Wu, S., & Stvilia, B. (2014). Exploring the Development and Maintenance Practices in the Gene Ontology.
Advances In Classification Research Online, 24(1). doi:10.7152/acro.v24i1.14675

# Exploring the Development and Maintenance Practices in the Gene Ontology

Shuheng Wu
sw09f@my.fsu.edu
Besiki Stvilia
bstvilia@fsu.edu
School of Library & Information Studies, Florida State University,
Tallahassee, FL, 32306-2100, United States

**ABSTRACT**

The Gene Ontology (GO) is one of the most widely used and successful bio-ontologies in biomedicine and molecular biology. What is special about GO as a knowledge organization (KO) system is its collaborative development and maintenance practices, involving diverse communities in collectively developing the Ontology and controlling its quality. Guided by Activity Theory and a theoretical Information Quality Assessment Framework, this study conducts qualitative content analysis of GO's curation discussions. The study found that GO has developed various tools and mechanisms to gain expert feedback and engage various communities in developing and maintaining the Ontology in an efficient and less expensive way. The findings of this study can inform KO system designers, curators, and ontologists in establishing functional requirements and quality assurance infrastructure for bio-ontologies and formulating best practices for ontology development.

**Keywords**

Gene Ontology, ontology development, ontology maintenance, data quality, Activity Theory, knowledge organization.

## INTRODUCTION

Due to the complexity of molecular biological entities (e.g., genes, proteins) and their relationships, there has been a trend towards the development and adoption of bio-ontologies in the biomedical and molecular biological communities (Rubin et al., 2006; Wu, Stvilia, & Lee, 2012). Rubin et al. (2006) defined *bio-ontologies* as collections of standardized, human-interpretable, and machine-processable representations of entities and relationships between these entities within a specific biological domain, providing scientists with an approach to annotating, analyzing, and integrating results of scientific and clinical research. Among many of the current bio-ontologies, the Gene Ontology (GO) is one of the most successful, and has been widely used for text mining and information extraction (Blaschke, Hirschman, & Valencia, 2002; Kelso, Hoehndorf, & Prüfer, 2010).

*Advances in Classification Research, 2013,* November 2, 2013, Montreal, Quebec, Canada.

Founded in 1998, GO consists of three ontologies describing the cellular components (CC), molecular functions (MF), and biological processes (BP) of genes and gene products in a species-neutral manner, and intends to provide each gene and gene product with a cellular context (Gene Ontology, 2013a; Gene Ontology Consortium, 2011). GO is open access, and the ontology data can be downloaded for free in different formats. Users can view and search GO terms and annotations (i.e., the association between a GO term and a gene or gene product supported by an evidence source) via a browser named Amigo (http://amigo.geneontology.org/cgi-bin/amigo/go.cgi).

A typical GO term record contains a GO term name and accession number (i.e., identifier), the ontology to which the term belongs, synonyms classified into different categories (i.e., exact, related, narrower, and broader) to aid searching, definition of the GO term and reference source of the definition, comments provided by the curators, the subset (e.g., prokaryote-specific) to which the GO term belongs, usage comments to the GO term provided by users on the GONUTS wiki, ancestors and children of the GO term linked with different types of relationship displayed in a number of different views (e.g., tree view), and links to annotations using the GO term (Gene Ontology Consortium, 2011). GO uses four types of relationship between terms: 'is-a', 'part of', 'regulates' (positively regulates and negatively regulates), and 'has-part' (Gene Ontology, 2013c; Gene Ontology Consortium, 2009).

The development and maintenance of bio-ontologies usually rely on curators reading and interpreting scientific literature and extracting concepts and relationships between these concepts from the literature (Kelso et al., 2010). However, these processes are time-consuming and financially costly without community engagement (Greenberg, Murillo, & Kunze, 2010). What is special about GO as a knowledge organization (KO) system is its collaborative development and maintenance practices, involving diverse communities in collectively developing the Ontology and controlling its quality. GO has created a


ISSN: 2324-9773

number of data-related and software-related request trackers hosted at SourceForge (http://sourceforge.net/) to allow any individual to provide feedback on the Ontology, such as suggesting a new term or definition, reorganizing a section of the Ontology, and reporting errors or omissions in GO annotations (Gene Ontology, 2013d; Gene Ontology Consortium, 2006, 2007). GO curators review individual requests and implement edits where appropriate.

