Overview of the Interactive Task in BioCreative V

Qinghua Wang [1,2], Shabbir Syed Abdul [3], Lara Almeida [4], Sophia Ananiadou [5], Yalbi Itzel Balderas-Martínez [6], Riza BatistaNavarro [5], David Campos [7], Lucy Chilton [8], Hui-Jou Chou [9], Gabriela Contreras[6], Laurel Cooper [10], Hong-Jie Dai [11], Juliane Fluck [12], Socorro Gama [6], Georgios Gkoutos [13], Afroza Khanam Irin [14], Lars Juhl Jensen [15], Silvia Jimenez [16], Toni Rose Jue [17], Ingrid Keseler [18], Sumit Madan [12], Sérgio Matos [4], Peter McQuilton [19], Matthew Mort [20], Jeyakumar Natarajan [21], Evangelos Pafilis [22], Emiliano Pereira [23], Shruti Rao [24], Fabio Rinaldi [25], David Salgado [26,27], Onkar Singh [28], Raymund Stefancsik [29], Chu-Hsien Su [30], Suresh Subramani [21], Hamsa Dhwani Tadepally [31], Loukia Tsaprouni [32], Nicole Vasilevsky [33], Xiaodong Wang [34], Andrew Chatr-aryamontri [35], Stan Laulederkind[36], Sherri Matis-Mitchell[37], Johanna McEntyre[38], Sandra Orchard[38], Sangya Pundir[38], Raul Rodriguez-Esteban[39], Kimberly Van Auken[34], Zhiyong Lu [40], Mary Schaeffer [41],Lynette Hirschman[36], Cecilia Arighi [1,2*]

1 Department of Computer and Information Sciences, University of Delaware, USA; 2 Center of Bioinformatics and Computational Biology (CBCB), University of Delaware, USA; 3 International Centre of Health Information Technology, Taipei Medical University; 4 IEETA/DETI, University of Aveiro, Portugal; 5 National Centre for Text Mining, University of Manchester, UK; 6 Facultad de Ciencias, Universidad Nacional Autónoma de México, México; 7 BMD Software, Aveiro, Portugal; 7 BMD Software, Aveiro, Portugal; 8 Northern Institute for Cancer Research, Newcastle University, UK ; 9 Rutgers University-Camden, New Jersey, USA; 10 Dept. of Botany and Plant Pathology, Oregon State University Corvallis, OR 97331, USA; 11 Department of Computer Science and Information Engineering, National Taitung University, Taiwan, R.O.C.; 12 Fraunhofer Institute for Algorithms and Scientific Computing, Schloss Birlinghoven, Germany; 13 Department of Computer Science, Aberystwyth University Aberystwyth, UK ; 14 Life Science Informatics, University of Bonn; Bonn, Germany; 15 Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark; 16 Atheris S.A., Switzerland; 17 Prince of Wales Clinical School University of New South Wales NSW, Australia; 18 SRI International, Menlo Park, California, USA; 19 Oxford e-Research Centre University of Oxford, Oxford, UK; 20 HGMD, Institute of Medical Genetics, Cardiff University, Heath Park, Cardiff, UK; 21 Department of Bioinformatics, Bharathiar University, Coimbatore, Tamil Nadu, India; 22 Institute of Marine Biology, Biotechnology and Aquaculture, Hellenic Centre for Marine Research, Heraklion, Crete, Greece; 23 Microbial Genomics and Bioinformatics Group, Max Planck Institute for Marine Microbiology, Bremen, Germany; 24 Innovation Center for Biomedical Informatics (ICBI), Georgetown University, Washington D.C. USA; 25 Institute of Computational Linguistics, University of Zurich, Zurich, Switzerland; 26 Aix-Marseille Université, Marseille, France; 27 Inserm, Marseille, France; 28 Taipei Medical University

Graduate Institute of Biomedical informatics, Taipei City, Taiwan.; 29 University of Cambridge, Department of Genetics, Cambridge, UK; 30 Institute of Information Science, Academia Sinica, Taiwan, R.O.C.; 31 Freelance Scientific Curator , Ohio, USA ; 32 Institute of Sport and Physical Activity Research (ISPAR), University of Bedfordshire, Bedford, UK; 33 Oregon Health & Science University, Ontology Development Group, Portland, Oregon, USA; 34 WormBase Consortium, Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, USA; 35 Institute for Research in Immunology and Cancer, Université de Montréal, Montréal, Canada; 36 Medical College of Wisconsin, Winsconsin, US, 37 Reed Elsevier, Philadelphia, US, 38 European Bioinformatics Institute (EMBL-EBI), Hinxton, UK, 39 Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd, Switzerland; 40 National Center for Biotechnology Information, National Institutes of Health, Maryland, US, 41 MaizeGDB USDA ARS and University of Missouri, Columbia, US, 42 The MITRE Corporation, Bedford, Massachusetts, USA

wangq@udel.edu;
drshabbir@tmu.edu.tw;
{lara.monteiro,aleixomatos}@ua.pt;
{riza.batista,sophia.ananiadou}@manchester.ac.uk
yalbibalderas@ciencias.unam.mx;
david.campos@bmd-software.com;
hcgo139362@upemor.edu.mx;
lucy.chilton@ncl.ac.uk;
alwaysrejoice516@gmail.com;
cooperl@science.oregonstate.edu;
hjdai@nttu.edu.tw;
{sumit.madan,juliane.fluck}@scai.fraunhofer.de;
sgama@ccg.unam.mx;
g.gkoutos@gmail.com;
afroza1226irin@gmail.com;
lars.juhl.jensen@cpr.ku.dk;
silvia.jimenez@bluewin.ch;
t.jue@unsw.edu.au;
keseler@ai.sri.com;
peter.mcquilton@oerc.ox.ac.uk;
mortm@cardiff.ac.uk;
n.jeyakumar@yahoo.co.in;
pafilis@hcmr.gr;
pereiramemo@gmail.com;
sr879@georgetown.edu;
fabio.rinaldi@uzh.ch;
david.salgado@univ-amu.fr;
onkarnims2009@gmail.com;
rs573@cam.ac.uk;
jason@iis.sinica.edu.tw;
sureshsubramani@hotmail.com;
hamsa_t@yahoo.com;

loukia.tsaprouni@beds.ac.uk;
vasilevs@ohsu.edu;
xdwang@caltech.edu;
aryamontri@gmail.com;
slaulederkind@mcw.edu;
sherrimatismitchell@gmail.com;
{orchard,sangya.pundir,mcentyre}@ebi.ac.uk;
raul.rodriguez-esteban@roche.com;
luzh@ncbi.nlm.nih.gov;
Mary.Schaeffer@ars.usda.gov;
lynette@mitre.org
arighi@dbi.udel.edu

