

# Facets and measures of Gene Ontology annotation quality in model organism databases

**W. John MacMullen**

School of Information and Library Science, University of North Carolina at Chapel Hill  
100 Manning Hall, CB#3360, Chapel Hill NC 27599-3360. Email: macmw@ils.unc.edu

**Model organism databases are important repositories of data and information for biomedical research, but are useful to scientists only if the information they contain meets certain levels of quality. This methodology paper describes six facets of information quality applicable to Gene Ontology (GO) annotations in model organism databases, and defines corresponding metrics to be used in measuring the quality of annotations made by one or more human database curators. The defined facets and measures of annotation quality are: consistency, reliability, specificity, completeness, accuracy, and validity. Contextual factors, and factors affecting internal and external validity, are also discussed.**

## Introduction

Model organism databases (MODs) aggregate and store data and information about organisms of interest to biomedical researchers. The MODs for major model organisms, such as the mouse, fruitfly, and yeast, have professional curatorial staffs that create and maintain the information in the databases. Much of the information in MODs is described as ‘annotations’ to the primary genetic sequence data of the underlying organism. This information is of varying forms and types, including curated literature used as evidence when making Gene Ontology annotations (Gene Ontology Consortium, 2006). MODs and organizational structures that link them, such as ontologies, are useful to scientists, but only if the information they contain meets certain levels of quality.

This paper defines generalized facets and measures of Gene Ontology (GO) annotations common to multiple MODs so that standards and outcomes can be compared across organisms. The focus of this work is on human-generated annotations, which can be studied in experimental settings with tasks carried out by one or more human annotators, or by other quantitative and qualitative analyses of annotation artifacts, as explored in other recent studies (MacMullen 2005, 2006). GO annotations are important to biology because they facilitate the database-independent cross-species integration of knowledge.

## Definitions

For the purposes of this paper, ‘annotations’ in general are the end products of the process of identifying and extracting claims and evidence from the scientific literature and linking them to related biological entities, such as genes. Specifically, the focus of this work is constrained to GO annotations, which have a formal structure (Gene Ontology Consortium, 2006). Each of the main GO Consortium MODs has in excess of 5,000 GO annotations for each of the three ontological aspects of GO – Molecular Function, Biological Process, and Cellular Component (GO, 2006). The sources of evidence used in making annotations are called ‘papers’, and refer to individual experimental articles published in scientific journals. ‘Claims’ are the individual statements of factual evidence made by authors that are manually extracted from papers by human annotators and used to create specific instances of annotations. The relationship of papers to claims is variable (each paper can contain zero to many claims relevant to the underlying organism, or to GO, or to both). The number of possible instances of GO annotations from each paper to each of the three GO aspects is also variable. The measures defined below are evaluated at the instance level, meaning one formal GO annotation to one ontological aspect.

## Quality facets and measures

In the research and methodology literatures of information and library science, biomedical informatics, psychology, and other domains, terms such as ‘inter-indexer consistency’, ‘inter-coder agreement’, or ‘inter-rater reliability’ often focus on a single quality measure, or conflate a variety of different measures. Often, only simple counts of agreement are used, not chance-adjusted measures such as Cohen’s kappa and its variants. This section operationalizes the multidimensional concept of information quality into six different quality facets relevant to GO annotations in MODs: consistency, reliability, specificity, completeness, accuracy, and validity. Each of these facets can be further decomposed into multiple attributes, which are parameterized below. These facets relate to the annotations themselves, not to the underlying evidence sources from which they are derived, or the truth value of the claims used in their construction. All except Validity require each paper

to be curated two or more times in order to allow comparison; thus those five measures cannot be applied to extant GO annotations if their underlying evidence sources have been curated only once (which is the typical case).

### *Definitions*

Let  $P_i$  denote a scientific paper used as an evidence source, and let  $A_i$  be an annotation instance extracted from that paper. A particular paper may have 0- $n$  annotation instances, and each may be made to any one (but only one) of the three GO aspects. For each  $A_i$ , two or more curators, denoted as  $C_{ij}$ , may have created annotation instances. For each annotation  $P_iA_iC_{ij}$ , the following measures are defined below:  $P_iA_iC_{ij}$  RELIABILITY,  $P_iA_iC_{ij}$  CONSISTENCY,  $P_iA_iC_{ij}$  SPECIFICITY,  $P_iA_iC_{ij}$  COMPLETENESS,  $P_iA_iC_{ij}$  ACCURACY, and  $P_iA_iC_i$  VALIDITY.

