

# Contextual Analysis of Variation and Quality in Human-curated Gene Ontology Annotations

W. John MacMullen  
University of North Carolina  
100 Manning Hall, CB# 3360  
Chapel Hill NC 27599-3360  
macmw@email.unc.edu

## ABSTRACT

This work employs a multi-methods approach combining prospective randomized controlled experiments with broad contextual analysis (including observations, concurrent verbal reports, document content analysis, workflow analysis, and interviews) to investigate the nature and extent of variation in human-curated annotations of the scientific literature using the Gene Ontology (GO), a standardized cross-organism controlled vocabulary. Data obtained to date include 4,000 GO annotation instances generated by 33 biological curators, and approximately 20 hours of audio recordings from observations and interviews. The results of this work will inform the development of organization- and organism-independent measures of GO annotation quality.

## Categories and Subject Descriptors

J.3 [LIFE AND MEDICAL SCIENCES]: – Biology and genetics; H.3.4 [Systems and Software]: Performance evaluation (efficiency and effectiveness); H.3.1 [INFORMATION STORAGE AND RETRIEVAL]: Content Analysis and Indexing; D.2.8 Metrics; H.3.7 Digital Libraries

## General Terms

Measurement, Performance, Experimentation, Human Factors, Standardization, Verification.

## Keywords

Annotation, Gene Ontology (GO), Information quality, Inter-indexer consistency, Inter-rater reliability.

## 1. INTRODUCTION

The wide variety of information in model organism databases has great potential utility for biomedical researchers, but only if it is of high quality. Human-curated annotations made in model organism databases are viewed as important links between an

organism's underlying genomic data and the experimental scientific literature. Gene Ontology (GO) annotations in particular are expected to provide significant value in reducing disciplinary fragmentation, and facilitating cross-organism information integration. Analysis of GO annotations could also be automated to enable inferencing and hypothesis generation. However, a common question associated with manually-curated knowledgebases (such as GO, model organism databases (MODs), and even PubMed/MEDLINE) is to what degree variation in human curators' annotations affects the overall quality of information in the resource. Information quality is a complex concept, with many facets that have interdependencies. While annotation quality is of great importance to the genomics community (see, e.g., [1]), very little is known about the types and amounts of GO annotation variation in MODs.

## 2. RESEARCH QUESTIONS

A broad question implied by the preceding is, "How and why do human curators differ in their GO annotation processes and outcomes?" This project investigates this through three more specific questions:

1. **How significantly do curators differ in annotation outcomes?** This question is investigated by conducting individual and consensus GO annotation experiments and comparing differences in curators' outcomes in the form of formal GO annotations against a parameterized model of annotation quality.
2. **Do differences in curators' educational-, training-, and research backgrounds influence their GO annotation performance?** This question is investigated by comparing the formal GO annotation data from the individual and consensus annotation experiments with demographic data collected from questionnaires, and in individual semi-structured interviews and focus groups.
3. **Do differences in curators' personal annotation behaviors influence their GO annotation performance?** This question is investigated by comparing formal GO annotation data from individual and consensus annotation experiments with data on personal annotation behaviors (such as workflows and resources employed) that were obtained from individual interviews with curators, focus group discussions, observations, and artifact analyses.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

*i-Conference Doctoral Colloquium*, Oct. 15–17, 2006, Ann Arbor, MI.

### 3. RESEARCH DESIGN

#### 3.1 Annotation Experiments

Two separate prospective randomized controlled experiments were conducted, one with biological curators from a single model organism database (MOD), and one with curators from multiple MODs and other databases. In the single-MOD study, 8 expert biological database curators from the same MOD manually extracted GO annotations from a corpus of 48 experimental articles at two levels of redundancy (16 papers at 4x and 32 papers at 2x). The experimental design randomly assigned each curator into one of two groups, and assigned each curator 16 articles to be curated over a four week period. In the 4x coverage set, each pair of curators who annotated each article subsequently rationalized their individual annotations to a single consensus annotation. Two additional curators each re-annotated two papers they had previously curated; this data will be used to evaluate to the Reliability quality measure (section 4.2, below).

The design of the multi-MOD study was similar to the single MOD study. Twenty-three curators from 11 databases annotated a corpus of 10 articles at full redundancy (i.e., every paper was annotated by every curator), over a four week period. Following the individual annotations, the pairs of curators who represented each database rationalized their individual annotations to a single consensus annotation for each paper.

#### 3.2 Contextual Analysis

The origins of the observed variation in curators' annotations cannot be determined solely from the formal GO annotations generated in the two studies described above. A variety of contextual data was collected during the studies to assist in explaining the variation. Data collection methods included unobtrusive observations of annotation creation and consensus discussions, concurrent verbal reports, document content analysis, workflow analysis, and individual curator interviews.

### 4. WORK IN PROGRESS

#### 4.1 Data Collection

The studies and data collection described above yielded a total of 400 individual and consensus GO annotations generated by 33 biological curators. Each 'annotation' is composed of one or more gene-specific instances of a formal GO annotation, resulting in nearly 4,000 instances. The studies also yielded approximately 650 pages of manually-annotated paper articles and associated intermediate notes, as well as approximately 20 hours of digital audio recordings from observations, individual curator interviews, and a focus group (at the end of the single-MOD study). Data on work practices and workflows will in turn be extracted from the audio files.

