BioCreative Challenges

Information Retrieval and Text Mining for Biology

SIB

June 4, 2015.

Cecilia N. Arighi, PhD

Research Associate Professor

Protein Information Resource

CBCB, University of Delaware

arighi@dbi.udel.edu

http://www.biocreative.org

Outline

- Overview of BioCreative Effort
- Evolution of Tasks
- Impact of BioCreative
- BioCreative User Interactive Task
- Concluding Remarks

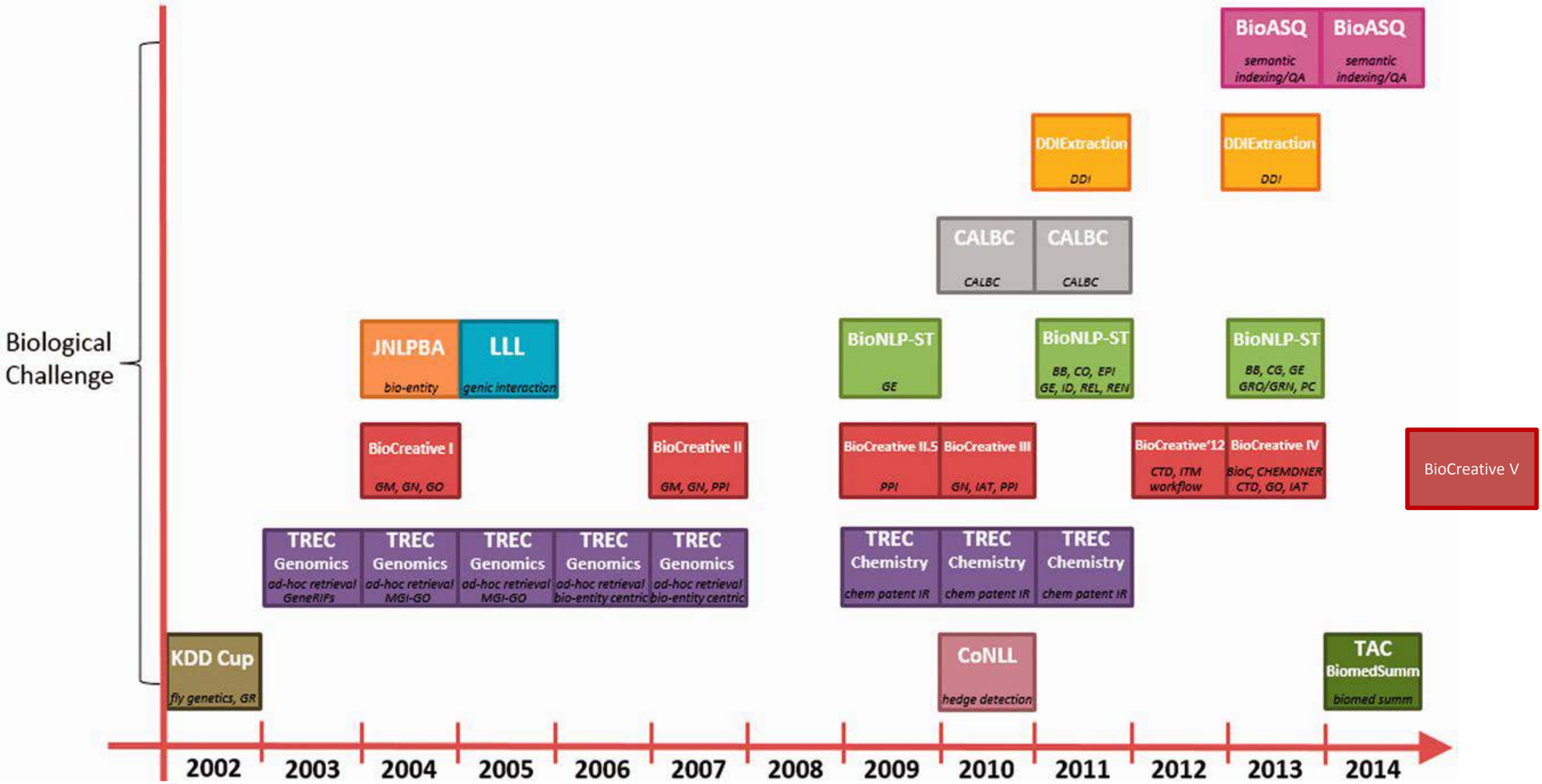
BioCreative

Community-wide effort for evaluating text mining systems applied to the biomedical domain

