

# Searching for Uses and Users in Gene Ontology Research

W. John MacMullen

Graduate School of Library & Information Science

University of Illinois, 501 E. Daniel St, MC-493, Champaign IL 61820-6211 USA wjohn@illinois.edu

## ABSTRACT

### Introduction:

Despite millions of dollars in investment over the past decade in the creation and maintenance of the Gene Ontology (GO), little is known about how (or even if) its intended end users – biomedical researchers – actually employ the ontology and its related databases and interfaces in their work. This project is a preliminary investigation of what evidence exists in the literature of specific uses of GO by researchers, and of use cases proposed for researchers by system designers. This work will help inform future in-depth studies of the specific information needs and research questions that researchers might use GO and other similar knowledge structures to address. It also provides to library and information science researchers and practitioners some insight into the quantity, sources, and breadth of publications about GO that exist.

### Methods:

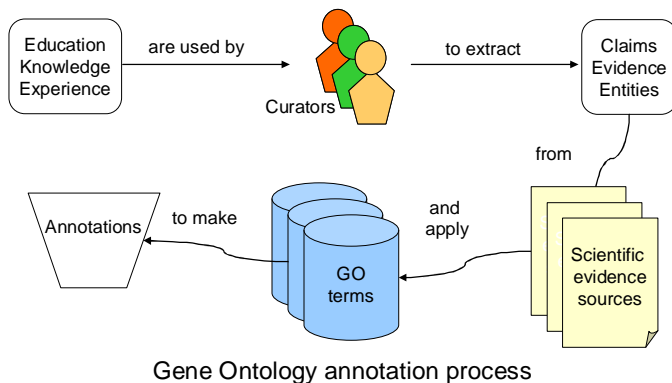
To better understand actual use, this exploratory literature-based study investigated the existence of user studies of biomedical scientists' use of the Gene Ontology; experimental articles that cite the use of GO; GO infrastructure and -application articles that describe user needs assessment; and articles that provide GO use cases. Numbers of GO-related articles were tabulated from searches in several bibliographic databases (PubMed/MEDLINE, EMBASE, BIOSIS, ISI Web of Science, Scopus, the ACM Digital Library, IEEE Xplore, and Library and Information Science Abstracts). A manual examination of the GO *Bibliography*, a database of GO-related articles maintained on the GO site, was used to supplement the bibliographic database searches. The GO *Bibliography* categorizes papers into 19 types, including 8 that explicitly refer to types of 'use'. The full bibliographic records for all citations from each of the eight 'use' categories were retrieved from PubMed. These were loaded into a purpose-built MySQL database and the intersection of the eight sets was computed in order to identify articles duplicated across categories.

### Results:

There were 1,342 distinct articles in the GO *Bibliography*'s 8 use types. Biomedical bibliographic database searching yielded similar quantities of GO-related articles, ranging from 1,234 to 1,556. Many GO papers assert the usefulness of GO and other ontologies for data integration across organisms, information retrieval, and knowledge discovery, but do not provide more precise needs and uses of end-user researchers. Specific use cases within a small number of the articles were identified and examined (see paper). Additionally, 30 articles were randomly selected from the GO *Bibliography* 'use' categories and characterized by a biologist by use case, biological question, methods, disciplinary affiliation, and other facets. Example use cases are shown here (far right).

### Acknowledgments

- This work was funded in part by the Graduate School of Library & Information Science. Use case data collection was performed in part by Shu-wen Huang, graduate research assistant.



Gene Ontology annotation process

## GO *Bibliography* 'use' categories and article quantities

Category	Articles
Use of GO in gene expression studies	857
Use of GO in clinical applications	485
Use of GO in biological databases	164
Use of GO in proteomics studies	125
Use of GO in network modeling and analysis	122
Use of GO in comparative genomics and evolutionary analysis	109
Use of GO in data or text mining	98
Use of GO to support predictions	59

Source: <http://www.geneontology.org/cgi-bin/biblio.cgi> (2008-02-26)