#### 4.2 GO Annotation Quality Measures

Six measures of GO annotation quality have been defined and parameterized [2]. The individual and consensus GO annotations

generated in the two studies will be evaluated with the following facets: consistency, reliability, specificity, completeness, accuracy, and validity, as summarized in Table 1.

**Table 1. GO Annotation Quality Facets**

Facet	Definition
Consistency	Variation across annotations made by different annotators
Reliability	Variation across annotations made by same annotator
Specificity	Relative granularity of annotation terms in relation to their source vocabulary
Completeness	Presence / absence of values for attributes in relation to standard annotation
Accuracy	Variation of individual annotations in relation to standard annotation
Validity	Variation of attribute values of individual annotations in relation to the GO annotation model

### 5. EXPECTED CONTRIBUTIONS

Original contributions expected from this work include:

- Definition and elaboration of the variation that exists in individual curators' formal GO annotations, and in their annotation practices, workflows, and related experiences.
- Evaluation of the fitness of the proposed measures of GO annotation quality.
- Content analysis of the scientific articles used in the studies, in relation to the outcomes of the quality measures, to assess whether structural facets may affect annotation creation.

### 6. ACKNOWLEDGMENTS

The author was supported in part by an Individual Biomedical Informatics Fellowship (F37 LM009194) from the National Library of Medicine, and a research fellowship from the Annotation of Structured Data project (Gary Marchionini, PI) at the School of Information and Library Science, University of North Carolina, Chapel Hill, which was funded by a gift from Microsoft Research.

### 7. REFERENCES

- [1] Dolan ME, Ni L, Camon E, Blake JA.. A procedure for assessing GO annotation consistency. *Bioinformatics*. 2005 Jun 1;21 Suppl 1:i136-i143. PMID: 15961450.
- [2] MacMullen, W.J. 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, to appear (Nov. 2006).

# Contextual Analysis of Variation and Quality in Human-curated Gene Ontology Annotations

W. John MacMullen

University of North Carolina, 100 Manning Hall, CB# 3360, Chapel Hill NC 27599-3360 macmw@email.unc.edu

## Abstract

This work employs a multi-methods approach combining prospective randomized controlled experiments with broad contextual analysis (including observations, concurrent verbal reports, document content analysis, workflow analysis, and interviews) to investigate the nature and extent of variation in human-curated annotations of the scientific literature using the Gene Ontology (GO), a standardized cross-organism controlled vocabulary. Data obtained to date include 4,000 GO annotation instances generated by 33 biological curators, and approximately 20 hours of audio recordings from observations and interviews. The results of this work will inform the development of organization- and organism-independent measures of GO annotation quality.

## RESEARCH QUESTIONS

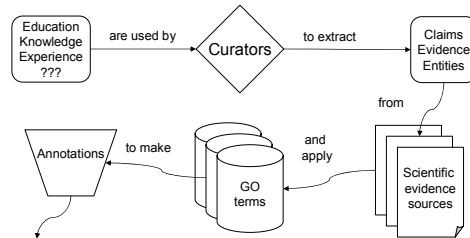
- How significantly do curators differ in annotation outcomes?** This question is investigated by conducting individual and consensus GO annotation experiments and comparing differences in curators' outcomes in the form of formal GO annotations against a parameterized model of annotation quality.
- Do differences in curators' educational-, training-, and research backgrounds influence their GO annotation performance?** This question is investigated by comparing the formal GO annotation data from the individual and consensus annotation experiments with demographic data collected from questionnaires, and in individual semi-structured interviews and focus groups.
- Do differences in curators' personal annotation behaviors influence their GO annotation performance?** This question is investigated by comparing formal GO annotation data from individual and consensus annotation experiments with data on personal annotation behaviors (such as workflows and resources employed) that were obtained from individual interviews with curators, focus group discussions, observations, and artifact analyses.

## RESEARCH DESIGN

Two separate prospective randomized controlled experiments were conducted, with scientific curators from model organism databases (MODs):

- Single-MOD study:** 8 curators from one MOD annotated a corpus of 48 experimental articles at two levels of redundancy (16 papers at 4x and 32 papers at 2x), with two pairs of curators creating consensus annotations.
- Multi-MOD study:** 23 curators from 11 databases annotated a corpus of 10 articles at full redundancy, with curators paired by database to create consensus annotations.
- Contextual data** about the curators' backgrounds, experience, and workflows were also collected, via unobtrusive observations of annotation creation and consensus discussions, concurrent verbal reports, document content analysis, workflow analysis, and individual curator interviews. Hand-annotated paper articles from six participants of the multi-MOD study were obtained for analysis.