### *Consistency*

The Consistency facet measures simple pairwise or cumulative agreement among the attributes of individual GO annotation instances, for two or more annotators. This is a dichotomous measure of the agreement between the attribute values (i.e., same/different). The attributes of which Consistency is composed are the mandatory GO annotation fields ‘Gene product’, ‘GO ID’, ‘GO term’, and ‘Evidence code’, and the optional fields ‘Qualifier’, ‘and With/from’. For each evidence source or each annotation instance, Consistency may be measured at the attribute level or the facet level (i.e., as an overall consistency score).

### *Reliability*

The Reliability facet measures simple pairwise or cumulative agreement among the attributes of the original and one or more repeated annotation instances, as made by the same annotator at different time points. Reliability is thus essentially an ‘intra-annotator’ version of Consistency. Both Reliability and Consistency require reference to GO to determine whether any changes made in the GO vocabularies over the time intervals between compared annotations had material effects on term selection. The Discussion section addresses other aspects of repeated measures that may influence performance and confound analysis.

### *Specificity*

The Specificity facet measures the degree of agreement between the GO IDs in two or more annotation instances, by GO aspect. Degree of agreement is measured at three levels of granularity: first, categorically, following Camon, et al. (2005), as being 1) an exact match; 2) different terms, but with the same lineage; or 3) different terms, not in the same lineage; second, as a binary distinction of one term being broader or narrower than the other; and third, as the nodal distance(s) between the compared GO IDs in relation to the depth of the branch(es) on which they reside.

### *Completeness*

The Completeness facet measures simple presence or absence of attributes in individual annotation instances in relation to a consensus annotation. Measurement of this facet requires a paper to have been curated by at least two curators, who then rationalize their individual annotations to create a reference standard against which their (and other curators’) individual annotations can be compared. All mandatory and optional GO fields are compared for true and false positives (an annotation instance is present or not present in the reference annotation), and true and false negatives (an annotation instance is not present or present in the reference annotation). The semantic content of the annotation instances is not evaluated by this facet.

### *Accuracy*

The Accuracy facet measures simple pairwise or cumulative agreement between individual and consensus (standard) annotations at the attribute level. While Completeness may appear to be a form of Accuracy (i.e., an incomplete annotation can be thought of as inaccurate) Accuracy is distinguished from Completeness in that it evaluates annotation instances for factual agreement with a reference annotation, while Completeness checks only for the presence or absence of any annotation instances, accurate or not. As with Completeness, this measure requires a reference annotation for comparison.

### *Validity*

The Validity facet is a basic error-checking test that evaluates whether the values supplied for each attribute of an annotation instance fall within the range of allowed values. This is a dichotomous measure that evaluates an annotation against the GO annotation definition, not against other individual or consensus annotations. Some MODs perform this check as an internal function of their databases, but this measure is MOD- and database-independent.

### *Facet Applicability*

While the Consistency facet applies to multi-annotator cases, and the Reliability facet applies a similar measure to single-annotator cases, the Specificity, Completeness, and Accuracy facets are applicable to both cases. Validity is evaluated individually. The quality facets defined above are summarized in Table 1. The attributes and parameters for the facets are shown in Figure 1.

Table 1. Summary of GO annotation quality facets

Facet	Definition	Scope	Variable types
Consistency	Variation across annotations made by different annotators	2 or more individual annotations	discrete (nominal)
Reliability	Variation across annotations made by same annotator	2 or more individual annotations	discrete (nominal)
Specificity	Relative granularity of annotation terms in relation to their source vocabulary	2 or more individual annotations	discrete (nominal) & interval (ordinal)
Completeness	Presence / absence of values for attributes in relation to standard annotation	1 or more individual annotations compared to standard (consensus)	discrete (nominal)
Accuracy	Variation of individual annotations in relation to standard annotation	1 or more individual annotations compared to standard (consensus)	discrete (nominal)
Validity	Variation of attribute values of individual annotations in relation to the GO annotation model	1 annotation compared to GO data model	discrete (nominal)