## Quantities of GO-related articles by source, as of 2008-02-26

Database	Query	Quantity
GO <i>Bibliography</i> (distinct articles across all categories*)		1,824
– Distinct articles across all 'use' categories		1,342
PubMed/MEDLINE	gene ontology	1,408
- in title	gene ontology[ti]	171
- in title & abstract	gene ontology[tiab]	1,237
- using search strategy #1	[see paper]	145
- using search strategy #2	[see paper]	94
Scopus	TITLE-ABS-KEY ("gene ontology")	1,556
ISI Web of Science	Topic=("gene ontology")   Title=("gene ontology")	1,528
EMBASE	gene ontology	1,308
BIOSIS	gene ontology	1,234
ACM Digital Library	gene ontology ("gene ontology") & ("use case")	338 5
IEEE Xplore	gene ontology	119
Library and Information Science Abstracts (LISA)	gene ontology	3

\*Some articles are classified in multiple categories, so summing category totals yields 3,108.

## GO *Bibliography* distinct 'use' articles

### Articles by publication year

Publication year	Articles
2008*	90
2007	444
2006	351
2005	248
2004	122
2003	49
2002	14
2001	4

\*As of 2008-02-26

### Articles by journal (top 10 of 104)

Journal	Articles
Nucleic Acids Research	133
BMC Bioinformatics	127
Bioinformatics	110
PLoS Genetics	93
PLoS Computational Biology	64
BMC Genomics	61
Physiological Genomics	45
PLoS Biology	42
Genome Biology	41
Proteomics	33

## GO use case examples

Examples of GO use cases identified from 30 articles randomly selected from the GO *Bibliography*'s 'use' categories:

- Characterize genomic regions associated with transcription factor binding. (PMID: 18271625)
- Inferring protein complexes from global protein-protein interaction networks. (PMID: 18270078)
- Characterize chronic kidney disease genes. (PMID: 18266955)
- Identify up- and down-regulated genes with significant differential expression. (PMID: 18259788)
- Identify statistically overrepresented functional groups of genes. (PMID: 18258795)
- Categorize genes responding to the changing hormonal environment at the transcriptome level during the oestrous cycle. (PMID: 18239051)
- Assign appropriate GO functional annotations to proteins according to the interaction pairs in diverse species having the shared domain patterns. (PMID: 18253506)
- Identify associated pathways with differentially expressed genes between obese and non-obese subjects in both human and rat. (PMID: 18239588)

## Searching for Uses and Users in Gene Ontology Research

### W. John MacMullen

Graduate School of Library & Information Science, University of Illinois at Urbana-Champaign,  
501 E. Daniel St., MC-493, Champaign, IL 61820, USA. Email: wjohn@uiuc.edu

### Introduction

Despite millions of dollars in investment over the past decade in the creation and maintenance of the Gene Ontology (GO), little is known about how (or even *if*) its intended end users – biomedical researchers – actually employ the ontology and its related databases and interfaces in their work. This project is a preliminary investigation of what evidence exists in the literature of specific uses of GO by researchers, and of use cases proposed for researchers by system designers. This work will help inform future in-depth studies of the specific information needs and research questions that researchers might use GO and other similar knowledge structures to address. It also provides to library and information science researchers and practitioners some insight into the quantity, sources, and breadth of publications about GO that exist.

The Gene Ontology is intended as a means of cross-organism integration of knowledge of the molecular functions, biological processes, and subcellular localization of gene products (Gene Ontology Consortium, 2008a). Much of basic biomedical research is performed on organisms that serve as surrogates for humans, due to their relative biological simplicity and the inappropriateness of direct experimentation on humans. GO integrates knowledge about gene function, process, and location across such so-called “model” organisms as the fruitfly *Drosophila melanogaster*, the mouse *Mus musculus*, and budding yeast *Saccharomyces cerevisiae*.

While dozens of ontologies and other controlled vocabularies for biomedical research, clinical medical informatics, and general science are under development (Smith, et al., 2007), and many different systems and applications have been developed to employ these vocabularies, there is often a knowledge gap between ontology developers, maintainers, and system designers on the one hand, and the target audience of end users (Rubin, Shah & Noy, 2008). Ontology developers and curators in biomedical informatics are frequently subject matter experts, holding the same academic credentials as the putative end users, and often have similar backgrounds and laboratory experience. Although developers may have a user perspective in mind, they are often in the position of trying to promote adoption and use of such systems to practicing researchers (see, e.g., Rhee, Dickerson & Xu, 2006).

