tagtog
Interactive Human & Machine Annotations
Team #176

Juan Miguel Cejuela, juanmi@tagtog.net
Peter McQuilton, pam51@gen.cam.ac.uk

June 7, 2013

Abstract

This is a proposal submission to the BioCreative IV Track 5 – User Interactive Task (IAT). We present the *tagtog* system. tagtog is a web-based annotation framework. It leverages user manual annotations in combination with automatic machine-learned annotations. For this submission, we present in collaboration with the FlyBase database at Cambridge University, the task of identifying and extracting mentions of *drosophila* genes in full-text biomedical articles.

1 tagtog: System Description

tagtog (www.tagtog.net), is a web-based framework for the annotation of named entities. A user creates a project, defines a named-entity recognition task, and uploads a set of text documents to the system. Each document is then displayed in a web editor where the user can add, delete, or correct the information relevant to the annotation task. An example of the user interface is shown in Figure 1. The user can add the annotation of an entity by selecting the corresponding word(s) and delete it by clicking again on the selection. The application case for the user is to be able to extract, share, and exploit the annotated information. During the course of their work, the user needs to analyze thousands or even millions of documents, an undertaking that is impossible to do through manual means alone. To address this problem, the tagtog system leverages machine learning methods to perform the same type of annotations. Initially, the tool is trained with a small set of (user) manually-annotated documents. The tool is then ran on novel documents to generate automatic predictions that can be reviewed and corrected by the user. It is this continuous and interactive re-training of the machine learning methods with user feedback, that leads to an ever-improving performance in automatic prediction. Once optimized, the trained machine learning methods can be used to process and annotate a large volume of documents to a sufficiently accurate level. Finally, the annotated doc-

uments can be exported (in XML format) for the particular user's application needs.

In the following Section 1.1, we describe the system's current features and technical details. In Section 1.2 we describe the planned features to be ready by the curation process task proposed in this submission to the BioCreative IV track 5 – User Interactive Task (IAT).