* corresponding author

Abstract. Fully automated text mining (TM) systems promote efficient literature searching, retrieval, and review but are not sufficient to produce ready-to-consume curated documents. These systems are not meant to replace biocurators, but instead to assist them in one or more literature curation steps. To do so, the user interface is an important aspect that needs to be considered for tool adoption. The BioCreative Interactive task (IAT) is a track designed for exploring user-system interactions, promoting development of useful text mining tools, and providing a communication channel between the biocuration and the text mining communities. In BioCreative V, the IAT track followed a format similar to previous interactive tracks, where the utility and usability of TM tools, as well as the generation of use cases, have been the focal point. The tasks proposed are user-centric and formally evaluated by biocurators. In BioCreative V IAT, seven TM systems and 43 biocurators participated. Two levels of user participation with different commitment were offered to broaden curator involvement and obtain more feedback on usability aspects. The full level participation involved training on the system, curation of a set of documents with and without text mining assistance, tracking of time-on-task, and completion of a user survey. The partial level participation was designed to focus on usability aspects of the interface and not the performance *per se*. In this case, biocurators navigated the system by performing pre-designed tasks and they were asked whether they were able to achieve the task and the level of difficulty in completing the task. In this manuscript, we describe the development of the interactive task, from planning to execution, and discuss some findings for the systems tested.

Keywords: Biocuration; Text Mining; Usability; Information Retrieval; Information Extraction

1 Introduction

BioCreative: Critical Assessment of Information Extraction in Biology is an international community-wide effort that evaluates text mining (TM) and information extraction (IE) systems applied to the biomedical domain [1-5]. A unique characteristic of this effort is its collaborative and interdisciplinary nature, as it brings together experts from various fields, including TM, biocuration, publishing houses and bioinformatics. Therefore, each competition is tailored towards specific needs of these communities. BioCreative has been working closely with biocurators to understand the various curation workflows, the TM tools that are being used and their major needs [6, 7]. To address the barriers in using text mining in biocuration, BioCreative has been conducting user requirements analysis and user-based evaluations, and fostering standards development for TM tool re-use and integration. The BioCreative Interactive text mining Task (IAT) introduced in BioCreative III [8, 9] has served as a means to observe the approaches, standards and functionalities used by state-of-the-art text mining systems with potential applications in the biocuration domain. The IAT task also provides a means for biocurators to be directly involved in the testing of TM systems. The benefits are multifold, including: direct communication and interaction; exposure to new TM tools that can be potentially adapted and integrated into the biocuration workflow, contribution to the development of systems that meet the needs of the biocuration community, and dissemination of findings in proceedings and peer-reviewed journal articles. A User Advisory Group (UAG), representing a diverse group of users with literature-based curation needs, has been assisting in the design and assessment of the IAT[1]. The current article describes the IAT task, the workflow of the IAT activities, the participating systems, and the results from the user evaluation.

2 Methods

2.1 Task description

Teams were invited to present a web-based system that could address a biocuration task of their choice. The systems were expected to follow the system requirements proposed in the call of participation[2]. Selection of participating systems was based on the evaluation of a document containing the description of the system, including the relevance of the proposed task to the targeted community, use of standards (vocabularies and ontologies), and baseline evaluation of the system or its components.

In addition, we invited biocurators to participate in the evaluation of such systems via the International Biocuration Society mailing list, and with the help from UAG members. This study was conducted remotely. Two levels of participation were offered: *full* (total commitment time of approximately 12 hours per system) which involved training, performing pre-designed tasks, curating a set of documents, and completing a user survey; and *partial* (total commitment time of approximately 30 minutes to 1 hour

[1] http://www.biocreative.org/about/biocreative-v/user-advisory-group/

[2] http://www.biocreative.org/tasks/biocreative-v/track-5-IAT/

per system) which involved performing basic pre-designed tasks at the system's website, and providing feedback via a user survey. The timespan to complete the activity was 6 weeks. **Table 1** shows the suggested timeline for the full level participation activity.

Table 1. Activity workflow of full level participation

Week	Activity
Week 1	Training with guided exercises with text mining team
Week 2	Review of task guidelines with text mining team and coordinator.
Week 3	Pre-designed tasks exercise
Week 4	1h annotation (non-TM assisted) and 1h annotation (TM-assisted)
Week 5	1h annotation (non-TM assisted) and 1h annotation (TM-assisted)
Week 6	Survey and submission of data

2.2 Pre-designed tasks and surveys

For the usability test and surveys we reviewed and followed guidelines outlined in usability websites[3]. All surveys and activities were prepared and presented to the user via the SurveyMonkey interface[4] and responses were collected in CSV format, and some summary information and the Net Promoter Score (NPS, promoters-detractors) were directly calculated by the software. All the pre-designed tasks and surveys described in this section can be accessed from the BioCreative website[5].

A collection of pre-designed tasks was prepared for each system with feedback from the participating teams. With previous consent, we asked all users to perform short tasks in the system of choice to encourage them to navigate and provide initial feedback on their overall impressions about the system. Examples of pre-designed tasks included: i) accessing the TM tool ii) testing general functionalities (such as searching and sorting), iii) finding documentation, iv) editing capabilities, v) saving results, and vi) understanding semantics of icons/buttons/tabs. Each task was followed by questions on the user's ability to complete the task and difficulty in accomplishing the task. At the end we asked some general questions about the system, such as perception of assistance in the biocuration task proposed, and feedback for improvements, followed by a set of questions to address usability, and user satisfaction questions (rating experience with the

[3] http://www.usability.gov
[4] https://www.surveymonkey.com/
[5] http://www.biocreative.org/media/store/files/2015/IATsystemsurveys2015.pdf

system, and likeliness to recommend the system to others). Response to general questions were converted from a semantic scale to a numerical scale of 1 to 5, ranging from most negative to most positive feedback, respectively. We represented the data in terms of positive (with score > 3), negative (with score < 3), neutral (with score = 3), and skipped (questions not responded or not applicable). For investigating the possible correlation between the different questions the following correlation coefficients were calculated: Spearman Rho[6], and Kendall Tau[7].