Rubin, Shah & Noy (2008) classify biomedical ontology use into six functional types: search and query of heterogeneous biomedical data, data exchange among applications, information integration, natural language processing, representation of encyclopedic knowledge, and computer reasoning with data (76). While abstract examples of uses of ontologies in these areas are discussed, specific cases are not.

To better understand actual use, this exploratory literature-based study investigated the existence of user studies of biomedical scientists' use of the Gene Ontology; experimental articles that cite the use of GO; GO infrastructure and -application articles that describe user needs assessment; and articles that provide GO use cases.

## Methods

Numbers of GO-related articles were tabulated from searches in several bibliographic databases (PubMed/MEDLINE, EMBASE, BIOSIS, ISI Web of Science, Scopus, the ACM Digital Library, IEEE Xplore, and Library and Information Science Abstracts). Since MEDLINE primarily contains bioscience articles, the other bibliographic databases were searched to ensure user-focused studies from disciplines under-represented in PubMed (such as library and information science, sociology, ethnography, and social studies of science [SSS] / science and technology studies [STS]) were not overlooked. Search strategies for PubMed and the ACM Digital Library were created (Fig. 1) to locate articles related to uses of GO.

<p><b>PubMed:</b></p> <p>1.(gene ontology[ti] OR gene ontology[tiab]) AND (use[tiab] OR user[tiab] OR users[tiab] OR user study[tiab] OR user studies[tiab] OR use case*[tiab])</p> <p>2.(gene ontology[ti] OR gene ontology[tiab]) AND (user[tiab] OR user study[tiab] OR user studies[tiab] OR use case*[tiab])</p> <p><b>ACM Digital Library:</b> ("gene ontology") AND ("use case")</p>
---

Figure 1. PubMed and ACM Digital Library search strategies for user-focused GO articles

A manual examination of the *GO Bibliography* (Gene Ontology Consortium, 2008b), a database of GO-related articles maintained on the GO site, was used to supplement the bibliographic database searches. The *GO Bibliography* categorizes papers into 19 types, including 8 that explicitly refer to types of 'use' (Table 1). The full bibliographic records for all citations from each of the eight 'use' categories were retrieved from PubMed. These were loaded into a purpose-built MySQL database and the intersection of the eight sets was computed in order to identify articles duplicated across categories. Articles were characterized by publication year, journal name, and author affiliation.

Table 1. *GO Bibliography* 'use' categories and article quantities

Category	Articles
Use of GO in gene expression studies	857
Use of GO in clinical applications	485
Use of GO in biological databases	164
Use of GO in proteomics studies	125
Use of GO in network modeling and analysis	122
Use of GO in comparative genomics and evolutionary analysis	109
Use of GO in data or text mining	98
Use of GO to support predictions	59

Source: <http://www.geneontology.org/cgi-bin/biblio.cgi> (2008-02-26)

## Results

### *Literature characterization*

Table 2 presents the article quantities retrieved from the *GO Bibliography* and searches of the bibliographic databases, using both keyword searches on 'gene ontology' and the more focused search strategies from Figure 1. Very few articles were found when queries included terms related to use and users.

Table 2. Quantities of GO-related articles by source, as of 2008-02-26

Database	Query	Quantity
GO <i>Bibliography</i> (distinct articles across all categories*)		1,824
- Distinct articles across all 'use' categories		1,342
PubMed/MEDLINE	gene ontology	1,408
- in title	gene ontology[ti]	171
- in title & abstract	gene ontology[tiab]	1,237
- using search strategy #1	see Fig. 1	145
- using search strategy #2	see Fig. 1	94
Scopus	TITLE-ABS-KEY ("gene ontology")	1,556
ISI Web of Science	Topic=("gene ontology")   Title=("gene ontology")	1,528
EMBASE	gene ontology	1,308
BIOSIS	gene ontology	1,234
ACM Digital Library	gene ontology	338
	("gene ontology") & ("use case")	5
IEEE Xplore	gene ontology	119
Library and Information Science Abstracts (LISA)	gene ontology	3

\*Some articles are classified in multiple categories, so summing category totals yields 3,108.