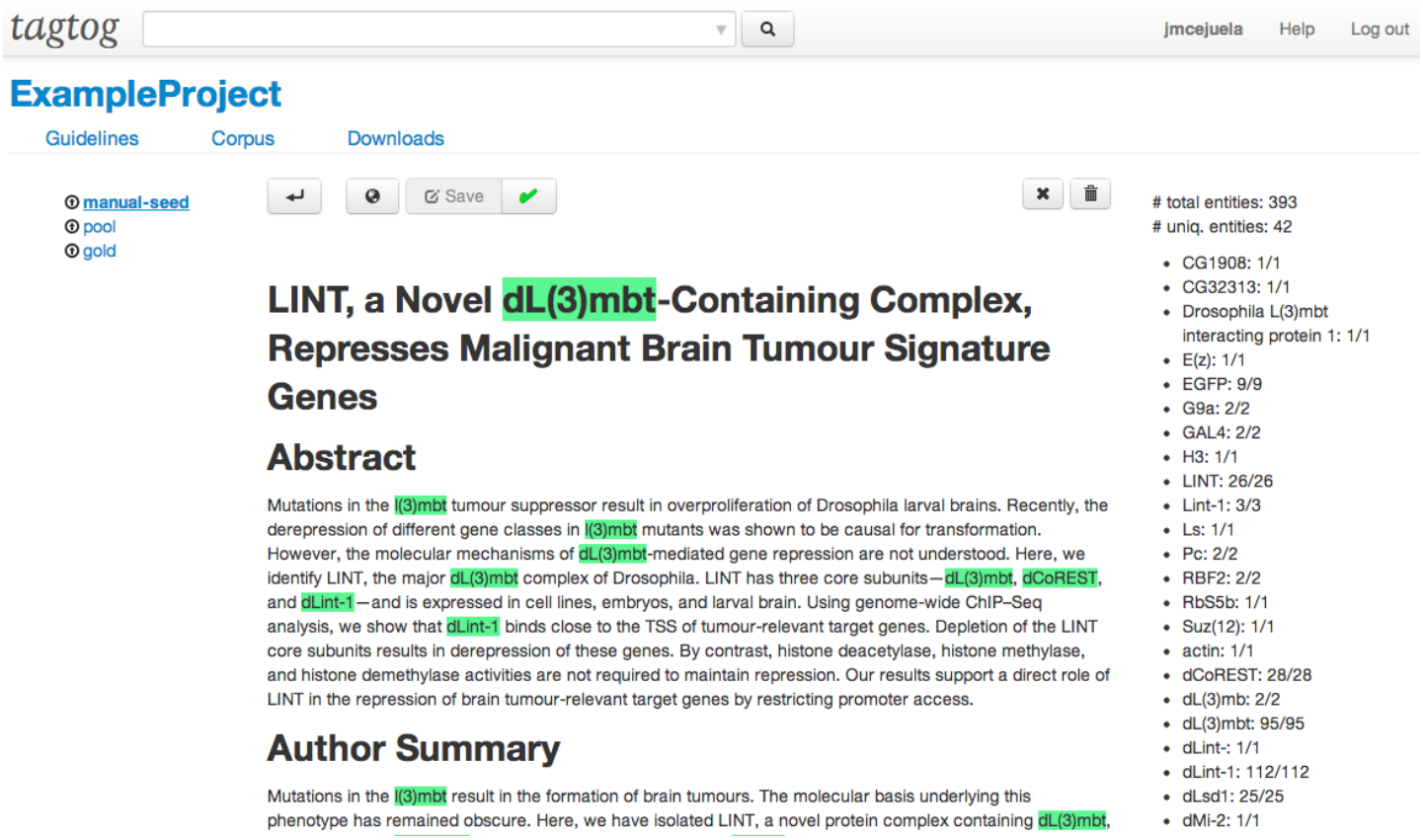


Figure 1: Example of the document display and editor in tagtog.

1.1 Current Features

The system (www.tagtog.net) runs on all major current browsers only requiring HTML5 and javascript. Chrome and Firefox are officially supported. Other browsers like Opera, Safari, and Internet Explorer (9 and 10) are regularly tested but lack official support. Access to the system is currently granted to hand-picked users. The BioCreative IV Track 5 committee are welcome to access the system. Users can create different annotation projects. Support for multiple users working on a same project is planned to be incorporated for the workshop evaluation, see Section 1.2.

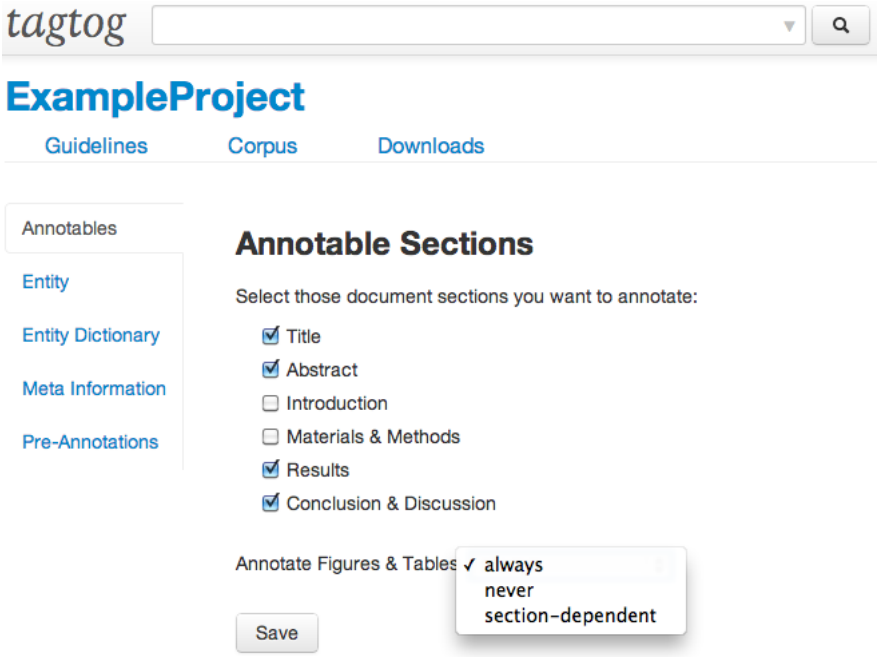


Figure 2: Project Guidelines.