The purpose of this empirical study is to examine GO's collaborative development and maintenance practices and identify GO's data quality issues, gaining an understanding of how GO engages different communities in contributing content and detecting and correcting data quality problems. One of the widely used definitions of *quality* is "fitness for use" (Juran, 1992). This paper reports on GO's data quality assurance activities as they are one of the major components of ontology development and maintenance. As an open, large-scale scientific KO system, GO's collaborative development and maintenance processes can be reused or extended for other ontologies or KO systems, and inform system designers, data curators, and ontologists in establishing the functional requirements and quality assurance infrastructure for bio-ontologies and formulating best practices for ontology development.

## RESEARCH QUESTIONS AND DESIGN

Guided by Activity Theory (Engeström, 1990; Leont'ev, 1978) and Stvilia's Information Quality Assessment Framework (Stvilia, Gasser, Twidale, & Smith, 2007), this paper reports on a qualitative content analysis (Schreier, 2012) of GO's community curation discussions at SourceForge to answer the following research questions: (a) What are the types of data quality problems present in GO? and (b) What are the collaborative development and maintenance processes present in GO, including communities, division of labor, actions, tools, rules, and mechanisms used?

GO's Ontology Requests Tracker is one of those data-related request trackers hosted at SourceForge providing different scientific communities with a means for data quality negotiations and discussions as well as collaborative quality control. Examining the negotiations and discussions in this tracker can help identify GO's typology of data quality problems and dimensions that are deemed important by the communities and their quality assurance practices (Stvilia, Twidale, Smith, & Gasser, 2008). A random sample of 320 requests from the past two years was drawn from GO's Ontology Requests Tracker. The sample size was determined using the technique introduced by Powell and Connaway (2004). The unit of analysis is individual requests submitted to the Tracker, most of which include curators' comments and the curation actions they had taken.

The qualitative content analysis was applied in two phases, similar to the one conducted in Stvilia et al. (2008) with a set of predefined themes developed based on Activity Theory (Engeström, 1990; Leont'ev, 1978) and Stvilia's Information Quality Assessment Framework (Stvilia, Gasser, Twidale, & Smith, 2007). The predefined themes include: communities, division of labor, types of data quality problems, actions, tools, and rules. During the first phase of content analysis, emergent codes of each theme were generated based on interpreting each request in the sample to form a coding scheme through iteratively clustering, comparing, and revising the codes. In the second phase, all the requests in the sample were recoded using the coding scheme.

## FINDINGS

### Types of Data Quality Problems and Corresponding Quality Assurance Actions

The study identifies a typology of 23 data quality problems in GO and the corresponding quality assurance actions suggested by the requesters or taken by GO Administrators and GO Developers (see Table 1). These data quality problems can be classified into three categories: semantic issues, structural issues, and linked data related issues. Missing a GO term is the most frequently occurred one, indicating the difficulty that the Ontology has to represent new or established knowledge. For example, a user submitted a request for a new GO term to annotate several genes discussed in the literature:

> I need a new term for annotating several genes described in PMID 22902739 where they investigate stalk morphogenesis.
>
> NEW: sorocarp stalk morphogenesis
> part_of
> GO:0031288 sorocarp morphogenesis
>
> Def: The process in the sorocarp stalk is generated and organized. An example of this process is found in Dictyostelium discoideum.

The user provided the new GO term with a definition and a reference (i.e., PMID), and linked the new term to an existing GO term with the 'part_of' relationship. Another identified data quality problem—incorrect selection of preferred terms—suggests the importance of choosing the most widely used terms to be included in ontologies to represent community data practices and to gain community acceptance (Hjørland, 2007). For example, a user requested to add a synonym—CENP-A loading—to a GO term 'GO:0034080 CenH3-containing nucleosome assembly at centromere', and also asked to change the GO term name to one that is more widely used and species-neutral:

> Actually, could the primary name be changed to CENP-A containing nucleosome assembly at centromere? This is used more universally (I think CenH3 is an organism specific name?).