For the full level participation, we modified the user survey from BioCreative IV [10] to include the questions needed to calculate the System Usability Scale (SUS [11]). The SUS is composed of ten statements, each having a five-point scale that ranges from Strongly Disagree to Strongly Agree, alternating positive and negative statements. A score of 68 is considered average, thus SUS scores higher than 68 can be considered above average[8]. As before, we also included the set of questions for the categories i) Comparison to similar systems, ii) System's ability to help complete tasks, iii) Design of application, and iv) other usability aspects. For each of the system, responses from users were aggregated for all questions related to a given category. We calculated the central tendency, using the median, the minimum and maximum values for the set (Min and Max in Tables 4-10, respectively), along with the 25% or lower quartile (splits off the lowest 25% of data from the highest 75%, Q1 in Tables 4-10) and the 75% or upper quartile (splits off the highest 25% of data from the lowest 75%, %, Q3 in Tables 4-10).

3 Results and Discussion

The Interactive Task (IAT) consisted of the demonstration and evaluation of web-based systems addressing literature curation tasks, evaluated by biocurators on performance and usability. One of the main goals was to collect data from biocurators testing the systems, and provide useful feedback to developers on possible enhancements and how to better tailor their system for biocuration. The UAG was engaged in multiple aspects of the task, including preparing the requirements for the systems, the reviewing of the user survey, recruitment of biocurators, and the testing of the systems. Each run of the IAT activity is modified based on previous BioCreative outcomes. BioCreative IV included a DOE-sponsored session on the text mining needs of the metagenomics community. The discussions from this session inspired the participation of a team for BioCreative V addressing the needs specific to the metagenomics community, namely extraction of environmental and species metadata from free text.

3.1 Systems and user recruitment

Seven teams participated in the IAT. **Table 2** summarizes some aspects of the participating systems. It is worth noting that this time, a common theme around gene and

[6] http://www.socscistatistics.com/tests/spearman/

[7] http://www.wessa.net/rwasp_kendall.wasp

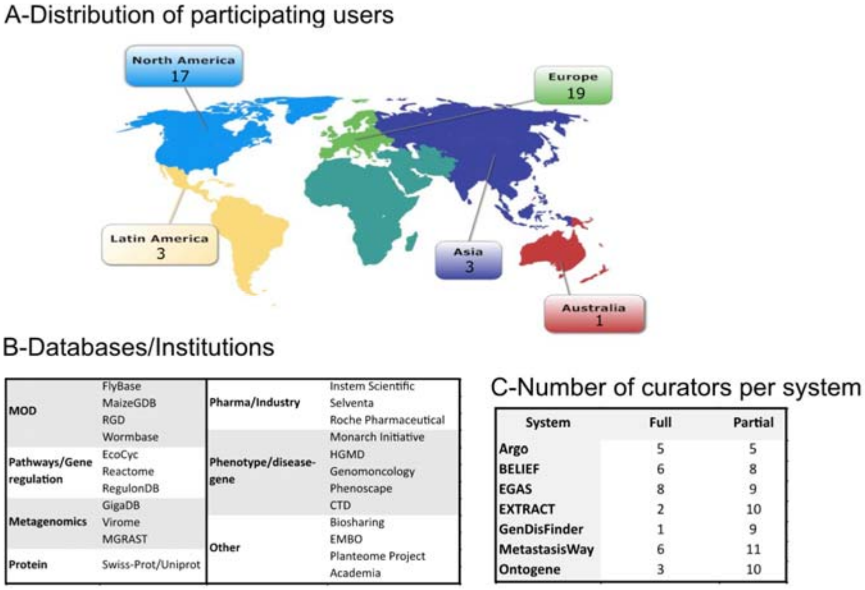
[8] http://uxpamagazine.org/sustified/

disease/phenotype annotation was prevalent (5 out of 7 systems). However, there was a great variability in the complexity of the systems presented. Some offered workflow design options (e.g Argo), management systems for curation (e.g egas and BELIEF), plug-ins/bookmarklets for the web browser (e.g. EXTRACT and MetastasisWay), and network visualization (e.g., GenDisFinder and MetastasisWay).

Table 2. Summary of IAT participating systems.

System	Description	Bioconcepts	Link to Standards/Databases	Relations captured	Text	Browser
Argo	Curation of phenotypes relevant to the chronic obstructive pulmonary disease (COPD)	-protein -medical condition -sign/symptom -drug	-UniProt -UMLS -ChEBI	-COPD-disease relations -COPD-drug relation -COPD-protein relation -COPD-sign/symptom relations	full-text	-Chrome -Firefox -Safari
Egas	Identification of clinical attributes associated with human inherited gene mutations, described in PubMed abstracts	-gene/protein -mutation -disorder/disease -zygocity -penetrance -ethnicity	-HGNC -OMIM -Human Phenotype Ontology -NCI Thesaurus	-gene/protein-mutation relation -gene/protein-disease relation -mutation-zygocity relation -mutation-penetrance relation	abstract	-Chrome -Firefox -Safari
GenDisFinder	Knowledge discovery of known/novel human gene-disease associations from biomedical literature	-gene -disease	-EntrezGene -OMIM	-gene/protein-disease relations -GDA-related action words and network association type	abstract	-Chrome -Explorer -Firefox -Safari
MetastasisWay	Look for the biomedical concepts and relations associated with metastasis and finally construct the metastasis pathway.	-gene/protein -metastasis -cancer -tissue -body part -microRNA -gene expression -cell line	-EntrezGene -Disease Ontology -MirTarBase	positive and negative regulations between biomedical concepts associated with metastasis	abstract	-Chrome
Ontogene	Curation of bioconcepts, such as miRNA, gene, disease and chemicals and their relations.	-microRNA -gene/protein -disease -organism	-RegulonDB ID -CTD -NCBI taxonomy		full-text	-Chrome -Firefox -Safari
BELIEF	Semi-automated curation interface which supports relation extraction and encoding in the modeling language BEL (Biological Expression Language).	-gene/protein -disease -chemical -biological processes	-HGNC/MGI/RGD -MeSH Diseases Branch -ChEBI -GO-Biological Process -GO-Complex -Selventa Protein/Family Names	Relations expressed in BEL (Biological Expression Language). Relations can be expressed between all of the detected entity types	abstracts, full-text	-Chrome -Firefox
EXTRACT	Lists the environment type and organism name mentions identified in a given piece of text.	-environment -organism -tissue -disease	-Environment Ontology -NCBI taxonomy -BRENDA Tissue Ontology -Disease Ontology		text snippets	-Chrome -Firefox

With the help of the UAG and the teams, we were able to recruit a wide variety of biocurators worldwide. A total of 43 biocurators participated in the IAT in different capacities. Fig.1 shows the distribution by geographical location (Fig.1A), examples of type of database or institution represented (Fig.1B), and distribution by system and level of participation (Fig.1C). All systems were inspected by at least 7 biocurators at some level (full/partial).