$P_i A_i C_{ij}$  RELIABILITY,  $P_i A_i C_{ij}$  CONSISTENCY,  $P_i A_i C_{ij}$  COMPLETENESS,  
 $P_i A_i C_{ij}$  ACCURACY, and  $P_i A_i C_i$  VALIDITY

$P_i A_i C_{ij}$  SPECIFICITY

Attributes	Values
Overall agreement	{0,1}
Gene product	{0,1}
GO ID	{0,1}
GO term	{0,1}
Evidence	{0,1}
Qualifier	{0,1,-}
With/from	{0,1,-}

Attributes	Values
Exact match	{0,1}
Mismatch, same lineage	{0,1}
Mismatch, different lineage	{0,1}
Broader	{0,1}
Narrower	{0,1}
Difference, in nodes	{0,n}

Figure 1. Attributes and parameters for GO annotation quality facets

## Discussion

This section reviews overall questions of validity and reliability of the facets and measures, as well as other factors influencing variation in annotations, and some conclusions and future work.

One potentially relevant quality attribute not addressed above is ‘currency’. This is another sense of the term ‘accuracy’, but in relation to scientific knowledge as a whole. Above, the Accuracy facet is measured at a single time point. What is not accounted for are effects related to the evolution of knowledge over time. If an annotation with perfect quality scores is demonstrated at a future time point to be incorrect, or is no longer the latest evidence available on a topic, that information needs to be evaluated in terms of its information quality. This is outside the scope of the current project, but is an important question for longer-term information quality evaluation in science. There is no reliable mechanism in the narrative text journal article paradigm whereby outdated knowledge at a relatively granular level is retrospectively indicated as such, but in the database paradigm, these kinds of changes become possible (although not entirely tractable at the present time).

### *Construct validity and reliability*

Factors affecting internal validity of these measures include what Campbell and Stanley (1966) call ‘history’ and ‘maturation’ effects: changes over time in annotators’ knowledge or skills that influence their behavior; this can also affect the assessment of annotations by expert judges. In addition to the other problems of the use of repeated measures, they can be biased by history effects. Changes in tools or workflows, or what Campbell and Stanley call ‘instrumentation’ (e.g., an ontology browser that makes it easier to find and select terms), ideally should be controlled for. However, these can also function as variables or controls in certain conditions. Since the number of professional curators in each MOD is relatively small (5-15), care must be taken to randomize paper assignment, paper order, and expert judge assignments. Papers themselves may have types or features that influence annotation, so a representative sample of papers must be defined and

selected to avoid oversampling of certain types (e.g., review articles, or specific methods used in experimental papers). As mentioned above, all facets except Validity require each paper to be curated by two or more people since the measures are comparative. In practice, this means that the measurements can be applied only to annotations that have been generated for evaluation purposes, so the use of experimental and control sets of papers to blind participants to the knowledge of which annotations are being evaluated should minimize observation effects.

Tests of the Reliability facet through the use of repeated annotation of the same paper at different time points only measures performance at those times on those papers, and extrapolation of those results to the rest of the annotator's work is potentially spurious. The papers chosen for re-annotation may not be representative of either the full collection of papers that curator has annotated, or of the overall performance of the annotator. When measuring both Consistency and Reliability, annotators may be stratified by facets such as years of annotation experience, years of experimental experience, and degree of experience with a particular organism. The evolution of the Gene Ontology vocabularies over time can also influence measurement of the Specificity facet.

External validity (i.e., generalizability) can be limited by the small number of participants. By focusing on GO annotation (which is relatively standardized as an intellectual task across MODs), as distinct from other MOD-specific annotation activities, differences relating to workflows and local customs should be minimized. However, as with internal validity, 'instrumentation' effects, such as differences in workflows or tools used to find and extract assertions from papers, and to find and assign terms from GO, may make cross-MOD comparisons unreliable. Due to their imprecision, these measures should not be used as performance evaluation criteria for employment considerations, either within or across MODs.