### Communities

The qualitative content analysis found a number of scientific communities actively participating in developing, maintaining, and using GO, which include but are not


ISSN: 2324-9773

| Problem types | Actions taken or suggested |
|---|---|
| Missing a GO term | Add, define, cite, place, comment |
| Typo in a GO term | Correct |
| Incorrect GO term name | Rename, cite |
| Incorrect selection of a preferred term | Replace, choose the most widely used form |
| Incorrect/incomplete definition of a GO term | Redefine, update, cite |
| Conglomeration in a GO term | Split, distinguish, add, obsolete |
| Redundant GO terms | Merge, reuse (as synonyms) |
| Invalid GO terms | Obsolete, replace |
| Missing synonym(s)<br>Incorrect synonym<br>Incorrect classification of a synonym | Add, cite<br>Remove, replace<br>Reclassify |
| Missing reference source<br>Incorrect reference source<br>Redundant reference source | Add<br>Update<br>Remove |
| Missing taxon constraint<br>Incorrect taxon constraint<br>Redundant taxon constraint | Add<br>Obsolete, remove<br>Replace |
| Missing a relationship between two GO terms<br>Incorrect type of relationship<br>Redundant relationship<br>Incorrect structural placement | Add<br>Replace<br>Remove<br>Reorganize, move, revert |
| Missing a type of GO relationship | Add, define, comment, exemplify |
| Complexity of a GO term | Link |

**Table 1. Data quality problem types and quality assurance actions taken or suggested.**

limited to FlyBase, PomBase, Saccharomyces Genome Database (SGD), WormBase, the Zebrafish Model Organism Database (ZFIN), the Ontology for Biomedical Investigations (OBI), CALIPHO, CellXP, Reactome, TAIR, UniProt, TermGenie, and PAINT. Among these communities, some are formed around model organism databases (e.g., FlyBase, PomBase,); some are bioinformatics resource centers (e.g., UniProt, TAIR); and others are tools or applications developed for using and maintaining GO (e.g., TermGenie, PAINT).

**Division of Labor**

The data analysis found three types of system accounts in GO's Ontology Requests Tracker—GO Administrators, GO Developers, and registered users—playing the roles of requesters, editors, reviewers, and commenters. As mentioned above, any registered user can be a requester submitting requests to any of GO's trackers. GO Administrators and GO Developers are registered users with specific privileges and permissions. Similar to journal editors, GO Administrators are responsible for reviewing requests submitted to different trackers and implementing edits where appropriate. They can also assign requests to appropriate GO Developers, who have specific domain knowledge to review those requests. Similar to journal reviewers, GO Developers are usually experts from the abovementioned communities helping GO Administrators develop and maintain the Ontology. However, GO Administrators and GO Developers usually cannot review requests submitted by themselves. Commenters are those participating in the conversation between the requester and the reviewer to support a request, express their viewpoints, provide more evidence, or oppose the request. Although they cannot make a decision whether to accept or reject a request, commenters may change the direction of the conversation, raise new quality issues, or become requesters.

**Tools**

According to Activity Theory (Engeström, 1990; Leont'ev, 1978), tools can be defined as the external objects or internal symbols that the communities use to detect and resolve data quality problems present in GO. The study identified the following categories of tools: biological literature (e.g., PubMed, PMC), other ontologies (e.g., Cell Ontology, Plant Ontology), data repositories (e.g., MetaCyc, UniProt), books (e.g., Wikibooks, textbooks), dictionaries and thesaurus, encyclopedias (e.g., Wikipedia), research or lab Websites, domain experts, and tools specifically developed for GO (e.g., GOCHE, QuickGO, TermGenie).

Particularly, GOCHE is a recently developed database used to check the structural representation of GO terms (Gene Ontology Consortium, 2011). Some of the complex GO terms, such as BP terms, contain chemicals. GOCHE curates these chemicals, which are linked to the Ontology of Chemical Entities of Biological Interest (ChEBI) and arranged into a structure aligning with that of ChEBI. GO curators use GOCHE to check if there are any misalignments between representation of GO terms and representation of those chemicals in ChEBI. When misalignments are found, GO curators will collaborate with ChEBI curators to resolve the discrepancies.

