

# **PubTator: A PubMed-like interactive curation system for document triage and literature curation**

Chih-Hsuan Wei<sup>1,2</sup>, Hung-Yu Kao<sup>2</sup>, Zhiyong Lu<sup>1,\*</sup>

<sup>1</sup>National Center for Biotechnology Information (NCBI), 8600 Rockville Pike, Bethesda, MD, 20894

<sup>2</sup>Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, R.O.C

\*Contact: [zhiyong.lu@nih.gov](mailto:zhiyong.lu@nih.gov)

## **1. Introduction**

PubTator is a Web-based tool that allows curators to create, save, and export annotations. As shown in our past study [1], manual curation can greatly benefit from (semi-)automated computer analysis. Hence, PubTator is equipped with multiple advanced computer algorithms for assisting two specific curation tasks: a) document triage and b) bioconcept annotation (e.g. genes).

PubTator is developed based on a prototype system that was previously used at the NCBI for various manual curation projects such as annotating disease mentions in PubMed abstracts. In response to call for participation in BioCreative 2012, we significantly extended our previous system in developing PubTator. First, relevance ranking and concept highlighting were added to ease the task of document triage. Second, state-of-the-art named entity recognition tools (e.g. winning gene normalization systems [2,3] in BioCreative III) were integrated to pre-tag bioconcepts of interest, as a way to facilitate the task of gene/disease/chemical annotation. Third, PubTator was developed to have a look-and-feel similar to PubMed, thus minimizing the learning efforts required for new users. Furthermore, a standard PubMed search option is made available in PubTator, which would allow our users to make a hassle-free move of their saved PubMed queries (a common practice for curators doing document triage) into this new curation system. Finally, by taking advantage of pre-tagging bioconcepts, PubTator also allows its users to do semantic search besides the traditional keyword based search, a novel feature not available in PubMed.

## **2. System description**

### **2.1 PubTator search page**

For the convenience of many PubMed users, by default PubTator allows the same search syntax and returns identical search results as PubMed. This is achieved by using the Entrez Programming Utilities Web service API.

In addition to the traditional keyword search, an advanced semantic search is featured in PubTator, which enables our users to retrieve articles associated with specific semantic bioconcepts. In the current implementation, a user can choose from one the three semantic categories: gene/proteins, diseases, and chemicals. This is to specifically address a known problem in biomedical literature search: a bioconcept is often associated with multiple different names. When using the semantic search, a user can retrieve all the papers relevant to a concept

without having to enumerate the entire set of possible aliases. For instance, searching for the breast cancer gene HER2 will also retrieve articles only mentioning its alternative names such as NEU or ERBB2 (e.g. See result #8 in Figure 2).

[Login by curator](#)

# PubTator



PubTator is a Web-based tool for assisting two specific biocuration tasks: 1) document triage, and 2) bioconcept annotation. It supports the standard PubMed search as well as advanced semantic search of gene, disease and chemical concepts. Click [here](#) for more information.

© 2011 National Center for Biotechnology Information (NCBI), U.S. National Library of Medicine  
8600 Rockville Pike, Bethesda MD, 20894 USA

**Figure 1:** The PubTator homepage.

Finally, we also provide a search option of using a list of PubMed identifiers (PMIDs). This is desirable when one or more articles have been judged to be relevant and need to be curated.

The link in the upper right hand corner is for the user to sign in. Once signed in, it will show the name of the curator (See Figure 2).

## 2.2 PubTator results page

Following the tradition of PubMed, by default PubTator returns search results in the reverse chronological order. However, only 15 results are returned per page in PubTator vs. 20 in PubMed, making room for quickly displaying the abstract. As shown in Figure 2, a user can click the ABSTRACT link below the PMID to take a peek at the abstract without having to go to a separate abstract page.

As shown in Figure 2, relevance-based ranking is an alternative option in PubTator when a curation team provides PubTator with their curation guidelines and training data (e.g. CTD data in BioCreative III Track I). In such cases, we will pre-compute a relevant score for each candidate article by using machine-learning algorithms (e.g. SVM) [4]. Next, the computed scores will be normalized and subsequently used for ranking search results.