1.1.1 Guidelines

Upon project creation, the first step for a user is to define the annotation guidelines, see Figure 2. The user currently has the following options:

- Annotables: select the sections of the full-text articles that can be annotated (and trained with). The annotation of figures' and images' captions is decided independently: *always*, *never*, or *section-dependent* (annotable if the enclosing section is annotable).
- Entity: choose the name of the entity to be annotated.
- Entity Dictionary: upload an user-defined dictionary/ontology of collected entity names. Users can for each entity indicate the corresponding unique id of their private databases for seamless integration, list recommended and alternative names.
- Meta Information: define a list of checkboxes to introduce for each document, e.g., whether the article contain disease mentions, yes/no.
- Pre-Annotations: if activated, upon entity selection or deselection, pre-annotate automatically similar names and require the user's confirmation for final acceptance.

Figure 3: Example of Pre-Annotations. Upon annotation of a single entity, similar names are automatically pre-annotated and require the user's confirmation (notice the check mark button for the hovered entity).

1.1.2 Corpus

The next step is to define the text corpus for annotation. For this, the user currently has the possibility to upload on a single or batch basis, any full-text XML document that follows the NCBI Journal Publishing Tag Set format (versions 2.x and 3.0) [1]. This format supports all open access PubMed Central [2] or PLOS [3] articles. The system's internal parser recognizes the documents' sections, subsections, figures, tables, and some extra meta information such as the paper's original URL. The project corpus can be augmented progressively as the user sees fit. Currently, documents are placed in 3 different folders, namely, *manual-seed*, a small set (∼20) of manually-annotated documents for the early training of the machine learning methods, the *gold* folder for manually-annotated documents for the evaluation of the machine learning methods' performance (these documents are never used for training), and the *pool* folder where all other documents are placed. The pool folder is iteratively polled by the machine learning methods to pick documents to be trained with (if the document's annotations were confirmed by the user) or to make predictions on new documents.

The system offers a search system to find relevant documents, see Figure 4. The current search fields are: words in the full text (with support for boolean logic), annotation complete (yes/no), document id, and annotated entities.

1.1.3 Download

The user can export the entire corpus upon request. Currently, documents are annotated in XML documents with an in-house-defined syntax, called *anndoc*. The current anndoc version (0.3) supports entity in-line annotations (without

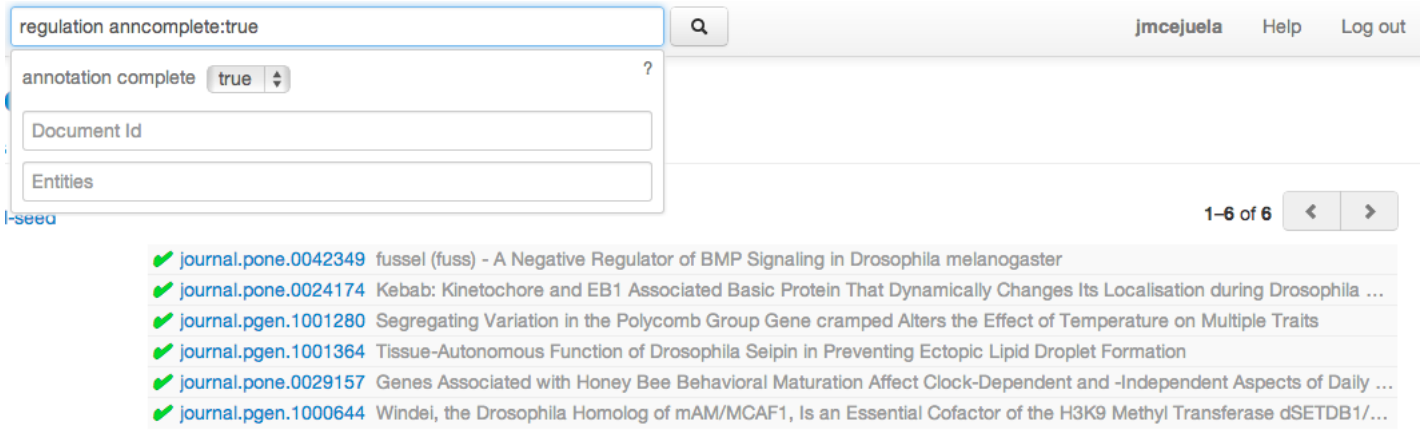


Figure 4: Document search and listing.