QuickGO (http://www.ebi.ac.uk/QuickGO) is a Web-based tool that allows users to construct a broad overview or subsets of GO and associated annotations using a set of


ISSN: 2324-9773

filters, such as taxonomic data and evidence codes (Binns et al., 2009). The number of concepts and associated annotations curated in GO has increased rapidly, and may overwhelm users (Jupp et al., 2012). Users may be interested in a small subset of GO to perform specific tasks (e.g., over expression analysis). QuickGO can help users navigate GO and provide views of GO that contain a set of tailored data for specific tasks. The findings showed that GO users and curators use QuickGO to generate graph views of a subset of GO to facilitate their quality discussions and negotiations. For example, a user submitted a request to remove the relationship between two GO terms—'GO:0007103 spindle pole body duplication' and 'GO:0005635 nuclear envelop'. A commenter provided evidence to support the request, including a link to a graph view of the GO term generated by QuickGO:

> Even when spindle pole body duplication [GO:0007103] occurs in the nuclear envelop [GO:0005635], not ALL of the components of the spindle pole body are embedded in the nuclear envelope making this parentage problematic… http://www.ebi.ac.uk/QuickGO/GTerm?id=GO:0007103#term=ancchart
> The spindle is associated with the nuclear envelope when it duplicates, but it isn't part of the nuclear envelope per se, it is in a fenestration…

Another widely used tool is TermGenie (http://go.termgenie.org/), which is a Web application using a pattern-based approach to help users and curators rapidly create new GO terms and place them appropriately in the Ontology (Gene Ontology, 2013b). The templates or patterns provided by TermGenie can ensure the completeness and accuracy of metadata in a GO term record.

### Rules

Rules refer to explicit or implicit norms, conventions, and regulations enabling or limiting actions and interactions of the communities (Engeström, 1990). Most of the rules identified in the content analysis are documented in GO Website, GO wiki, GO logs, and GO's request trackers. One interesting finding is a rule regulating new GO term requests among the communities. Similar to the literary warrant in the Library of Congress Subject Headings, requesters are required to provide pertinent literature reference, usually a PubMed identifier (i.e., PMID) or a data repository identifier, to validate the existence of a new term. Curators may review the literature or the data entry to decide whether to accept or reject the new term request. For example, a user submitted a new term request without providing any references. The reviewer asked, "Hi, Susan [pseudonym], Do you have a reference so we can see what the mechanism is?"

## DISCUSSION

Not surprisingly, some of the identified data quality problems in Table 1 are ontology-specific, such as missing a GO term and those structural representation issues. These findings imply ontology development should focus on representing domain concepts and their relationships to reflect established knowledge and keep up to date with new knowledge. Besides concentrating on knowledge representation, GO has endeavored to develop various tools and mechanisms to involve diverse communities in collaboratively developing and maintaining the Ontology and help users search, browse, view, use, and integrate ontological data. GOCHE is an instance of GO collaborating with another ontology to control its quality and form a broader Web of knowledge. Similar to user tagging (Trant, 2009), TermGenie empowers the user community to create new GO terms to represent the knowledge of their interests. Meanwhile, TermGenie's pattern-base approach can standardize the creation process and ensure the validity, completeness, and structure of user-contributed GO term records. GO's request trackers at SourceForge provide users and other communities with a platform to communicate with curators and participate in ontology development and maintenance. A set of rules (e.g., literary warrant, species-neutrality) has also been established among these communities to guide their curation activities and bridge community gaps. Informed by the development and maintenance practices in GO, KO system designers in libraries, archives, and museums can consider investing on developing tools or applications that can empower users to contribute contents and provide feedback, and support the use/reuse of bibliographic data for a wider range of tasks and user communities.

## CONCLUSION

This paper examines the collaborative development and maintenance practices in GO by analyzing the curation discussions in GO's Ontology Requests Tracker. Future research includes conducting quantitative content analysis on that Tracker (e.g., the distribution of data quality problems, the distribution of communities participating in ontology development) to gain a statistical profile of GO's quality value structure (Stvilia, 2007; Wu, 2013); interviewing stakeholders (e.g., GO Administrators, GO developers, GO users from different communities); and proposing a set of context-specific quality metrics to assess the quality of GO.

## REFERENCES

Binns, D., Dimmer, E., Huntley, R., Barrell, D., O'Donovan, C., & Apweiler, R. (2009). QuickGO: A web-based tool for Gene Ontology searching. *Bionformatics, 25*, 3045-3046. doi:10.1093/bioinformatics/btp536

Blaschke, C., Hirschman, L., & Valencia, A. (2002). Information extraction in molecular biology. *Briefings in Bioinformatics, 3*, 154-165.


ISSN: 2324-9773

Engeström, Y. (1990). *Learning, working and imagining: Twelve studies in activity theory*. Helsinki, Finland: Orienta-Konsultit Oy.