A-Distribution of participating users

B-Databases/Institutions

MOD	FlyBase MaizeGDB RGD Wormbase	**Pharma/Industry**	Instem Scientific Selventa Roche Pharmaceutical
Pathways/Gene regulation	EcoCyc Reactome RegulonDB	**Phenotype/disease-gene**	Monarch Initiative HGMD Genomoncology Phenoscape CTD
Metagenomics	GigaDB Virome MGRAST	**Other**	Biosharing EMBO Planteome Project Academia
Protein	Swiss-Prot/Uniprot		

C-Number of curators per system

System	Full	Partial
Argo	5	5
BELIEF	6	8
EGAS	8	9
EXTRACT	2	10
GenDisFinder	1	9
MetastasisWay	6	11
Ontogene	3	10

Fig. 1. Distribution of biocurators (A) by geographic area, (B) by type of database/institution, and (C) by level of participation. A total of 43 biocurators participated in this activity. Notice that the total number in (C) is higher because some biocurators tested more than one system, and all curators participated in the partial activity.

3.2 Evaluation

It should be noted that the IAT activity is a demonstration task, which yields qualitative rather than quantitative results. In addition, given the diversity of biocuration tasks proposed and varied complexity of the systems, the results should not be directly compared, but taken each within its specific context. Therefore we present the data highlighting some general trends or important differences.

Pre-designed tasks.

The pre-designed task activity was customized for each system. By reviewing the answers to questions about the ability to complete task, its difficulty, and confidence, specific problems with the system can be identified. Table 3 shows the percentage of users who completed each task per system (n/a means we have no data for that field). In general, users were able to accomplish the tasks requested. Some cases where users failed to accomplish tasks were related to: inability to install or access the system; functionality that did not work properly at the time; the formatting of the input text, and, in a few cases, the user simply did not understand the task. In the case of BELIEF system, which produce expressions in BEL language, some of the users reported that they were unfamiliar with the BEL language, and therefore, felt less confident in some of the tasks (e.g. editing and exporting the statements).

Table 3. -Results on task completion in the pre-designed tasks for each system.

TASK	% users completed task	Based on those who completed task	
		% found it difficult	% non-confident
Argo (5 curators)			
TASK1-Launching Argo	100	0	0
TASK2-Find the page with tutorial for curation task	80	0	0
TASK3-Managing files in Argo	100	0	0
TASK4-Open a file	80	25	0
TASK5-Edit annotations	80	25	0
TASK6-Saving annotations	80	25	0
BELIEF (8 curators)			
TASK1-Find information about BEL	100	13	13
TASK2-Find and open project. Understanding content of page	100	0	13
TASK3-Edit the BEL statements and select for export	75	33	17
TASK4-Export the document	100	0	13
TASK5-Add document to project	88	14	0
egas (9 curators)			
TASK1-Log in and access the project	100	0	0
TASK2-Find project status (private vs public)	89	0	13
TASK3-Finding help	100	0	0
TASK4-Edit annotation	100	0	0
TASK5-Export and opening file	33	0	0
EXTRACT (10 curators)			
TASK1-Install bookmarklet	100	0	0
TASK2-Extract on a piece of text	100	0	0
TASK3-Review annotations and information	90	0	0
TASK4-Save Extract table	100	n/a	n/a
TASK5-Finding help	100	0	0
GenDisFinder (9 curators)			
TASK1-Find information on format	100	0	0
TASK2-Find GenDisFinder gene-disease associations in a given abstract	33	0	0
TASK3-Understand annotations and network	33	0	0
TASK4-Edit annotation	56	20	20
TASK5-Export annotation	67	n/a	n/a
MetastasisWay (11 curators)			
TASK1-Register and install MAT	82	33	22
TASK2-Find information about vocabularies used*	89	13	50
TASK3-Review and edit annotations*	67	17	17
TASK4-Save annotation*	89	n/a	n/a
*calculations based on the 9 curators who were able to install the application			
Ontogene (10 curators)			
TASK1-Open a document in Ontogene	100	10	0
TASK2-Find information about panels	100	10	0
TASK3-Using filters in panels	100	0	0
TASK4-Validate annotation	80	0	0
TASK5-Export annotations	100	0	0

The results collected from overall assessment of each system is shown in Fig.2. Many of the systems show a high proportion of skipped answers in the error message category, meaning that the user did not encounter any error messages along the way.