### *Contextual factors*

Apart from the quantitative measures defined above, there are many qualitative (or 'contextual') factors influencing human annotation behavior that must be investigated to achieve a fuller explanation of variation. While quantitative methods help answer 'what' questions, contextual approaches investigate 'why' questions: Why do some annotators have high (or low) reliability but low (or high) consistency? What factors influence how some annotators choose broader or narrower terms than other annotators, or have greater (or lesser) completeness than others? Additionally, contextual facets such as gene type, or the relative degree to which some genes have been studied compared to others, may affect annotations. Content analyses of the evolution of certain features of annotations over time may provide additional information.

### *Conclusions and future work*

Greater quantitative and qualitative understanding of all of the above quality facets is important to the utility of model organism databases, in terms of overall information integrity, assistance to curators in performing their work, and the development of automated tools to improve workflows and accelerate the end use and discovery processes. Despite the identified problems, controlled experiments where repeated measures are taken and multiple annotations of the same papers are made may be useful in better understanding human annotation behavior, and why people make certain choices. These experiments are currently being performed by this author with curators from multiple MODs, and this data is being coupled with contextual information from interviews, observations, task analysis, and artifact analysis.

## **ACKNOWLEDGMENTS**

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

## **REFERENCES**

- Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally. Cited in Miller & Salkind, 2002, pp. 50-51 (below).
- Camon EB, Barrell DG, Dimmer EC, Lee V, Magrane M, et al. (2005). An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. *BMC Bioinformatics* 6 Suppl 1:S17. PMID: 15960829.
- Gene Ontology Consortium (2006). The Gene Ontology (GO) project in 2006. *Nucleic Acids Research*, 34: D322-D326. PMID: 16381878.
- GO (2006). Gene Ontology web site. Table of current annotations by MOD. Available: <http://geneontology.org/GO.current.annotations.shtml> (Accessed: 2006-02-24.)
- MacMullen, W.J. (2006). Quantifying literature citations, index terms, and Gene Ontology annotations in the *Saccharomyces* Genome Database to assess results-set clustering utility. Submitted to the American Society for Information Science and Technology (ASIS&T) 2006 Annual Meeting.
- MacMullen, W.J. (2005). Inter-database annotation linkages in model organism databases. In *Proceedings of the 68th Annual Meeting of the American Society for Information Science & Technology (ASIS&T)*, Vol. 42.
- Miller, D. C., & Salkind, N. J. (2002). *Handbook of research design and social measurement* (6th ed.). Thousand Oaks & London: Sage Publications.

# Facets and measures of Gene Ontology annotation quality in model organism databases

W. John MacMullen

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

## Abstract

Model organism databases are important repositories of data and information for biomedical research, but are useful to scientists only if the information they contain meets certain levels of quality. This project describes six facets of information quality applicable to Gene Ontology (GO) annotations in model organism databases, and defines corresponding metrics to be used in measuring the types and degrees variation of GO annotations made by one or more human database curators. The defined facets and measures of annotation quality are: Consistency, Reliability, Specificity, Completeness, Accuracy, and Validity. Contextual factors, and factors affecting internal and external validity, are also discussed.

## GO ANNOTATION QUALITY FACETS

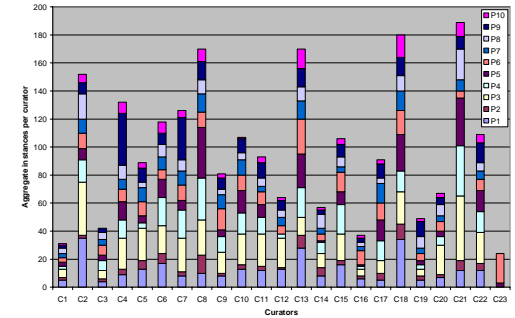
Facet	Definition	Scope	Variable types
Consistency	Variation across annotations made by different annotators	2 or more individual annotations	discrete (nominal)
Reliability	Variation across annotations made by same annotator	2 or more individual annotations	discrete (nominal)
Specificity	Relative granularity of annotation terms in relation to their source vocabulary	2 or more individual annotations	discrete (nominal) & interval (ordinal)
Completeness	Presence / absence of values for attributes in relation to standard annotation	1 or more individual annotations compared to standard (consensus)	discrete (nominal)
Accuracy	Variation of individual annotations in relation to standard annotation	1 or more individual annotations compared to standard (consensus)	discrete (nominal)
Validity	Variation of attribute values of individual annotations in relation to the GO annotation model	1 annotation compared to GO data model	discrete (nominal)

## Definitions

- Annotation** – The set of annotation instances made from a single evidence source (e.g., a scientific article)
- Annotation instance** – An individual claim annotated to a particular GO aspect (Molecular Function, Biological Process, or Cellular Component).
- A particular paper may have 0-n annotation instances, and each may be made to any one (but only one) of the three GO aspects.