Collaborative and interdisciplinary effort

University of Delaware 1743
NCBI
cnio
MITRE
University of Colorado Denver | Anschutz Medical Campus
Universitas Turicensis
RGD
FlyBase
Bio Cre at ive
Fraunhofer
SGD
WormBase
MaizeGDB Genome Browser
tair
EMBL-EBI
Pfizer
IntAct
ctd
MGI
AstraZeneca Health Connects Us All
BioGRID
MINT
cima
CENTRO DE INVESTIGACIÓN MÉDICA APLICADA
UNIVERSIDAD DE NAVARRA
ELSEVIER

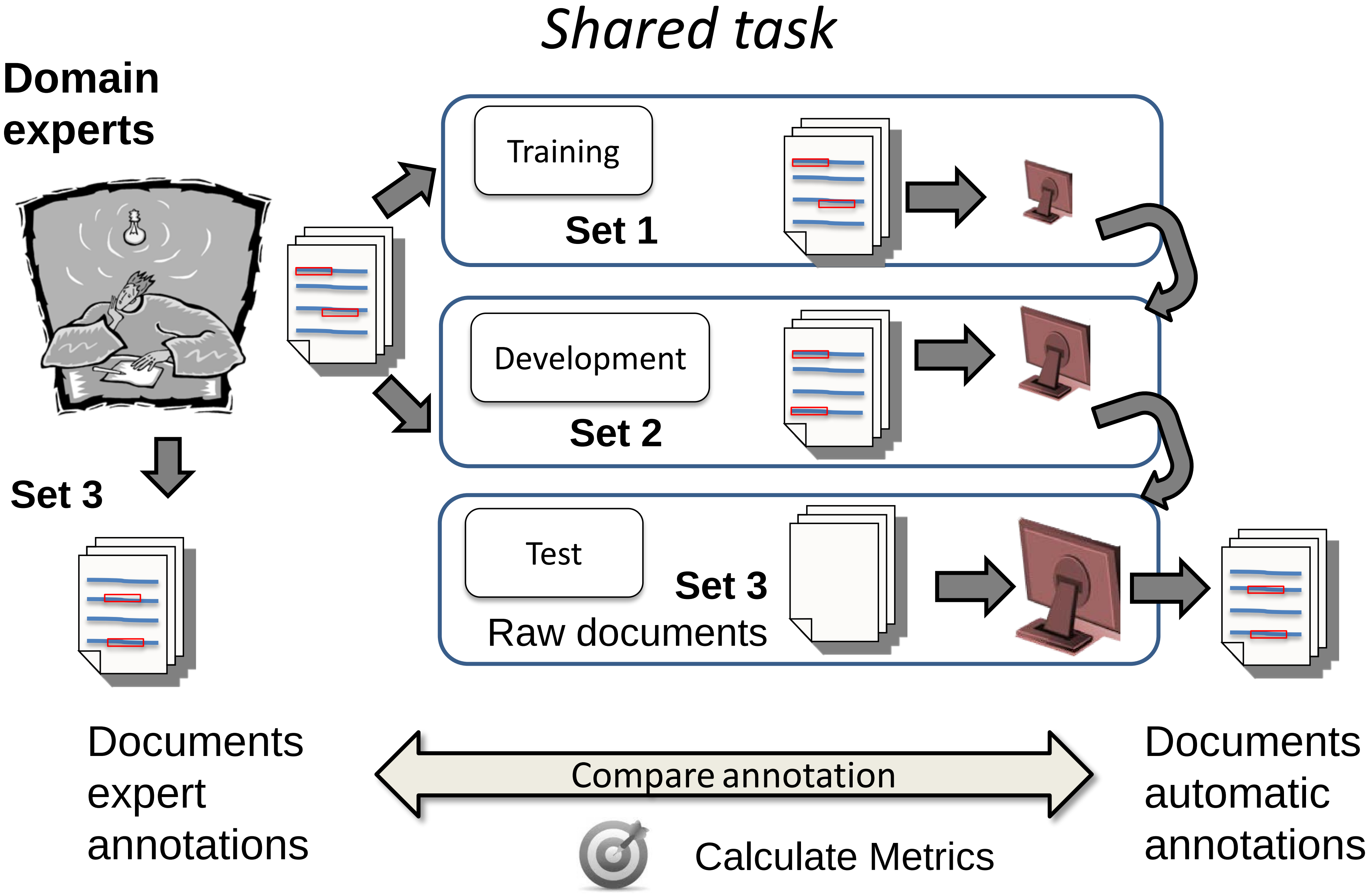
BioNLP challenges in chronological order


