, although more sophisticated representations and methodologies may be required in the future as its size further increases. One such example under consideration is GONG (Gene Ontology Next Generation), in which GO has been expressed as a Description Logic (specifically, DAML+OIL), thus enabling formal definition of concepts as well as the ability to reason over and automatically classify these term definitions [5]. Furthermore, the Open Biological Ontologies Consortium (http://obo.sourceforge.net/), of which GO is a member, has recently proposed that DAML+OIL's successor, OWL, be adopted as one of their ontology exchange languages.

The growth of GO over the last five years has been impressive. We are not implying that an ontology must have all of the previously described characteristics of GO to be

successful, but rather imparting what we believe has contributed to its success with the hope that it may be of use to ontology projects in other domains. With the use of a host of successful ontologies such as GO, many more resources, on the Web and elsewhere, could be more readily and reliably exploited.

References

[1] L. Badea, Functional discrimination of gene expression patterns in terms of the Gene Ontology, Pacific Symposium on Biocomputing 8 (2003) 565-576.

[2] The Gene Ontology Consortium, Gene Ontology: tool for the unification of biology, Nature Genetics 25 (2000) 25-29.

[3] The Gene Ontology Consortium, Creating the gene ontology resource: design and implementation, Genome Research 11 (2001) 1425-1433.

[4] O. D. King, J. C. Lee, A. M. Dudley, D. M. Janse, G. M. Church, and F. P. Roth, Predicting phenotype from patterns of annotation, Bioinformatics 19 (2003) I183-I189.

[5] C. J. Wroe, R. Stevens, C. Goble, and M. Ashburner, A Methodology to Migrate the Gene Ontology to a Description Logic Environment Using DAML+OIL, Pacific Symposium on Biocomputing 8 (2003) 624-635.