being nested), meta information, hierarchy of sections, figures, tables, and their captions. It follows a HTML-like format, with a header for meta information and a body for the document's text and annotations. The tags have been so selected to match those of HTML5's for an easy display of the document in current browsers. An extract example of the anndoc format can be seen in Figure 5. In the future, we are keen to adopt other standard formats if there is sufficient user demand.

```
<anndoc origid="journal.pone.0040912" i="aIxUwUH0_o.Q3LkJTSaRWjKXi5UO-journal.pone.0040912">
▼<div class="anndochead">
   <meta name="source" content="http://dx.doi.org/10.1371/journal.pone.0040912"/>
   <meta name="parser" parsername="com.jmcejuela.bio.NcbiJournalArticleParser$" version="0.1"/>
   <meta name="anncomplete" content="false"/>
 </div>
▼<div class="anndocbody">
  ▼<div class="section" depth="1" type="title">
    ▼<h2>
       Similarities of Drosophila rab GTPases Based on Expression Profiling: Completion and Analysis of the
       <span title="(PRE-SELECTED) Annotated by: user:jmcejuela" who="user:jmcejuela" class="selected ner_genprot">rab-
       Gal4</span>
       Kit
     </h2>
   </div>
  ▼<div class="section" depth="1" type="abstract">
     <h2>Abstract</h2>
    ▼<div class="content">
      ▼<p>
         We recently generated
         <span who="user:jmcejuela" class="selected ner_genprot pre-selected">rab-Gal4</span>
         lines for 25 of 29 predicted Drosophila rab GTPases. These lines provide tools for the expression of reporters,
         mutant rab variants or other genes, under control of the regulatory elements of individual rab loci. Here, we
         report the generation and characterization of the remaining four
         <span who="user:jmcejuela" class="selected ner_genprot pre-selected">rab-Gal4</span>
```

Figure 5: Export anndoc XML format.

1.1.4 Machine Learning

A core defining characteristic of the system is that the users can choose the entity type to annotate, this not being predefined. The system boasts a general-purpose named-entity recognizer (nevertheless better suited for the biomedical domain and the English language). The methods are customized to the prediction task at hand by means of the user annotations and by a submittable ontology of entity terms, see Section 1.1.1. The system is so configured to later on be easily expanded with new machine annotators via plugins and so better adapt to more specific tasks and to new language domains.

For this submission to the BioCreative Track 5, we propose in Section 2 the recognition of *drosophila* genes (thus organism-specific). To demonstrate the flexible capabilities of the machine learning system, tagtog will also be used to participate at the BioCreative Track 2 (*CHEMDER*) for the recognition of chemical compounds and drugs.

In addition, if desired by the user, the machine learning component can be turned off to work exclusively with the manual document editor interface.

1.2 Planned Features

The following are upcoming features that are needed and will be ready by the start of the Biocuration curation process in (planned) August:

- *Export annotations as tab-separated list of terms linked to PMIDs.*
- *Multi-user*: support for multiple users adding and correcting annotations in the same project.
- *Active learning*: actively ask for user's feedback for those predicted annotations that the machine methods were least sure about in order to train with interesting information and free the user of time-consuming non-relevant revisions. A proposed mechanism was already developed in an early version of tagtog, presented last year at the BioCreative 2012 workshop [4].

Support for the annotation of *multi-entities* in the same project is also desired and expected to be ready by the BioCreative 2013 conference in October.