Biological Challenge
BioASQ semantic indexing/QA
BioASQ semantic indexing/QA
DDIExtraction DDI
DDIExtraction DDI
CALBC CALBC
CALBC CALBC
JNLPBA bio-entity
LLL genic interaction
BioNLP-ST GE
BioNLP-ST BB, CG, EPI GE, ID, REL, REN
BioNLP-ST BB, CG, GE GRO/GRN, PC
BioCreative I GM, GN, GO
BioCreative II GM, GN, PPI
BioCreative II.5 PPI
BioCreative III GN, IAT, PPI
BioCreative'12 CTD, ITM workflow
BioCreative IV BioC, CHEMDNER CTD, GO, IAT
BioCreative V
TREC Genomics ad-hoc retrieval GeneRIFs
TREC Genomics ad-hoc retrieval MGI-GO
TREC Genomics ad-hoc retrieval MGI-GO
TREC Genomics ad-hoc retrieval bio-entity centric
TREC Genomics ad-hoc retrieval bio-entity centric
TREC Chemistry chem patent IR
TREC Chemistry chem patent IR
TREC Chemistry chem patent IR
KDD Cup fly genetics, GR
CoNLL hedge detection
TAC BiomedSumm biomed summ
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014

Adapted from Chung-Chi Huang, and Zhiyong Lu Brief Bioinform 2015;bib.bbv024

Briefings in Bioinformatics

Traditional Tracks in BioNLP Challenges



What is special about BioCreative?

BioCreative Goals

Intends to:

- Attract researchers from both natural language processing and biomedical domain
- Address problems of importance to the biology and bioinformatics community (focus on biocuration)
- Create legacy training and test data suites that could be used for development and benchmarking of future applications
- Allow the assessment of the state-of-the-art on real biological tasks