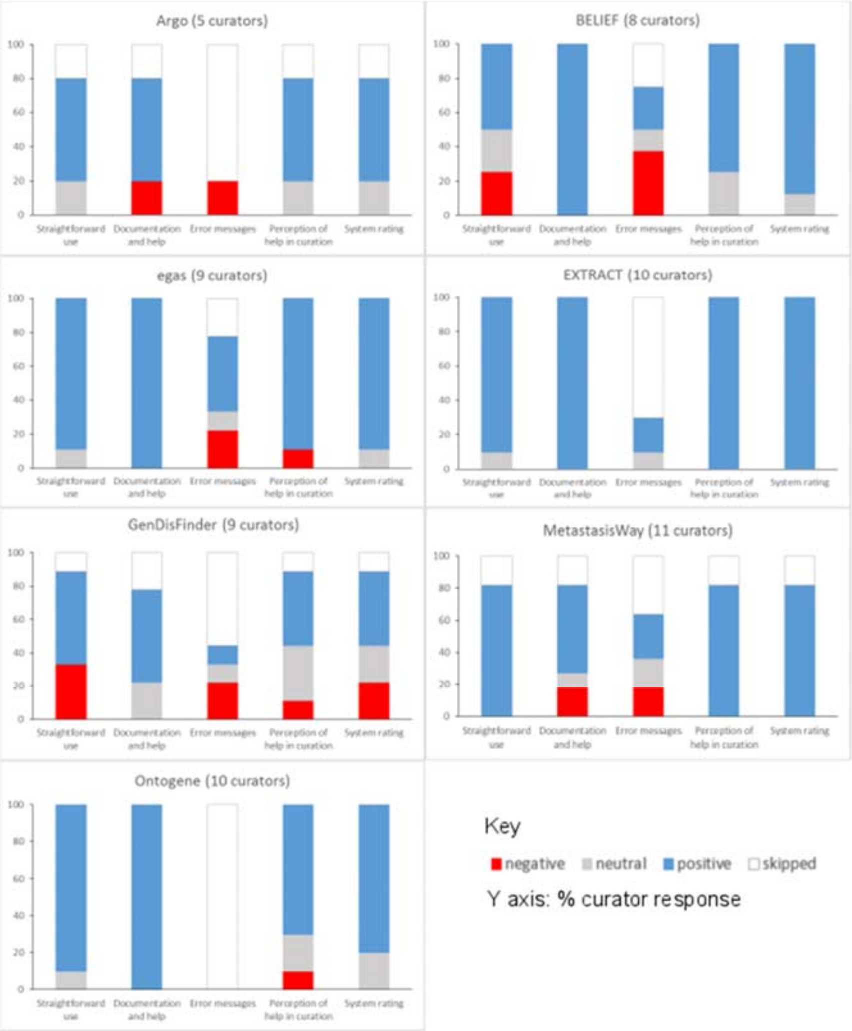


Fig. 2. Pooled responses to questions related to system perception of usability from the pre-designed task activity.

We investigated if the perception of the system helping in the biocuration task correlated with the system rating by calculating the correlation between the collective responses for each of these questions. The result shows that there is weak positive correlation between the perception of the system helping in the biocuration task and the rating of the system (Spearman's R=0.3996 and 2-sided p=0.0023; Kendall tau=0.3614, 2-sided p=0.0227), suggesting that the users would be more likely to rate the system higher if he/she perceives that the system would assist in biocuration task.

When we compare the Net Promoter Scores (NPS, bars in Fig.3) in response to the question about likelihood of recommending the system to a colleague/friend, with the median of the system rating (black dots in Fig.3), we don't find a consistent trend (Fig.3). Although all systems have positive median ratings, they are not always accompanied by their recommendation to others.

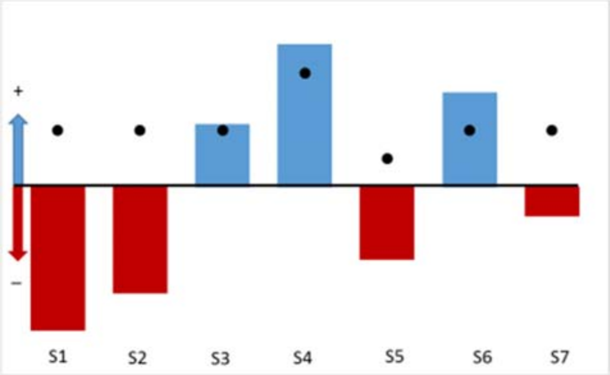


Fig. 3. Plot of the NPS score and the median for the system rating for each system (S1-S7). The y-axis represents whether the NPS and median are positive (for NPS, positive means NPS>0, for system rating median >3) or negative (for NPS, negative means NPS<0, for system rating median <3) The NPS score is represented with bars, blue and red color indicate positive and negative scores, respectively. The median for the system rating is represented with black dots.

Full curation task by system

Argo[9] (Team 277: Batista-Navarro, Carter, and Ananiadou)
Description: A generic text mining workbench that can cater to a variety of use cases, including the semiautomatic curation of information from literature. It enables its technical users to build their own customized TM solutions by providing a wide array of interoperable and configurable elementary components that can be seamlessly integrated into processing workflows. With Argo's graphical annotation interface, domain experts can then make use of the workflows' automatically generated output to curate information of interest.
Task: Five domain experts utilized Argo for the curation of phenotypes relevant to the Chronic Obstructive Pulmonary Disease (COPD). Specifically, they carried out three curation subtasks, namely: (1) the markup of phenotypic mentions in text, e.g., medical conditions, signs or symptoms, drugs and proteins, (2) linking of mentions to relevant vocabularies/ontologies, i.e., normalization, and (3) annotation of relations between COPD and other mentions.
Corpus: Based on 30 COPD relevant PubMed Central Open Access papers which were annotated as part of previous work [12]. The corpus was split into two subsets

[9] http://argo.nactem.ac.uk

with 15 papers each: one for training the text mining tools underpinning the semiautomatic COPD phenotype curation workflow, and another from which the documents for curation were drawn. Since the time constraints did not make the annotation of entire full-text papers feasible, we defined a document as a smaller chunk of text (e.g., section paragraphs according to each paper's metadata). Based on automatic random selection, 124 such documents were set aside for the curation task. The first 62 were used for purely manual curation while the remaining were exploited in the text mining (TM) assisted mode of the task. All of the biocurators were working on the same data set.

Results: Results from the performance and survey are summarized in Table below.

Table 4. Argo metrics from full level evaluation.

Performance			**Ave # documents/hour**					
Curators	**Annotation**		**non-TM assisted**		**TM assisted**		**Ave. IAA**	
5	concept		9		14		68.12%	
	relation		25		35			
Survey	**median**	**Q1**	**min**	**max**	**Q3**		**Ave.**	**St. Dev**
Task	4	4	2	5	4.5	**SUS**	71	3.6
Design	3	3	2	4	4	**Usability**	72.5	3.5
Usability	4	3	2	5	4	**Learnability**	65	8

Using the concept annotations (e.g., text span boundaries and semantic types) of the expert who voluntarily curated all of the 124 documents in the data set, we evaluated the performance of the Argo workflow which formed the basis of the text mining support provided to the biocurators. The overall micro-averaged precision, recall and F score values are 68.17, 63.96, and 66.97, respectively. These results are quite encouraging especially considering that the F-score (66.97) is very close to the measured inter-annotator agreement (68.12), indicating that the automatic concept annotation workflow performs comparably with human curators. The usability score is just slightly higher than the average; the learning component seems to have the highest variability.

BELIEF[10] (Team 333: Madam, Hodapp and Fluck)

Description: Semi-automated curation interface which supports an expert in relation extraction and encoding in the modeling language BEL (Biological Expression Language). BEL can represent biological knowledge in causal and correlative relationships that are triples. A triple consists of a subject, a predicate (relationship) and an object.

Corpus: In total 20 PubMed abstracts were chosen for the curation. The documents were selected from different areas with different entities but consistent with the context for which BELIEF was created. All users worked with the same set of data divided into two sets (Set1 and Set2) containing 10 documents each.