As a novel feature to help document triage, we also highlight key concepts in the title and abstract. Currently, four different concepts are pre-annotated and highlighted: Gene (purple), Chemicals (green), Diseases (orange), and Species (blue).

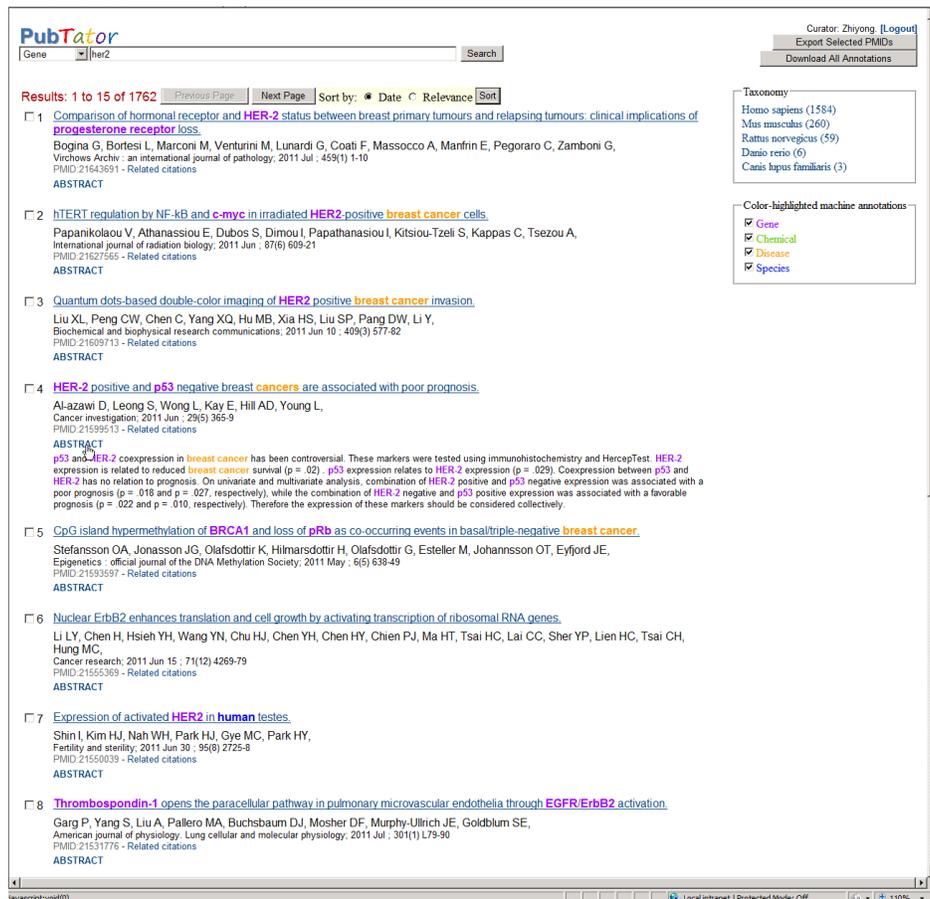


Figure 2: The PubTator results page.

To the right of the search results, we show two advanced search options. On the top panel, users can refine their search results by taxonomy. This feature is useful for those curation teams who work with a specific organism because by default we show results across all species. In the lower panel, one can choose to turn off one or more highlighted concepts if desired.

For the document triage task, a curator can select the relevant papers from the search results by simply checking the box next to its number. To further examine an article or perform the detailed annotation task, a curator then needs to go to the abstract page as described below.

### 2.3 PubTator abstract page

When an article title is clicked in the results page, PubTator returns its abstract page in response. Concepts are annotated in this page as follows: 1) a piece of text is color-highlighted and assigned to a semantic category; and 2) a standard database identifier is searched and assigned to the selected text mention.