• Design of tracks based on user needs

GO curation: identification of articles with curatable GO information

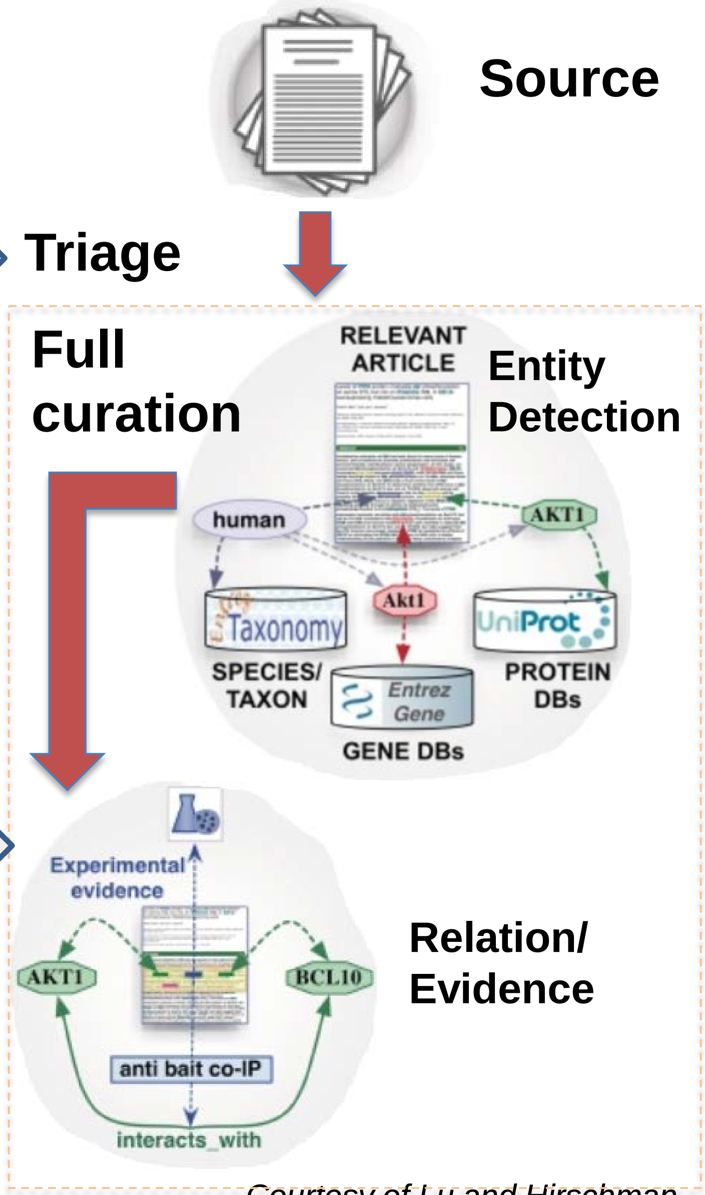
PPI curation: selection of relevant PPI articles from PubMed

CTD: identification of gene, chemical, disease and link to CTD vocabulary

PPI curation: Extracting interactant pair, PPI method

ChemDNER: Identification of chemical entities

Generic biocuration workflow



Source

Triage

Full curation

Entity Detection

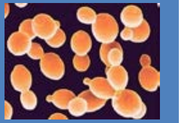
Relation/ Evidence

Courtesy of Lu and Hirschman

• Builds on previous tracks

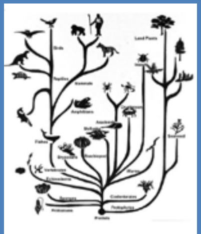
Gene normalization (linking a gene mention to database identifier)

BCI Abstracts



BCII Abstracts

BCIII Full-length



Species are known

Closer to real scenario

Chemical recognition

(identifying compound names in text)

BCIV Abstracts

BCV Patents

Noisy data

• Fosters interoperability

Some problems faced by shared-task challenges:

- Many different formats
- Many new projects start over
- An atmosphere of competition

The Needs?

- Common format
- Simple-to-learn software to access the format
- Sufficient resources to motivate users