Results: There were two distinct groups of users, those who had previous experience with BEL statements and were familiar with the BELIEF, vs. those who were new to both the BEL language and the annotation interface. Annotators in the first group had

[10] http://belief.scai.fraunhofer.de/BeliefDashboard/

a higher throughput per hour (approximately 5 documents) than the novice (1-2 documents). There does not seem to be a consistent difference of the tool speeding up curation of BEL statements, but this could be due to the learning curve for the BEL language and the interface, and the low number of documents that are therefore annotated. The final survey shows that the learnability, as computed based on the SUS questionnaire, gives the lowest score with the highest variability, which depends on the user experience. This is in agreement with the results shown for the pre-designed tasks.

Table 5. BELIEF metrics from full level evaluation.

Survey	median	Q1	min	max	Q3		Ave.	St. Dev
Task	4	3	2	5	4	**SUS**	66.67	15.28
Design	3.5	3	2	5	4	**Usability**	67.19	13.54
Usability	3	3	2	4	4	**Learnability**	64.58	31.25

Egas[11] (Team 286: Matos, Campos, Pinho, Silva, Mort, Cooper, Oliveira)
Description: Egas is a web-based platform for text-mining assisted literature curation, supporting the annotation and normalization of concept mentions and relations between concepts. Egas allows the definition of different curation projects with specific configuration in terms of the concepts and relations of interest for a given annotation task, as well as the ontologies used for normalizing each concept type. Egas may be described as an “annotation-as-a-service” platform. Document collections, users, configurations, annotations and back-end data storage, are all managed centrally, as are the tools for document processing and text mining. This way, a curation team can use the service, configured according to the annotation guidelines, taking advantage of a centrally managed pipeline.
Task: Identification of human inherited gene mutations and associated clinical attributes, such as inheritance mode and penetrance, described in PubMed abstracts. Seven curators were selected and were asked to annotate documents that were pre-analyzed by an automatic concept recognition tool (half of the corpus), and raw documents (the remaining corpus), in order to evaluate the added benefit of text mining-assisted curation. Three curators annotated the complete corpus, two followed a four hour time-limited work plan, and other two curators annotated a small portion of the corpus (13 and 9 documents).
Corpus: 100 abstracts randomly selected from search previously tailored to the Human Gene Mutation Database (HGMD).
Results: It took in general a shorter time to curate documents that had been previously annotated by the concept recognition tool, although the results are not conclusive (Table 6). The inter-annotator agreement is acceptable for this task. In terms of perception metrics, the usability SUS score is above average for this system, and consistently positively rated in all aspects evaluated.

[11] https://demo.bmd-software.com/egas/

Table 6. Egas metrics from full level evaluation.

Performance				
Curators	**Annotation**	**non-TM assisted**	**TM assisted**	**Ave. IAA**
7	concept	449	495	0.74
	relation	105	138	
	time/article (s)	222.8	198.6	p-value 0.21
	time/concept (s)	12.9	9.6	p-value 0.07

Survey	**median**	**Q1**	**min**	**max**	**Q3**		**Ave.**	**St. Dev**
Task	4	4	3	5	5	**SUS**	77.14	9.69
Design	4	4	3	5	5	**Usability**	76.34	9.18
Usability	4	3	3	5	4	**Learnability**	80.36	13.26

EXTRACT[12] (Team 327: Pafilis, Buttigieg, Schnetzer, Arvanitidis, and Jensen)
Description: Interactive annotation tool, which helps curators, during browsing, to identify and extract standard-compliant terms for the annotation of the source environment of metagenomics and other sample records. Behind the web-based user interface, the system combines components from published systems for named entity recognition of environments, organisms, tissues and diseases.
Task: The two full evaluators were asked to investigate if the EXTRACT bookmarklet can help them locate sampling environment information in a document and if it can accelerate the metagenomics record metadata annotation process. In particular, they were asked to annotate samples as recommended by the standards. Annotated metadata included filling in the ‘environmental feature, environmental material and biome’ describing a sample’s source environment. The evaluators performed this task with and without the assistance of EXTRACT and compared the time taken in both cases. The goal was to assess the curation acceleration that EXTRACT offers when evaluators work as closely as possible to their actual workflow.
Corpus: The full evaluators were asked to try EXTRACT with records they would in any case annotate as part of their normal curation tasks. In response to this, each evaluator curated eight multiple metagenomics record related full-text articles.
Results: Comparison of fully manual and text-mining-assisted curation revealed that EXTRACT speeds up annotation by 15–25% and helps curators detect terms that would otherwise have been missed. The quality of the tagging results for species and environments has previously been evaluated on gold-standard corpora consisting of Medline abstracts and of Encyclopedia of Life species summary pages, respectively [13, 14]. Counted at the level of individual mentions, the SPECIES and ENVIRONMENTS taggers showed precisions of 83.9% and 87.8%, recalls of 72.6% and 77.0%, and F1 scores of 78.8% and 82.0%. The quality of the NER of tissues and diseases has not been benchmarked directly; however, these NER components have shown to give good results when used for co-mentioning-based extraction of protein–tissue and protein–disease associations [15, 16]. In terms of perception metrics, the evaluators generally found the system to be intuitive, useful, well documented and sufficiently accurate to be helpful in spotting relevant text passages and extracting organism and environment terms (Fig.

[12] https://extract.hcmr.gr

3 and Table 7). The SUS score is above average but with high variability as it is the result of two users using EXTRACT in their own different curation pipelines.

Table 7. EXTRACT metrics from full level evaluation.