2 Proposed User Curation Task

In collaboration with the FlyBase database [5, 6] at Cambridge University we propose the task of identifying and extracting mentions of genes of the drosophila genus (fruit flies) in full-text biomedical articles. Starting in September 2012, the FlyBase database has been testing tagtog for its adoption in their curation pipeline. One member of FlyBase, the first proposed biocurator for this task, used tagtog to prepare a set of 139 full-text articles with manual and automatic annotations. The system's performance with this set was evaluated, see Section 3. Currently, FlyBase has two well-defined application cases for tagtog:

1. To link Drosophila genes and publications. The planned feature of a tab-separated export of annotated entities will be used for this, see Section 1.2.

2. To skim-curate (i.e. generate gene-to-publication links and triage the data within the paper) new documents to enable document triage and the formation of curation lists. This is integral to the FlyBase curation pipeline and will use the meta information checkboxes described in the Guidelines, Section 1.1.1.

The proposed biocuration task will consist of primarily 1) expanding the corpus set with more annotated documents, especially machine-predicted documents, 2) improving the performance of the machine learning methods, and 3) showing how FlyBase could integrate tagtog with their curation pipeline, addressing specific practical details such as how to consume the generated output formats, what is the subjective user experience with the tool on a daily basis, and what short-comings and improvements can be identified.

We propose the following two biocurators from FlyBase at Cambridge University to perform the annotation and evaluation experiments:

1. Peter McQuilton. He already prepared a set of 139 annotated documents with tagtog and is well acquainted with the tool. He therefore will not require any adaptation time.

2. *Undecided.* The group is committed to propose a second biocurator. This user will be new to the tool and will require some adaptation time.

3 System Performance

The tool has undergone two benchmark iterations with different document sets. The documents used are all full-text articles and were collected from different PLOS journals, namely, *ONE* [7], *Biology* [8], and *Genetics* [9], for publication dates between 2011 to 2013. The following document sections were annotated: title, abstract, results, materials and methods, and figure and image legends. Other sections (for example, the introduction) were not annotated and therefore not considered for the evaluation results. Annotations were done partially manually by a sole curator (Peter McQuilton) and automatically by the system. All the manual annotations and corrections were performed using tagtog's document editor interface. The user also made use of the semi-automatic pre-annotations feature, see Section 1.1.1.

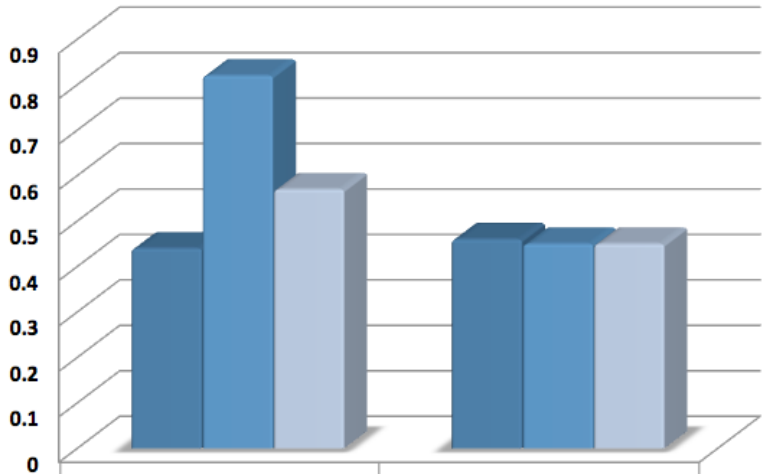
The document sets of the two iterations are the following:

1. Iteration 1: the curator first manually annotated a *training set of 20 articles*. Trained with these documents, the system was applied to predict an unlabeled *test set of 99 articles* (the user submitted 100 test articles but one was rejected by the system as it recognized it as a duplicate from the training set). The curator then went through the test set and, corrected, added, or removed the predicted annotations when appropriate. In the end, mismatched annotations between the original predictions and the revised annotations were counted as errors.

2. Iteration 2: the previous two sets were united to form a *training set of 119 articles*. As just described, this set contained automatic predictions although revised and confirmed by the user. They were all considered as

ground truth. For evaluation, the user manually annotated a *test set of 20 articles*. The system was trained on the 119 articles and benchmarked against the 20 test articles. In contrast to Iteration 1, in this case prediction errors could be read off directly from mismatches against the test set.