A Solution

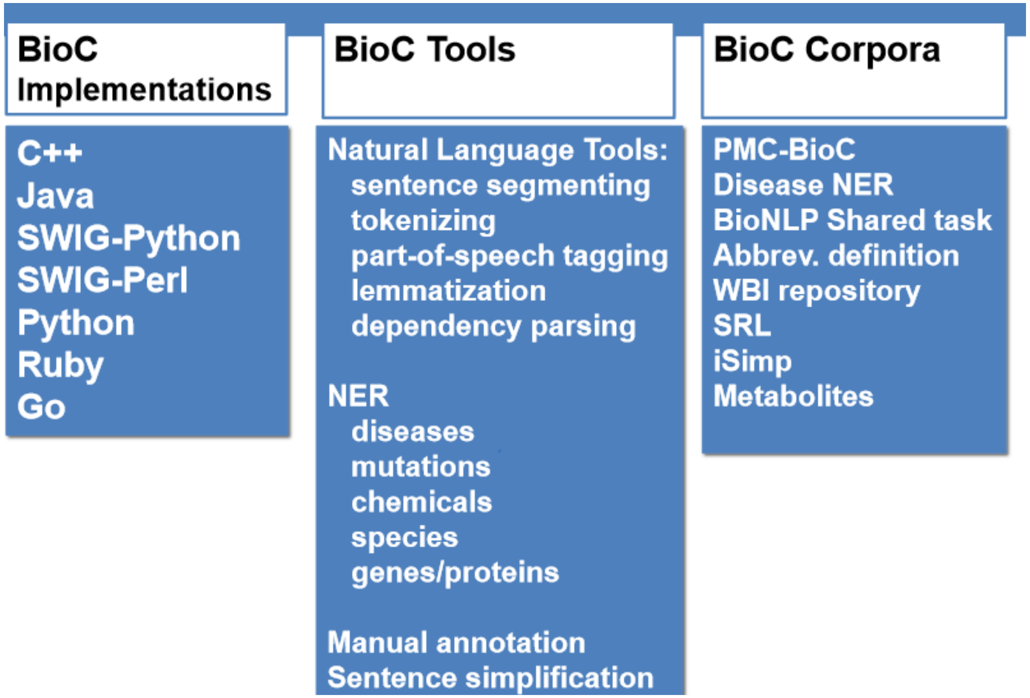
- A convenient format to share text documents and annotations
- A library to promote interoperability of data and tools

• Fosters interoperability

BioC, a BioCreative interoperability initiative, is a simple extensible XML language format to share text data and annotations

Goals:

- o simplicity
- o interoperability
- o broad use and reuse



BioC Implementations

C++
Java
SWIG-Python
SWIG-Perl
Python
Ruby
Go

BioC Tools

Natural Language Tools:
sentence segmenting
tokenizing
part-of-speech tagging
lemmatization
dependency parsing

NER
diseases
mutations
chemicals
species
genes/proteins

Manual annotation
Sentence simplification

BioC Corpora

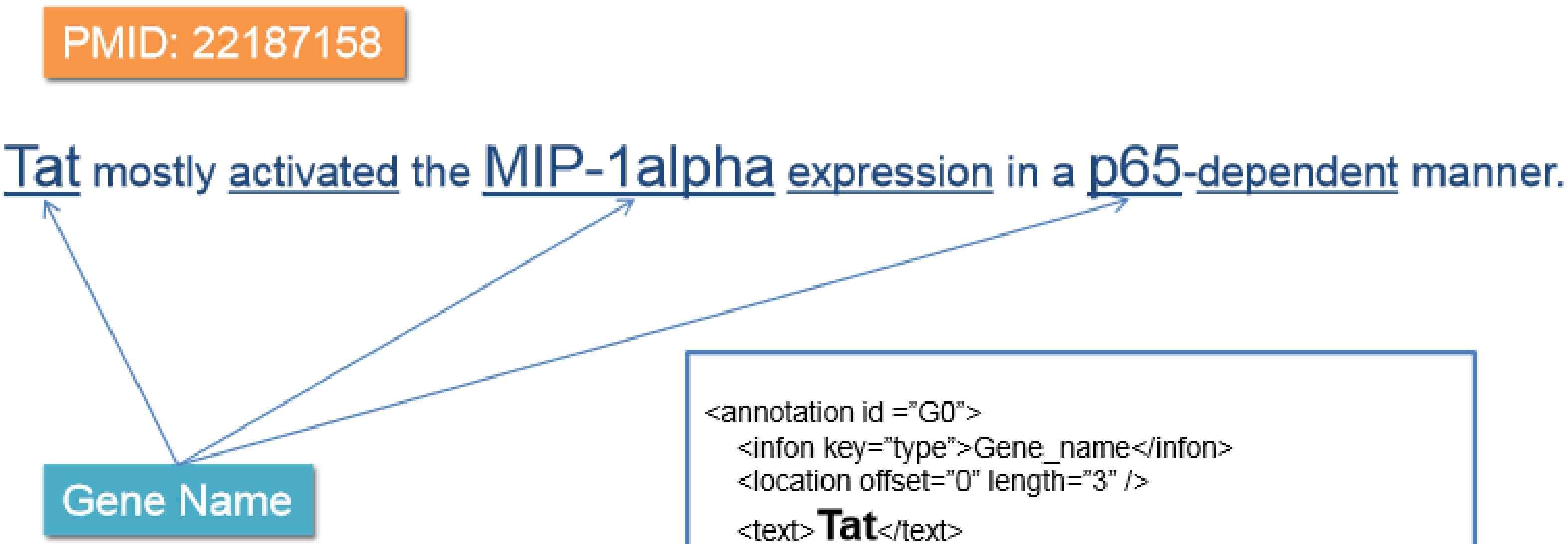
PMC-BioC
Disease NER
BioNLP Shared task
Abbrev. definition
WBI repository
SRL
iSimp
Metabolites