Survey	median	Q1	min	max	Q3		Ave.	St. Dev
Task	4	3.25	1	4	4	**SUS**	77.5	20.0
Design	4.25	3.75	2	5	5	**Usability**	76.6	20.3
Usability	4	4	4	4	5	**Learnability**	81.2	18.7

GenDisFinder[13] (Team 294: Subramani and Natarajan)
Description: Web-based text mining tool that aids in the extraction of known/novel human gene-disease associations from biomedical literature and further categorizes them using networks analysis. GeneDisFinder has four different modules for the above tasks: 1) gene mention and normalization of gene/protein names with NAGGNER and ProNormz [17], respectively, 2) disease mention identification and normalization using OMIM based normalized disease phenotype dictionary, 3) identification and extraction of semantic relations between genes and diseases using a relation keyword dictionary, and 4) construction of gene-disease association networks and further categorization. To our knowledge, GenDisFinder is the first tool which integrates text mining with network analysis to discover novel genes associated with a disease and provides an interface to view the interaction network.
Task: Curate a set of abstracts for gene-disease association. Curate genes, disease and gene-disease association relations. Also validate the categorization of the abstract into novel, unknown or known gene-disease associations.
Corpus: In-house curated gene-disease association corpus called the Human Gene-Disease Association (HGDA) corpus which is available on-line from the website. From GeneRIF database, we randomly selected 500 sentences which were manually annotated by three domain experts in our lab with gene name, disease name relation type, and gene-disease association information and called the HGDA corpus for our text mining methodology evaluation. The HGDA corpus contains PubMedID, corresponding sentences, EntrezGene HGNC approved gene entries, OMIM phenotype based disease entries, relation types such as genetic variation, altered expression, regulatory modification, negative association or ‘any’. The final HGDA corpus contains 157 unique genes, 96 unique diseases and 206 relations between them from 182 sentences.
Results: Note that only one curator participated in the full annotation task. Based on this unique user, the SUS score is lower than the average 68, and it seems to be mostly related to usability aspects, as learnability item has a score of 75. Other questions also related about usability and help in task completion were mostly neutral (value 3).

[13] http://biominingbu.org/GenDisFinder

Table 8. GenDisFinder metrics from full level evaluation.

Survey	median	Q1	min	max	Q3		Ave.	St. Dev
Task	3	3	3	3	3	**SUS**	57.50	n/a
Design	3.5	3	3	4	4	**Usability**	53.12	n/a
Usability	3	3	3	3	3	**Learnability**	75.00	n/a

MetastasisWay[14] (Team 311: Dai, Su, Lai, Chang, and Hsu)

Description: Curation tool developed as a Chrome browser extension which allows curators to review and edit concepts and relations related to metastasis directly in PubMed. PubMed users can view the metastatic pathways integrated from the large collection of research papers. The text mining services support a wide range of biomedical concepts including gene, microRNA, neoplasm metastasis, cytoskeleton, cell movement, cell adhesion, neoplasms, tissues and organ. Based on the recognized concepts, the relations among them are determined and sent for visualization in the client side browser.

Task: Annotate abstracts with the nine biomedical concepts related to metastasis described above and also any relation within or between those concepts of the type positive regulation, negative regulation, or neutral regulation.

Corpus: To collect a set of articles related to metastasis and its regulation, we searched PubMed with the query term “EMT[title/abstract] AND TGF-β[title/abstract]”. From the result, 300 abstracts were randomly selected as the curation dataset for the interactive text mining task. The data was split among six curators who participated in the task with overlapping sets.

Results: The annotation throughput of the non-TM assisted task (but using BRAT[15]) vs. the TM assisted task is slightly higher for the non-TM assisted (Table 9). This unexpected result could be due to difference on extent of annotation (MetastasisWay annotates all bioconcept mentions and relations along with links to identifiers, whereas, in manual mode the user concentrate only the sentences containing the relations, and in some cases they did not normalize the annotated concepts). Despite the results above, the perception of usability measures are overall positive for this system with SUS score within the average range, consistent with results from the pre-designed task.

[14] http://btm.tmu.edu.tw/metastasisway

[15] http://brat.nlplab.org/standoff.html

Table 9. MetastasisWay metrics from full level evaluation.

Performance								
Curators	Annotation	non-TM assisted	TM assisted					
6	#abstracts Week1	46	40					
	#abstracts Week2	49	44					
Survey	me-dian	Q1	min	max	Q3		Ave.	St. Dev
Task	4	3.25	1	5	4	SUS	68.75	5.41
Design	4	4	3	5	5	Usability	68.75	7.29
Usability	4	3	2	5	4	Learnability	68.75	14.58

Ontogene[16] (Team 364: Balderas-Martinez, Rinaldi, Contreras, Solano, Sanchez-Perez, Gama-Castro, Collado-Vides, Selman, and Pardo)
Description: Ontogene is a platform for the curation of bioconcepts, such as miRNA, gene, disease and chemicals and their relations.
Task: Use the OntoGene text mining pipeline and the ODIN curation system to curate miRNAs in relation to one particular respiratory disease, idiopathic pulmonary fibrosis from full length articles. Annotate miRNA name, target genes, transcription factors associated, organism, disease, level of miRNA and some characteristics of the sample.
Corpus: For the miRNA corpus the articles were selected by PubMed search with the query: idiopathic pulmonary fibrosis AND microRNA. The final corpus contained 62 articles.
Results: Note that this system was specifically tailored for the RegulonDB curation pipeline, and was tested at the full level by RegulonDB curators. The results are very positive, the throughput of articles curated using Ontogene platform is much higher than the non-TM assisted mode. Also the SUS score and other subjective measures are quite high for this system. This shows that the integration of Ontogene in the curation pipeline has been successful.

Table 10. Ontogene metrics from full level evaluation.

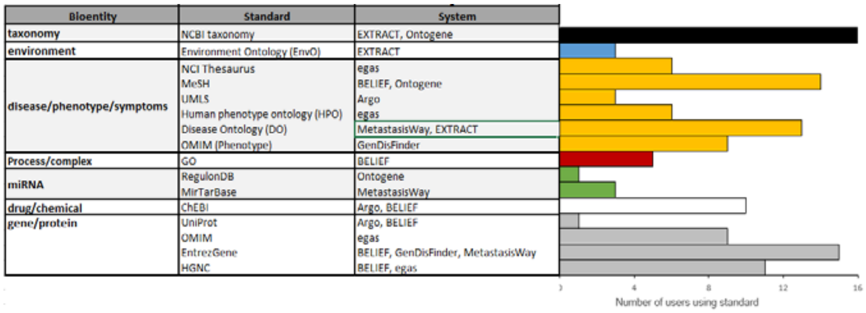
Performance								
Curators	Annotation	non-TM assisted	TM assisted					
3	Ave.articles/day	1	12					
Survey	me-dian	Q1	min	max	Q3		Ave.	St. Dev
Task	4	3	3	5	5	SUS	91.67	4.44
Design	4	3.75	3	5	4.25	Usability	90.62	6.25
Usability	3	3	3	4	4	Learnability	95.83	5.55

[16] http://www.ontogene.org

General Observations

One of the important aspects of the interactive activity is to do a reality check between what the systems offer and what the users need. We looked, for example, at what standards the systems offer for annotation (Table 2) and asked the set of curators who participated in the full task what standards they use or intend to use in their work. The results are presented in Fig.4. The table on the left side lists the bioentity types and standards used with the bar graph on the right side depicting the number of curators using such standard. It is very positive to see that the standards implemented by the systems are used by the community.