The total number of articles across all GO *Bibliography* 'use' categories is 2,019, but due to the classification of some articles into multiple categories, and duplicated references in some lists, there are 1,342 distinct articles. Table 3 shows the distribution of distinct use-focused articles by publication year. Year-over-year growth has been significant, more than doubling every year in the early years, and slowing to 26% from 2006 to 2007 (still more than 1 new article per day). Table 4 shows the distribution of distinct use-focused articles by the top 10 (of a total of 404) journals, which account for 56% of the total number of articles, following a classic power law probability distribution. The top ten journals are a mix of biological science journals and bioinformatics and computational biology journals.

Table 3. Articles by publication year

Publication year	Articles
2008*	90
2007	444
2006	351
2005	248
2004	122
2003	49
2002	14
2001	4

\*As of 2008-02-26

Table 4. Articles by journal (top 10 of 104)

Journal	Articles
Nucleic Acids Research	133
BMC Bioinformatics	127
Bioinformatics	110
PLoS Genetics	93
PLoS Computational Biology	64
BMC Genomics	61
Physiological Genomics	45
PLoS Biology	42
Genome Biology	41
Proteomics	33

#### *Details of use cases*

Many GO papers assert the usefulness of GO and other ontologies for data integration across organisms, information retrieval, and knowledge discovery, but do not provide more precise needs and uses of end-user researchers. Specific use cases within a small number of the articles were identified and examined.

Sahoo, et al (in press) use an example of the complex problem of integrating information about the genetic components of nicotine dependence to illustrate the need for ontologies in the

creation of semantically integrated information resources. They provide three specific research questions a scientist might be interested in if pursuing this topic: “Which genes participate in a large number of pathways? Which genes (or gene products) interact with each other? Which genes are expressed in the brain?” (4) More detail is provided on each of the questions as the features of the system are described.

The “use case development” section in Shegogue & Zheng (2005) describes the use of class, responsibility, and collaboration (CRC) cards to gather requirements information which was then used to create a detailed document describing the main and alternate steps in a model. While it provides details of a biological process of interest, this is not a narrative scenario that describes a biological problem, question, or information need.

The “use cases” presented in Shah, et al. (2005) and Blake and Bult (2006) are examples of possible tasks and queries that a user might perform (e.g., a “known item” search using an accession number for a biological object such as a sequence), but specific motivations for these types of tasks, or the underlying biological significance or user relevance, are not discussed. Shah et al. do describe a use case where their system can be used to compile a list of reagents for use in identifying orthologous human genes that exist in model organisms, which “represent essential genes which are candidates for human disease agents” (11). Blake and Bult note that examination of ontological annotations in databases “can be extremely useful for experimental biologists making crucial decisions as to allocation of research resources for further characterization of specific genes (317)”.

## **Discussion**

### *Limitations*

The fact that few of the articles explicitly mention users or use cases indicates that little has been published on the information needs and practices of biomedical researchers. However, since many of the articles were published in biology-focused journals, it is possible that a substantial proportion of the articles could in fact be examples of researchers using GO to address biological problems, or the nature of the biological questions could be implicit in the design of the project being reported. More work is needed to explore the nature of these papers, and to assess the differences in the types of problems and questions explored in the underlying projects.

### *Future Work*

Both bibliographic and human subject research are needed to understand more about the information needs and user motivations for the use of GO. A fuller characterization of the literature identified above should be carried out to categorize types of uses, identify specific use cases, and segment the document sets into articles produced by ontology developers, system developers, and end-users.