Syntax is XML defined by a Document Type Definition (DTD)
Key file describes content of XML file

Comeau et al., Database (Oxford);2013:bat064.

http://bioc.sourceforge.net/

• DTD example



PMID: 22187158

Tat mostly activated the MIP-1alpha expression in a p65-dependent manner.

Gene Name

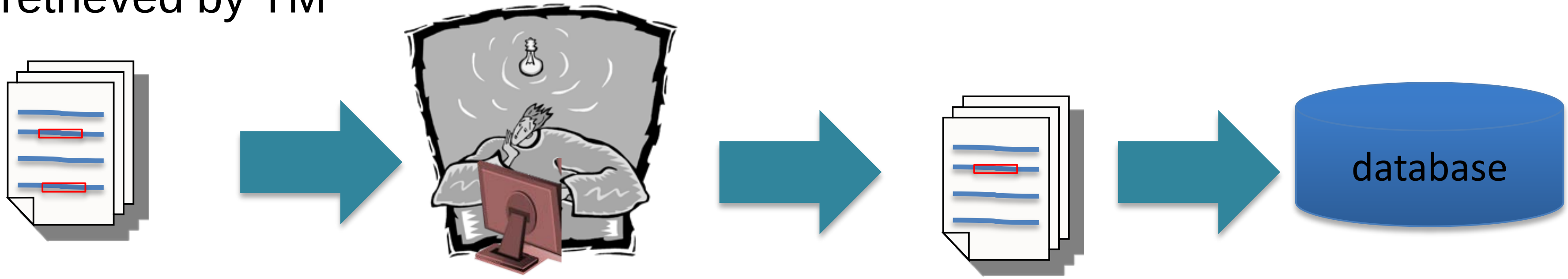
```
<annotation id =”G0”>
   <infon key=”type”>Gene_name</infon>
   <location offset=”0” length=”3” />
   <text>Tat</text>
</annotation>

<annotation id =”G1”>
   <infon key=”type”>Gene_name</infon>
   <location offset=”25” length=”10” />
   <text>MIP-1alpha</text>
</annotation>

<annotation id =”G2”>
   <infon key=”type”>Gene_name</infon>
   <location offset=”52” length=”3” />
   <text>p65</text>
</annotation>
```

http://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/BioC_ISMB2014.pd.

• Explores text mining systems via users

BioNLP evaluations have focused on isolated tasks, they have emphasized ‘off-line’ accuracy measures

But in the biocuration world….



