A characterization of manual literature annotations by biocurators

W. John MacMullen, Sarah J. Meadows, Shu-wen Huang
Graduate School of Library & Information Science
University of Illinois at Urbana-Champaign
Champaign, IL 61820, USA
wjohn@illinois.edu

Abstract
This paper describes work in progress to characterize the manual annotations biocurators make to journal articles when curating the scientific literature using the Gene Ontology. We examined a corpus of 87 experimental journal articles from the fruitfly literature and characterized biocurators’ manual annotations by location within the articles and by annotation type. We observed a total of 5,746 annotations over 800 pages of text. Most annotations were localized to the Results and Methods sections, and most were highlights and checkmarks of text. These findings support our prior work in other model organism literatures, and should be of use to the NLP community for the development of curation support tools.

1 Introduction
Much effort has been expended by the NLP community in creating automated or semi-automated methods for augmenting or replacing human literature curation in biomedical databases, typically using approaches such as named entity recognition (NER) and relation extraction (see, e.g., Chen, Müller & Sternberg, 2006; Crangle et al., 2007; Laskowski, 2007; Ling, et al., 2007). The utility of these approaches to actual biocuration tasks remains largely unstudied, and most do not investigate the underlying cognitive behaviors of human biocurators when developing their methods and systems. Even those that involve human interaction are usually limited to activities such as curators taking the role of relevance judges to create a ‘gold standard’ against which to evaluate system performance (e.g., Krallinger et al., 2008).

A common work task for biocurators (Howe, et al., 2008) is reading narrative journal articles to assess the authors’ claims and the evidence supplied to support them, and capturing this information in distilled, structured secondary formats, such as Gene Ontology (GO) annotations (GO, 2007), genetic- or protein interactions, or phenotypic characterizations. The cognitive activities related to this intellectual process have not been well-studied, and the documentation of the human information interactions between the narrative journal article and the final recording of GO annotations is sparse (MacMullen, 2007).

Manual document annotations here function as one surrogate for more direct investigations of those annotators’ intellectual activities that are ultimately manifested in database records. While many biocurators read electronic versions of articles and annotate directly into curation interfaces, a substantial number of curators still read paper articles and make manual annotations on them as an intermediate step in their curation workflow. Characterization of these manual annotations may help illuminate the origins of curators’ final annotations, and provide explicit links back to specific locations and passages in the primary documents from where the resulting annotations were derived. Characterization of manual annotations is also one
source of evidence for use in creating effective, evidence-based systems for augmenting literature curation, as well as for potentially improving end-user exploration of literature, as Hearst et al. (2007) suggest. Additionally, content- and linguistic analyses may provide basic evidence about the structure and composition of biomedical text that could influence searching, browsing, and information extraction, as Shah et al. (2003) have argued.

To begin to investigate these questions, this project analyzed a corpus of 87 journal articles that were manually annotated by several biocurators from one prominent model organism database in the course of their normal work tasks around GO annotation. Since there is a trend in biocuration toward the reading of electronic articles, it was important that this work be conducted now while manual annotations are still being created. In an all-electronic curation scenario, more of the cognitive processes underlying reading and annotation behaviors could become hidden and more difficult to examine without substantial qualitative investigations, such as task analyses or concurrent verbal report (i.e., ‘think-aloud’) studies.

## 2 Methods

In earlier work (MacMullen, 2007), 23 biocurators from 11 model organism databases curated a common set of 10 journal articles for GO annotations. Two articles from each of 5 model organism literatures were selected by biocurators as being typical of the articles they normally curate. Six biocurators also provided their manually-annotated paper copies of these articles for analysis. A shallow content analysis of the annotations in the article set (60 articles, 615 total pages) was performed, and the manual annotation features were classified by paper section (i.e., Introduction, Methods, Results, etc.) and annotation type (i.e., underlining, highlighting, etc.) These classifications were informative and were applied here, as described below.

For this project, 87 experimental journal articles that were manually annotated by biocurators in 2007 were provided in paper form to the authors by the FlyBase model organism database (Tweedie, et al., 2009). While originating in different journals, all were associated with *Drosophila melanogaster* (fruitfly) biology and genetics. No selection criteria were applied; the set was composed of all articles curated by numerous FlyBase biocurators in late 2007. The articles were digitized, yielding 800 scanned pages. The PubMed unique identifiers (PMIDs) for the papers were obtained, and the full MEDLINE citations were downloaded using the NCBI’s eFetch scripts. The FlyBase GO gene association file was obtained, and both that file and the citations from the annotated collection were loaded into a purpose-built MySQL database for bibliographic analysis.

Two coders (the second and third authors) counted and classified the manual annotations present in the papers. The coders were given instructions on how to count annotations, and how to identify annotations by type, and completed a coding exercise using annotated papers from a different corpus. Consistency was calculated using the intra-class correlation (ICC) function of the R package 'psy' (Falissard, 2008). For the full paper set, inter-coder reliability was evaluated by randomly assigning them 18 papers in common out of the 87 (20% redundancy). The coders did not know which of their assigned papers were in the reliability subset, and their coding was performed independently of one another. An evaluation of the two raters’ aggregate annotation counts for the 18 papers using ICC showed agreement of 0.946.