STATE: Not curatable    Gene  Chemical  Disease  Species

PMID:21599513 **HER-2 positive and p53 negative breast cancers are associated with poor prognosis.**  
 Author: Al-azawi D, Leong S, Wong L, Kay E, Hui AD, Young L,  
 Publication: Cancer investigation, 2011 Jun ; 29(5) 365-9

TITLE:  
**HER-2 positive and p53 negative breast cancers** are associated with poor prognosis.  
 ABSTRACT:  
**p53** and **HER-2** coexpression in **breast cancer** has been controversial. These markers were tested using immunohistochemistry and HercepTest. **HER-2** expression is related to reduced **breast cancer** survival (p = .02). **p53** expression relates to **HER-2** expression (p = .029). Coexpression between **p53** and **HER-2** has no relation to prognosis. On univariate and multivariate analysis, combination of **HER-2** positive and **p53** negative expression was associated with a poor prognosis (p = .018 and p = .027, respectively), while the combination of **HER-2** negative and **p53** positive expression was associated with a favorable prognosis (p = .022 and p = .010, respectively). Therefore the expression of these markers should be considered collectively.

Type	Mention	Identifier	Nonclementure
Gene	HER-2	2064	NCBI Gene
Gene	p53	7157	NCBI Gene
Disease	cancers	D009369	MeSH
Gene	p53	7157	NCBI Gene
Gene	HER-2	2064	NCBI Gene
Disease	breast cancer	D001943	MeSH
Gene	HER-2	2064	NCBI Gene
Disease	breast cancer	D001943	MeSH
Gene	p53	7157	NCBI Gene
Gene	HER-2	2064	NCBI Gene
Gene	p53	7157	NCBI Gene
Gene	HER-2	2064	NCBI Gene
Gene	HER-2	2064	NCBI Gene
Gene	p53	7157	NCBI Gene
Gene	HER-2	2064	NCBI Gene
Gene	p53	7157	NCBI Gene

**Figure 3:** The PubTator abstract/annotation page.

As shown in Figure 3, at the very top of the page, the paper’s current curation status is shown (curatable or not). Clicking on the button next to it will readily change its status, giving our user an option to perform document triage also in the abstract page. Immediately below is some publication metadata including the PMID, title, author(s), journal, and publication date.

Under the metadata, the title and abstract are displayed in a text box where a user can manipulate annotations in a number of different ways:

1. To create an annotation: selecting a piece of text and click one of the four semantic categories (e.g. gene).
2. To remove an annotation: selecting an existing annotation and click ‘Clear’.
3. To reset annotations: by clicking ‘Reset’, system will return to the results that were last modified.
4. To commit annotations: by clicking ‘Confirm’, all highlighted text mentions will be added to the Table immediately below where the second step of concept annotation—assigning the concept id for the highlighted textual mention—is required.

To facilitate the concept annotation process, we pre-tag all concepts in the title and abstract using state-of-the-art text mining tools (See more in Section 5). However, if one or more concept categories are not needed for a specific task, those pre-computed concepts could be removed by clicking the x icon in front of the corresponding category. For the sample article shown in Figure 3, most machine generated annotations are correct; the only manual work is to correct an

annotation in the title (change ‘cancers’ to ‘breast cancers’) and its MeSH ID accordingly in the Table below where database identifiers are assigned to the corresponding selected text mentions. After accepting or correcting concept ids, the user can click to save or export all annotations (both text spans and concept ids) of the article. In either case, the time information for this annotation is also saved into the PubTator system.

### **3. Proposed tasks for BioCreative 2012 Track III**

We propose two general tasks that can be achieved using PubTator. Once a user is committed, a customized version will be provided should they have any specific requirements (See more about our system adaptability in Section 4).

#### **1. Document triage**

This task will assess our system for assisting human curators to prioritize papers for more detailed curation. A curator with a specific need will decide a query (e.g. a chemical name in the case of CTD curation) and search it in PubTator. The curator will then examine the returned search results and mark relevant papers to be curated. The experience with PubTator can be compared with the system they are currently using or general-purpose systems like PubMed with respect to productivity and effectiveness.