Standard evaluation measures for named-entity recognition (NER), namely, precision (P), recall (R), and F1-Measure (F1) were used. Only exact matches between the predictions and the test annotations were counted as correct. That is, the predictions had to match the same exact word boundaries. The correct and erroneous annotations were collated on a document basis and finally averaged for all documents. Two types of counts were considered: 1) only unique entities on a document basis. That is, for an entity X, the predictions were right if it at least one mention of that entity could be identified, wrong otherwise. Equivalently, all (unique) entities identified by the predictions but not present on the test annotations were counted as errors. 2) All entity mentions on a document basis. That is, for all entity mentions, matching predictions and test annotations were counted as correct. Mismatched mentions, either false positives or false negatives, were counted as errors.

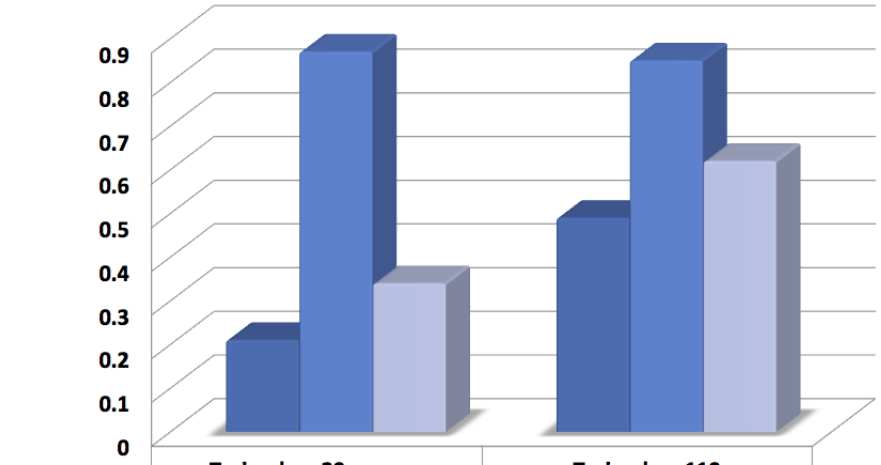


	Trained on 20 papers, tested on 99 papers	Trained on 199 papers, tested on gold standard (20)
Recall	0.44	0.46
Precision	0.82	0.45
F1-measure	0.57	0.45

Figure 6: Entity recognition performance for unique entities, i.e., the ability to identify the presence of a gene at least once within a paper.

Figure 6 shows the performance results for Iteration 1 and 2 for the unique entity counts. Figure 7 shows the performance results for Iterations 1 and 2 considering the counts of all mentions. The performance for unique entities, that is, the ability to identify the presence of a gene at least once within a

paper, appeared to remain the same or slightly drop from Iteration 1 (less training instances) to Iteration 2 (more training instances). The performance for all mentions, that is, the ability to identify a gene every time it is mentioned within a paper, appeared to considerably increase from Iteration 1 (F1 = 0.34) to Iteration 2 (F1 = 0.62).



	Trained on 20 papers, tested on 99 papers	Trained on 119 papers, tested on gold standard (20)
Recall	0.21	0.49
Precision	0.87	0.85
F1-measure	0.34	0.62

Figure 7: Entity recognition performance for all mentions, i.e., the ability to identify a gene every time it is mentioned within a paper.

Although overall a moderate performance, we see these early evaluation results as promising. First of all, to our knowledge these results represent one of the first NER evaluations with a substantial amount of full-text articles in the biomedical field. NER with full-text articles is deemed considerably more difficult than for abstracts, a problem more studied in the past [10, 11]. Second of all, it must be noted that the machine learning method employed was but for the ontology of terms (given by the user) not specialized to the problem. In particular, there was no special treatment for the fact that only genes for a single organism, drosophila, should be annotated. Finally, although not yet for unique entities, prediction performance appears to increase with an increase in the volume of training data and there seems to be more room for improvement, mainly for recall. The continuous learning of tagtog is designed to generate cheaper (as for manual curation effort) training data, by taking advantage of semi-automatically annotated data. Still, a central goal for the curation task proposed in this submission, see Section 2, is to further improve the performance of the tagtog system so that FlyBase can incorporate it into their pipeline at a sufficiently accurate level.