This project was reviewed and approved by the University of Illinois Institutional Review Board as being exempt from human subjects review as a retrospective study (IRB Protocol # 08587).

## 2.1 Supplementary Data

The citations for the 87 articles analyzed in this study are available online as supplementary data, along with additional data: [URL removed for review]

## 3 Results

### 3.1 Bibliographic analysis

The 87 annotated articles came from 24 distinct journals. Fourteen of the 24 were single occurrences, meaning that the remaining 73 articles came from 10 journals. The top three journals account for more than 50% of the 87 articles. These

---

1. [http://pubmed.gov](http://pubmed.gov)
data were contrasted with the distribution of articles and journals in the overall FlyBase gene association data file, which had 4,555 distinct articles with PubMed identifiers. There were 205 distinct journals with PubMed IDs, and there is substantial overlap in the lists of journals from the gene association file and the sample.

### 3.2 Total annotations by paper

The papers ranged in length from 3 to 18 pages (mean: 9.2, median: 10, mode: 6). The coders observed a total of 5,746 manual annotations over all papers, ranging from a low of 6 to a high of 221 (mean: 66, median: 59). Plotting numbers of pages against total annotations per paper yields an $R^2$ value of -0.0029, which indicates that paper length is not a good predictor of numbers of annotations made.

### 3.3 Annotations by paper section

The earlier work in MacMullen (2007) explored manual annotations by paper section. We used the same section typology here, which consisted of Abstract, Introduction, Methods, Results, Discussion (or, alternatively, a combined ‘Results and Discussion’ section), Acknowledgements, Supplementary Materials, and References. Some papers used alternate language for sections equivalent to these (e.g., ‘Background’ for Introduction, ‘Materials and Methods’ for Methods), but since the document set was composed of experimental articles, and did not include other article types, such as reviews, the overall structures of the papers were substantially similar, following the ‘IMRAD’ model (Sollacci & Pereira, 2004). We also counted numbers of annotations made to figures and tables, to article titles, and within whitespace areas of the articles, and counted them as being associated with the sections in which they appeared.

In the aggregate, the Results sections of papers had the highest numbers of annotations, with a cumulative 3,009, or 3,718 when Results and Discussion annotations are included (65% of IMRAD, min: 9, max: 178, mean: 43, median: 37). The Methods section was next with 1,347 annotations (23% of IMRAD, min: 0, max: 84, mean 17, median: 11). As Figure 1 shows, the other sections had substantially fewer annotations. Annotations to Figures and Tables accounted for 767 and 205 annotations, or 13.3% and 3.6% of total annotations, respectively. Not all papers had annotations in every section.

### 3.4 Annotations by type

We used the typology from MacMullen (2007) when classifying annotations made by biocurators. The 7 types of annotations observed in that work were Highlight, Underline, Circle, Emphasis, Arrow, Check mark, and Textual note. We observed each of those types to varying degrees in the current document set, and no new types. Figure 2 shows that Highlights are by far the most frequently observed annotations (3,239 or 56% of all annotations), with Check marks making up one quarter of the total. Not all curators used every annotation type.

4 **Discussion**

These results from one model organism confirm our earlier findings from several different model organism literatures.
That most manual annotations were found in the Results and Methods sections is consistent with the fact that GO annotations are meant to identify and formalize authors’ claims about gene products and the evidence supporting those claims. In interviews with curators in MacMullen (2007), the abstract was cited as a locus for the most important claims in the article, and was usually the first place curators looked for claims and evidence. The proportionately smaller number of manual annotations from abstracts in this document corpus could be explained by the smaller size of abstracts relative to the full Methods and Results sections. In the same interviews, curators described Introduction sections as containing claims by other authors, and Discussion sections as containing more speculative and unsupported statements, which is consistent with the much smaller annotation counts observed here for those sections.

Our results show that the types of annotations made by biocurators and the locations where annotations appear within Drosophila papers are consistent with our earlier work on biocurators in multiple literatures. We do not attempt to interpret the semantic meaning of the annotation types in this paper.

4.1 Future work

We are currently extending this work by examining the text of the corpus articles from deeper semantic and syntactic perspectives, in order to explore the claims, entities, and evidence the curators annotated, compared to the other claims within the papers, at section, paragraph, phrase, and word levels. For the corpus from MacMullen (2007), we are performing a 6-way comparison of the biocurators’ independently-derived manual annotations to investigate whether common types of semantic and syntactic constructions are identified by curators as being important sources of GO annotations. These analyses may have implications for the design of methods for claim identification and extraction, and other approaches to curation automation.

Acknowledgments

This project was funded by a Seed Research Grant from the Graduate School of Library and Information Science at the University of Illinois. We thank Steven Marygold from the FlyBase resource for making the document corpus available to us.

References