Gene Ontology. (2013a). *An introduction to the Gene Ontology*. Retrieved from http://www.geneontology.org/GO.doc.shtml

Gene Ontology. (2013b). *GeneOntology TermGenie Web application*. Retrieved from http://go.termgenie.org/

Gene Ontology. (2013c). *GO Ontology relations*. Retrieved from http://www.geneontology.org/GO.ontology.relations

Gene Ontology. (2013d). *The Gene Ontology at SourceForge*. Retrieved from http://geneontology.sourceforge.net/

Gene Ontology Consortium. (2006). The Gene Ontology (GO) project in 2006. *Nucleic Acids Research, 34*, D322-D326. doi:10.1093/nar/gkj021

Gene Ontology Consortium. (2007). The Gene Ontology (GO) project in 2008. *Nucleic Acids Research, 36*, D440-D444. doi:10.1093/nar/gkm883

Gene Ontology Consortium. (2009). The Gene Ontology in 2010: Extensions and refinements. *Nucleic Acids Research, 38*, D331-D335. doi:10.1093/nar/gkp1018

Gene Ontology Consortium. (2011). The Gene Ontology: Enhancements for 2011. *Nucleic Acids Research, 2011*, 1-6. doi:10.1093/nar/gkr1028

Greenberg, J., Murillo, A., & Kunze, J. A. (2012). Ontological ownership: Empowerment and sustainability. *Advances in Classification Research Online, 23*, 47-48.

Hjørland, B. (2007). Semantics and knowledge organization. *Annual Review of Information Science and Technology, 41*, 367-405. doi:10.1002/aris.2007.1440410115

Juan, J. (1992). *Juan on quality by design*. New York, NY: The Free Press.

Jupp, S., Gibson, A., Malone, J., Parkinson, H., & Stevens, R. (2012, July). *Taking a view on bio-ontologies*. Paper presented at the 3rd International Conference on Biomedical Ontology, Graz, Austria. Retrieved from http://ceur-ws.org/Vol-897/session4-paper22.pdf

Kelso, J., Hoehndorf, R., & Prüfer, K. (2010). Ontologies in biology. In R. Poli, M. Healy, & A. Kameas (Eds.), *Theory and applications of ontology: Computer applications* (pp. 347-371). New York, NY: Springer. doi:10.1007/978-90-481-8847-5_15

Leont'ev, A. (1978). *Activity, consciousness, personality*. Englewood Cliffs, NJ: Prentice–Hall.

Powell, R. R., & Connaway, L. S. (2004). *Basic research methods for librarians* (4th ed.). Westport, CT: Libraries Unlimited.

Rubin, D. L., Lewis, S. E., Mungall, C. J., Misra S., Westerfield, M., Ashburner, M., … Musen, M. A. (2006). National Center for Biomedical Ontology: Advancing biomedicine through structured organization of scientific knowledge. *OMICS, 10*(2), 185-198.

Schreier, M. (2012). *Qualitative content analysis in practice*. Thousand Oaks, CA: Sage Publications.

Stvilia, B. (2007). A model for ontology quality evaluation. *First Monday, 12*(12). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2043/1905

Stvilia, B., Gasser, L., Twidale, M., & Smith, L. C. (2007). A framework for information quality assessment. *Journal of the American Society for Information Science and Technology, 58*, 1720-1733. doi:10.1002/asi.20652

Stvilia, B., Twidale, M., Smith, L. C., & Gasser, L. (2008). Information quality work organization in Wikipedia. *Journal of the American Society for Information Science and Technology, 59*, 983-1001. doi:10.1002/asi.20813

Trant, J. (2009). Studying social tagging and folksonomy: A review and framework. *Journal of Digital Information, 10*(1). Retrieved from http://arizona.openrepository.com/arizona/handle/10150/105375

Wu, S. (2013). A model for assessing the quality of Gene Ontology. In W. Moen (Chair), *Proceedings of iConference 2013* (pp. 953-956). Fort Worth, TX. Champaign, IL: iSchools. doi:10.9776/13492

Wu, S., Stvilia, B., & Lee, D. J. (2012). Authority control for scientific data: The case of molecular biology. *Journal of Library Metadata, 12*, 61-83. doi:10.1080/19386389.2012.699822


ISSN: 2324-9773