The knowledge gained from the bibliographic research can be employed in studies of potential and actual end-user scientists’ motivations for using (and not using) GO in their work. Some researchers are relying upon GO annotations to validate computationally-derived predictions of biological activity (e.g., Kaminker, et al, 2007), which suggests that the perceived information quality (Nicolaou & McKnight, 2006) of GO is high. However, as MacMullen (2006) notes, the quality of GO annotations, like their use, is largely unquantified, and may have a variety of facets whose weights vary depending upon specific uses. Future work in this area could include awareness assessments conducted with prospective GO end-users based upon ideas such as the technology acceptance model (TAM) (e.g., Davis, 1989). Established constructs such as perceived ease of use (PEOU) and perceived usefulness (PU), which have been explored in relation to adoption of bioinformatics tools by scientists (see, e.g., Shachak, Shual & Fine, 2007) could be supplemented with others, such as the perceived advantage of using GO instead of another approach, or the use of one GO interface over another, and perceived goal attainment (PGA) (Glaser & Backer, 1980) – how can the use of GO help a researcher achieve her goals?

As Blake and Bult (2006) and Sahoo, et al. (in press) note, the true power of ontologies for driving industrial-scale e-Science research processes that repurpose the vast amounts of extant data has not yet been realized. User-centered and problem-based research by social scientists that addresses the range of researchers' information problems requires a far deeper and richer understanding of the specific information needs and tasks scientists face in the spectrum of their work, from large data management activities down to the individual experiments that comprise "the long tail of science" (Palmer, et al., 2007) performed in thousands of labs worldwide.

## REFERENCES

- Blake JA, Bult CJ (2006). Beyond the data deluge: data integration and bio-ontologies. *Journal of Biomedical Informatics* 39(3):314-320. PMID: 16564748
- Davis FD (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319-340.
- Gene Ontology Consortium (2008a). The Gene Ontology project in 2008. *Nucleic Acids Research* 36(Database issue):D440-444. PMID: 17984083.
- Gene Ontology Consortium (2008b). *GO Bibliography*. Available: <http://www.geneontology.org/cgi-bin/biblio.cgi> (Accessed: 2008-02-26).
- Glaser EM, Backer TE (1980). Durability of innovations: how goal attainment scaling programs fare over time. *Community Mental Health Journal*. 16(2):130-43. PMID: 7389300
- Kaminker JS, Zhang Y, Watanabe C, Zhang Z (2007). CanPredict: a computational tool for predicting cancer-associated missense mutations. *Nucleic Acids Research* 447: 714-719. PMID: 17537827.
- MacMullen WJ (2006). Facets and measures of Gene Ontology annotation quality in model organism databases. In *Proceedings of the 69th Annual Meeting of the American Society for Information Science & Technology (ASIS&T)*, Vol. 43.
- Nicolaou AI & McKnight DH (2006). Perceived Information Quality in Data Exchanges: Effects on Risk, Trust, and Intention to Use. *Information Systems Research* 17(4): 332-351. DOI: 10.1287/isre.1060.0103
- Palmer CL, Cragin MH, Heidorn PB, Smith LC (2007). Data Curation for the Long Tail of Science: The Case of Environmental Sciences. 3rd International Digital Curation Conference, Washington, DC, Dec. 13, 2007.
- Rhee SY, Dickerson J, Xu D (2006). Bioinformatics and its applications in plant biology. *Annual Review of Plant Biology* 57: 335-360. PMID: 16669765
- Rubin DL, Shah NH, Noy NF (2008). Biomedical ontologies: a functional perspective. *Briefings in Bioinformatics* 9(1):75-90. PMID: 18077472
- Sahoo SS, Bodenreider O, Rutter JL, Skinner KJ, Sheth AP (in press). An ontology-driven semantic mash-up of gene and biological pathway information: Application to the domain of nicotine dependence. *Journal of Biomedical Informatics*. DOI: 10.1016/j.jbi.2008.02.006
- Shachak A, Shuval K, Fine S (2007). Barriers and enablers to the acceptance of bioinformatics tools: a qualitative study. *Journal of the Medical Library Assoc.* 95(4):454-458. PMID: 17971896
- Shah SP, Huang Y, Xu T, Yuen MM, Ling J, Ouellette BF (2005). Atlas - a data warehouse for integrative bioinformatics. *BMC Bioinformatics* 6:34. PMID: 15723693
- Shegogue D, Zheng WJ (2005). Integration of the Gene Ontology into an object-oriented architecture. *BMC Bioinformatics* 6:113. PMID: 15885145
- Smith B, Ashburner M, Rosse C, Bard J, Bug W, et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology* 25(11):1251-1255. PMID: 17989687