Documents automatically annotated, or retrieved by TM

Reviewed by curator

Curated document

Information stored in DB

Interactive Task (IAT): Evaluation of text mining systems by potential users and report on performance and usability

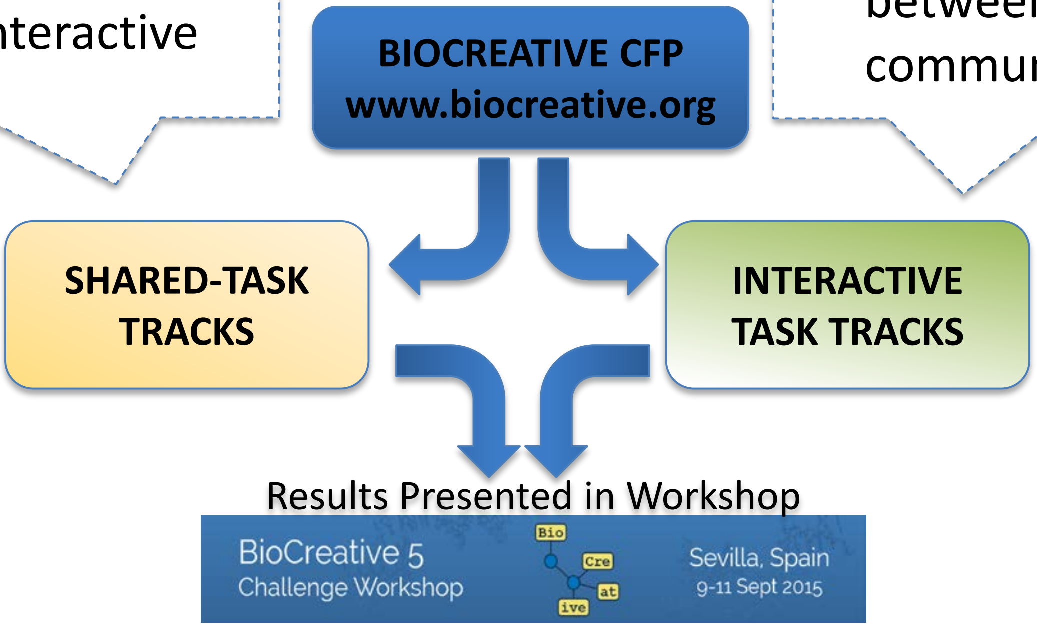
• BioCreative goes beyond the shared task

Competitive

- Task relevant to biomedical domain
- Drive state-of-the-art TM tools development
- Provide building modules for systems in interactive task

Non-competitive

- System development for literature curation tasks
- Tested by users
- Foster interaction between bioNLP and user communities