## GENERIC GO ANNOTATION PROCESS, WITH EXAMPLE ANNOTATION INSTANCES FROM ONE SCIENTIFIC ARTICLE



Gene	Ontology	GO ID	GO Term	Evidence	With/From
PSF1	Process	GO:0006261	DNA-dependent DNA replication	IMP	
PSF1	Component	GO:0008111	GIN5 complex	IPI	PSF2, PSF3, SLD5
PSF2	Process	GO:0006261	DNA-dependent DNA replication	IGI	SLD5
PSF2	Component	GO:0008111	GIN5 complex	IPI	PSF1, PSF3, SLD5
PSF3	Process	GO:0006261	DNA-dependent DNA replication	IGI	PSF1
PSF3	Component	GO:0008111	GIN5 complex	IPI	PSF1, PSF2, SLD5
SLD5	Process	GO:0006261	DNA-dependent DNA replication	IMP	
SLD5	Component	GO:0008111	GIN5 complex	IPI	PSF1, PSF2, PSF3

## QUANTITATIVE DATA

Aggregate GO annotation data points by study	Single-MOD	Multi-MOD
Individual annotations per paper	*4	23
Total individual annotations	**128	230
Total consensus annotations	32	90
Pairwise comparisons per paper (annotation-level, individual & consensus)	45	528
Total paper-level pairwise comparisons	***752	5,280
Total individual instances	217	2,284
Total consensus instances	88	1,280
Overall instances	305	3,564
Total field-level pairwise comparisons	3,008	21,120

## CONTEXTUAL DATA

**Audio:** Approximately 20 hours of digital audio recordings were made during the two studies, in the following forms:

- 15 individual curator interviews
- 13 observations of paired consensus discussions
- 4 individual curator think-aloud observations
- 1 curator focus group

### Documents:

- Six sets of 10 manually annotated articles, plus additional personal notes (~650 pages total)
- 23 personal data questionnaires about background and experience
- Screenshots of the single MOD's curation interface, to assist in interpretation of the interviews and observations.
- Email records of the discussions leading to one pair of curators' consensus annotations.

## RESULTS

Preliminary descriptive analysis of the quantitative GO annotation data from the multi-MOD study suggests:

- Variation in the number of annotation instances by curator across papers is significant. (Range: 0 to 35; see Figure 1.) This may be influenced by attributes of the underlying papers.
- Variation in the number of annotation instances created by different curators for the same papers is very large. (Range: 0 to 46; see Figure 2.)

## GO ANNOTATION EVALUATION PROCESS

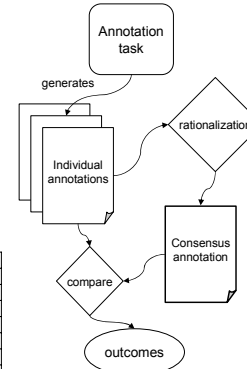


Fig. 1. Annotation instances per curator by paper (multi-MOD study)

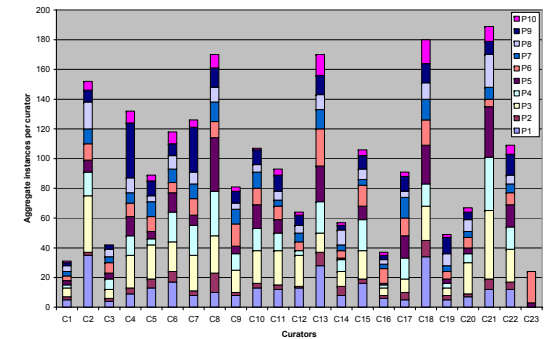
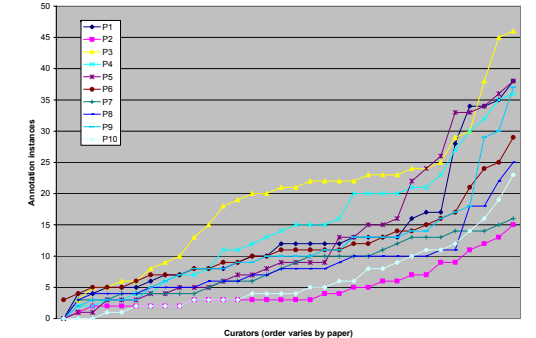


Fig. 2. Individual and consensus annotation instances per paper, in rank order (Multi-MOD study)



## GO ANNOTATION QUALITY FACETS

Facet	Definition
Consistency	Variation across annotations made by different annotators
Reliability	Variation across annotations made by same annotator
Specificity	Relative granularity of annotation terms in relation to their source vocabulary
Completeness	Presence / absence of values for attributes in relation to standard annotation
Accuracy	Variation of individual annotations in relation to standard annotation
Validity	Variation of attribute values of individual annotations in relation to the GO annotation model

## Acknowledgments

- This work benefited from discussions with curators from multiple model organism databases at the 2005 and 2006 Gene Ontology Annotation Camps held at Stanford University. The Stanford Department of Genetics and the Saccharomyces Genome Database (SGD) provided travel support.
- W.J.M. was supported in part by an Individual Biomedical Informatics Fellowship (F37 LM009194) from the National Library of Medicine, and an unrestricted research gift from Microsoft Research to the Annotation of Structured Data research group in the School of Information and Library Science at the University of North Carolina at Chapel Hill.