In a couple of cases the evaluation revealed important differences in the way the user and the system approaches the curation. In one case, the biocuration task asked curators to curate all the mentions extensively, including relations and normalization proposed by the system, whereas in reality the user would only be interested in curating the subset that is most relevant to them. In another case, the task included categorizing gene-disease association into known, novel and unknown. However, the definitions of novel and unknown were not intuitive to the users. The term novel was used by the system to indicate that the association of the gene to a disease was based on the association network, while for the user this would be an inference, not a novelty. Whereas the term "unknown" was used for gene-disease relations found in the text which are not yet in the system, so for the user this would be a novelty (experimental evidence of association).



Bioentity	Standard	System
taxonomy	NCBI taxonomy	EXTRACT, Ontogene
environment	Environment Ontology (EnvO)	EXTRACT
disease/phenotype/symptoms	NCI Thesaurus	egas
	MeSH	BELIEF, Ontogene
	UMLS	Argo
	Human phenotype ontology (HPO)	egas
	Disease Ontology (DO)	MetastasisWay, EXTRACT
	OMIM (Phenotype)	GenDisFinder
Process/complex	GO	BELIEF
miRNA	RegulonDB	Ontogene
	MirTarBase	MetastasisWay
drug/chemical	ChEBI	Argo, BELIEF
gene/protein	UniProt	Argo, BELIEF
	OMIM	egas
	EntrezGene	BELIEF, GenDisFinder, MetastasisWay
	HGNC	BELIEF, egas

Fig. 4. Usage of standards/databases proposed by the systems. The table describes most of the bioentities and standards/databases proposed by the different systems, and the bar graphs show the number of IAT evaluators using each standard/database. Note that environment is a specialized bioentity type which is only used by the microbial and metagenomics communities. Data from 25 users.

Overall we can say that the users had a satisfactory experience with the system tested, and in terms of performance and usability measures, a few systems have been consistent throughout the evaluation and seem to have promising potential for wider adoption. It is worth noting that this was mostly the case for the teams that worked very closely worked with the users. We should also highlight that the system tackling the metagenomics needs has been tested in the context of different biocuration pipelines and although an extensive evaluation could not be done, it seems that it is a promising tool, not only to the two curators but to the ten additional users who tried it during the partial task.

The interactive activities have gained traction in the last few years, not only in BioCreative. For example, in recognition of potential barriers that may inhibit the widespread adoption of biomedical software, the 2014 i2b2 Challenge introduced a special track, Track 3 - Software Usability Assessment, which indeed highlighted usability problems and therefore limitation of use/adoption of biomedical software [18]. Also in parallel to this interactive track, BioCreative V has introduced the Collaborative Biocurator Assistant Task (BioC) which explores the integration of the BioC format output from different TM modules to provide a system for literature curation of protein-protein interactions tailored for the BioGrid Database.

We have asked both the teams and the users about the experience in participating in the IAT activity. Although somewhat chaotic, both groups find participation a positive experience overall. We hope to improve the task based on the experience gained this year and results from the BioC track.

4 Acknowledgment

We would like to acknowledge all the users and team members who have participated in this activity. BioCreative is partially funded by NIH/NIGMS R13-GM109648-01A1; DOE: DE-SC0010838; NIH Intramural Research Program, National Library of Medicine.

5 References

1. Hirschman, L., et al., *Overview of BioCreAtIvE: critical assessment of information extraction for biology.* BMC Bioinformatics, 2005. **6**(Suppl 1): p. S1.
2. Krallinger, M., et al., *Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge.* Genome Biology, 2008. **9**(Suppl 2): p. S1.
3. Leitner, F., et al., *An Overview of BioCreative II.5.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2010. **7**(3): p. 385 - 399.
4. Arighi, C., et al., *Overview of the BioCreative III Workshop.* BMC Bioinformatics, 2011. **12**(Suppl 8): p. S1.
5. Wu, C.H., et al., *Editorial: BioCreative-2012 Virtual Issue.* Database (Oxford), 2012(bas).
6. Hirschman, L., et al., *Text mining for the biocuration workflow.* Database (Oxford), 2012. **2012**: p. bas020.
7. Lu, Z. and L. Hirschman, *Biocuration workflows and text mining: overview of the BioCreative 2012 Workshop Track II.* Database, 2012. **17**(10).
8. Arighi, C., et al., *BioCreative III interactive task: an overview.* BMC Bioinformatics, 2011. **12**(Suppl 8): p. S4.
9. Arighi, C. , et al., *An Overview of the BioCreative 2012 Workshop Track III: Interactive Text Mining Task.* Database (Oxford), 2012(bas).

10. Matis-Mitchell, S., et al. *BioCreative IV Interactive Task* in *Proceedings of the Fourth BioCreative Challenge Evaluation Workshop*. 2013. Bethesda, MD.
11. Bangor, A., P. Kortum, and J. Miller, *The System Usability Scale (SUS): An Empirical Evaluation.* International Journal of Human-Computer Interaction, 2008. **24**(6): p. 574–594.
12. Fu, X., et al., *Supporting the annotation of chronic obstructive pulmonary disease (COPD) phenotypes with text mining workflows.* Journal of Biomedical Semantics, 2015. **6**(1): p. 8.
13. Pafilis, E., et al., *The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text.* PLoS One, 2013. **8**(6).
14. Pafilis, E., et al., *ENVIRONMENTS and EOL: identification of Environment Ontology terms in text and the annotation of the Encyclopedia of Life.* Bioinformatics, 2015. **31**(11): p. 1872-4.
15. Pletscher-Frankild, S., et al., *DISEASES: text mining and data integration of disease-gene associations.* Methods, 2015. **74**: p. 83-9.
16. Santos, A., et al., *Comprehensive comparison of large-scale tissue expression datasets.* PeerJ, 2015. **30**(3).
17. Subramani, S., K. Raja, and J. Natarajan, *ProNormz--an integrated approach for human proteins and protein kinases normalization.* J Biomed Inform, 2014. **47**: p. 131-8.
18. Zheng, K., et al., *Ease of adoption of clinical natural language processing software: An evaluation of five systems.* J Biomed Inform, 2015. **22**(15): p. 00148-3.