Input: a concept (gene/disease/chemical) name/identifier OR any PubMed query

Output: a list of PMIDs that are selected for further annotation.

#### **2. Bioconcept annotation**

This task will assess our system for assisting manual annotation of various kinds of bioconcepts. A curator can use our system to create and export annotations with regard to specific concepts. For instance, annotating genes is a central task for many model organism databases. After entering a list of PMIDs, PubTator will return the corresponding articles with machine tagged pre-annotations. The curator can then accept/edit/remove them or create new annotations. The goal is to see if using PubTator can accelerate this labor-intensive manual process.

Input: a list of PMIDs

Output: PMIDs with corresponding annotations (database identifiers).

### **4. System adaptability and interactivity**

In developing PubTator, we decided to develop it as a general curation tool rather than a specialized one in order to reach a broader community. However, once a team decides to adapt our system into their curation pipeline, PubTator is quite robust for customization. For instance, any team can use PubTator to perform the document triage task via a simple query. In this case, search results are returned in reverse time order by default. However, in the case where teams wish to have search results ranked by relevance, we can easily achieve this based on team provided training data. Indeed, this is the case we are doing for the CTD document triage task (See BioCreative 2012 Track I for details).

Similarly for bioconcept annotation, we can customize PubTator for different needs of model organism databases. For instance, in default setting NCBI Gene database is used in gene ID assignment. However, this can be changed to any other organism-specific gene nomenclature such as the Arabidopsis Genome Initiative locus identifiers.

Our system is Web-based and involves a great deal of human interaction. As described earlier, users are involved in many curation aspects ranging from selecting/deselecting articles in document triage to creating/editing/deleting pre-tagged markups in bioconcept annotation.

## 5. System evaluation

With regard to the underlying algorithms employed in the PubTator, some have already been extensively evaluated with exceptional performance (See Table 1 for details). We are currently evaluating the recall and precision of our algorithms in finding disease and chemical concepts. In addition, using the CTD data, a machine-learning based algorithm will be separately assessed in ranking curatable articles for document triage. All these benchmark experiments will be completed before the user-testing period in March 2012.

Module Name	Targeted Use	Precision	Recall	F-measure
GeneTUKit [2]	Gene Mention (abstract)	86.73%	82.36%	84.49%
GenNorm [3]	Gene Normalization (full text)	56.23%	39.72%	46.56%
SR4GN [5]	Species Recognition (abstract)	85.42%	85.42%	85.42%

**Table 1:** Reported benchmark performance of computational modules used in PubTator

Through participation in the BioCreative 2012 track III, we plan to collect interactive data and subsequently perform comparative analysis of curation effectiveness using our system vs. manual or another curation system. Furthermore, the user-curated data during the system testing could be used as the gold standard to report performance metrics such as precision and recall as requested by the track III organizers.

**Acknowledgments** The authors are grateful to Smith L, Comeau D and Dogan R for building the prototype annotation system. We also thank Wilbur WJ and Kim S for helpful discussion and Comeau D for proofreading the manuscript. Funding: Intramural Research Program of the NIH, National Library of Medicine.

## References

1. Névéol A, Doğan RI, Lu Z (2010) Semi-automatic semantic annotation of PubMed queries: a study on quality, efficiency, satisfaction. *Journal of Biomedical Informatics* **44**: 310-318.
2. Huang M, Liu J, Zhu X (2011) GeneTUKit: a software for document-level gene normalization. *Bioinformatics* **27**: 1032-1033.
3. Wei C-H, Kao H-Y (2011) Cross-species gene normalization by species inference. *BMC Bioinformatics* **12**: S6.
4. Yeganova L, Comeau DC, Kim W, Wilbur WJ. Text Mining Techniques for Leveraging Positively Labeled Data; 2011. pp. 155-163.
5. Wei C-H, Kao H-Y, Lu Z (2011) SR4GN: a species recognition software tool for gene normalization. *Plos one*. Submitted