Notice

tagtog is privately developed and funded by Juan Miguel Cejuela. Future commercial support is planned.

References

[1] Journal Publishing Tag Set. URL `http://dtd.nlm.nih.gov/publishing/`.

[2] PubMed Central. URL `http://www.ncbi.nlm.nih.gov/pmc/`.

[3] PLOS. URL `http://www.plos.org/`.

[4] C. N. Arighi, B. Carterette, K. B. Cohen, M. Krallinger, W. J. Wilbur, P. Fey, R. Dodson, L. Cooper, C. E. Van Slyke, W. Dahdul, P. Mabee, D. Li, B. Harris, M. Gillespie, S. Jimenez, P. Roberts, L. Matthews, K. Becker, H. Drabkin, S. Bello, L. Licata, A. Chatr-aryamontri, M. L. Schaeffer, J. Park, M. Haendel, K. Van Auken, Y. Li, J. Chan, H. M. Muller, H. Cui, J. P. Balhoff, J. Chi-Yang Wu, Z. Lu, C. H. Wei, C. O. Tudor, K. Raja, S. Subramani, J. Natarajan, J. M. Cejuela, P. Dubey, and C. Wu. An overview of the BioCreative 2012 Workshop Track III: interactive text mining task. *Database (Oxford)*, 2013:bas056, 2013.

[5] McQuilton P, St Pierre SE, Thurmond J; FlyBase Consortium. FlyBase 101–the basics of navigating FlyBase. *Nucleic Acids Res.*, 40(Database issue):D706–714, Jan 2012.

[6] McQuilton P; FlyBase Consortium. Opportunities for text mining in the FlyBase genetic literature curation workflow. *Database (Oxford)*, 2012:bas039, 2012.

[7] PLOS ONE. URL `http://www.plosone.org/`.

[8] J. A. Eisen. PLoS Biology 2.0. *PLoS Biol.*, 6(2):e48, Feb 2008.

[9] PLOS Genetics. URL `http://www.plosgenetics.org/`.

[10] Larry Smith, Lorraine Tanabe, Rie Ando, Cheng J. Kuo, Fang I. Chung, Chun N. Hsu, Yu S. Lin, Roman Klinger, Christoph Friedrich, Kuzman Ganchev, Manabu Torii, Hongfang Liu, Barry Haddow, Craig Struble, Richard Povinelli, Andreas Vlachos, William Baumgartner, Lawrence Hunter, Bob Carpenter, Richard Tsai, Hong J. Dai, Feng Liu, Yifei Chen, Chengjie Sun, Sophia Katrenko, Pieter Adriaans, Christian Blaschke, Rafael Torres, Mariana Neves, Preslav Nakov, Anna Divoli, Manuel M. Lopez, Jacinto Mata, and John W. Wilbur. Overview of BioCreative II gene mention recognition. *Genome Biology*, 9(Suppl 2), 2008.

[11] Zhiyong Lu, Hung Y. Kao, Chih H. Wei, Minlie Huang, Jingchen Liu, Cheng J. Kuo, Chun N. Hsu, Richard Tsai, Hong J. Dai, Naoaki Okazaki, Han C. Cho, Martin Gerner, Illes Solt, Shashank Agarwal, Feifan Liu, Dina Vishnyakova, Patrick Ruch, Martin Romacker, Fabio Rinaldi, Sanmitra Bhattacharya, Padmini Srinivasan, Hongfang Liu, Manabu Torii, Sergio Matos, David Campos, Karin Verspoor, Kevin Livingston, and W. Wilbur. The gene normalization task in BioCreative III. *BMC Bioinformatics*, 12(Suppl 8):S2+, 2011. ISSN 1471-2105. URL `http://dx.doi.org/10.1186/1471-2105-12-S8-S2`.