## PRELIMINARY DATA FROM A MULTI-MOD RANDOMIZED CONTROLLED STUDY (23 CURATORS, 10 PAPERS)

### Annotation instances per curator by paper



## ATTRIBUTES AND VALUES OF QUALITY MEASURES

### Consistency, Reliability, Completeness, Accuracy, and Validity

Attributes	Values
Overall agreement	{0,1}
Gene product	{0,1}
GO ID	{0,1}
GO term	{0,1}
Evidence	{0,1}
Qualifier	{0,1,-}
With/from	{0,1,-}

### Specificity

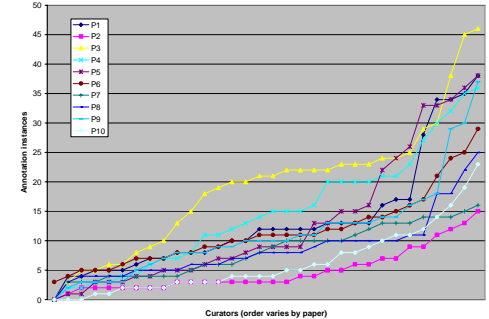
Attributes	Values
Exact match	{0,1}
Mismatch, same lineage	{0,1}
Mismatch, different lineage	{0,1}
Broader	{0,1}
Narrower	{0,1}
Difference, in nodes	{0,n}

## Model Organisms

Model organisms are biological organisms which have high research utility due to certain features, such as relative simplicity, small genome size, or functional similarity to aspects of human biology. Within biomedical research they are valued for their use as surrogates for human gene expression analysis. Model organism databases (MODs) provide rich collections of professionally curated information about specific model organisms. MODs which are members of the Gene Ontology Consortium include:

- DictyBase, for the mold *Dictyostelium discoideum*
- Flybase, for the fruitfly *Drosophila melanogaster*
- MGD, the Mouse Genome Database, for *Mus musculus*
- RGD, the Rat Genome Database, for *Rattus norvegicus*
- SGD, the Saccharomyces Genome Database, for *Saccharomyces cerevisiae* (yeast)
- TAIR, the Arabidopsis Information Resource, for *Arabidopsis thaliana* (mustard plant)
- Wormbase, for the roundworm *Caenorhabditis elegans*
- Zfin, the Zebrafish Information Network, for *Danio rerio*

### Individual and consensus annotation instances per paper, in rank order (curators vary)



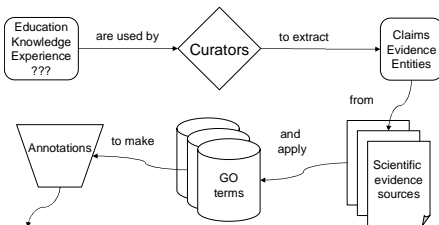
## Acknowledgments

- 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.

Poster and paper available at: <http://macmullen.com/conferences/asist>

IGI with: **Process:** S phase DNA damage checkpoint? RAD53 MRC1 TOF1 CSM3

Example of manual claim extraction from an experimental paper for use in creating a GO annotation (Christie, 2005)



Gene	Ontology	GO ID	GO Term	Evidence	With/From
PSF1	Process	GO:0006261	DNA-dependent DNA replication	IMP	
PSF1	Component	GO:0008111	GINS complex	IP1	PSF2, PSF3, SLD5
PSF2	Process	GO:0006261	DNA-dependent DNA replication	IGI	SLD5
PSF2	Component	GO:0008111	GINS complex	IP1	PSF1, PSF3, SLD5
PSF3	Process	GO:0006261	DNA-dependent DNA replication	IGI	PSF1
PSF3	Component	GO:0008111	GINS complex	IP1	PSF1, PSF2, SLD5
SLD5	Process	GO:0006261	DNA-dependent DNA replication	IMP	
SLD5	Component	GO:0008111	GINS complex	IP1	PSF1, PSF2, PSF3