BIOCREATIVE CFP
www.biocreative.org

SHARED-TASK TRACKS

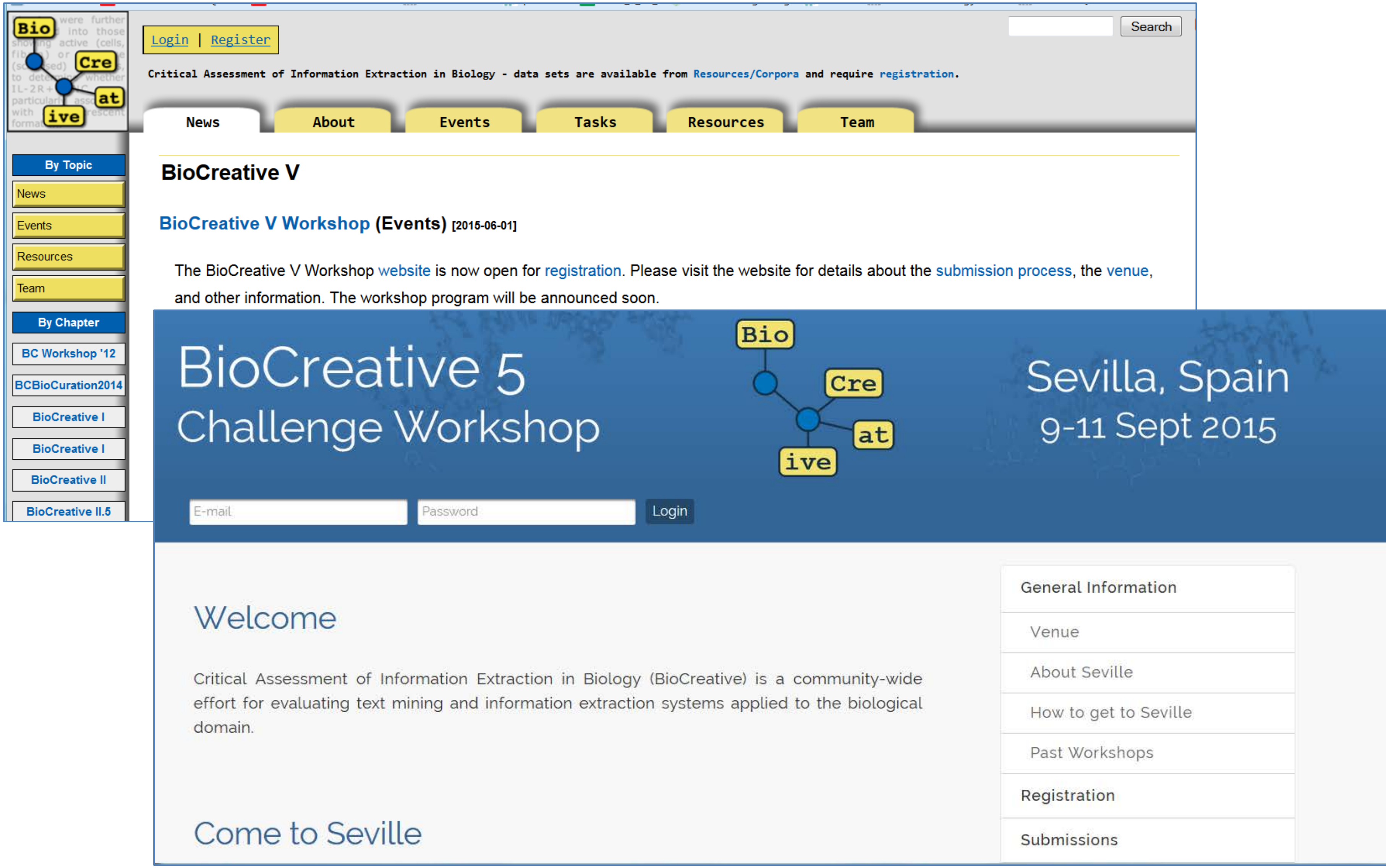
INTERACTIVE TASK TRACKS

Results Presented in Workshop

		Tasks	Publications
BC I	Spain, 2004	• Gene Mention • Gene Normalization • GO	BMC Bioinformatics 2005, 6 (Suppl 1)
BC II	Spain, 2007	• Gene Mention • Gene Normalization • Protein-protein Interaction	Genome Biology 2008, 9 (Suppl 2)
BC II.5	Spain, 2009	Protein-Protein Interaction: • Interactor Normalization • Interaction Pair • Article Categorization	IEEE Transactions in Computational Biology and Bioinformatics 2010
BC III	USA, 2010	• Gene Normalization • User Interactive Task • Protein-protein Interaction	BMC Bioinformatics 2011
BC 2012	USA, 2012	• CTD • Biocuration Workflow • User Interactive Task	Database Virtual Issue 2012
BC IV	USA, 2013	• Interoperability • ChemDNER • CTD • GO • User Interactive Task	Database Virtual Issue 2014 Chemical Informatics

BioCreative V Challenge Workshop

http://www.biocreative.org/



Critical Assessment of Information Extraction in Biology - data sets are available from Resources/Corpora and require registration.

BioCreative V

BioCreative V Workshop (Events) [2015-06-01]

The BioCreative V Workshop website is now open for registration. Please visit the website for details about the submission process, the venue, and other information. The workshop program will be announced soon.

BioCreative 5 Challenge Workshop

Sevilla, Spain
9-11 Sept 2015

Welcome

Critical Assessment of Information Extraction in Biology (BioCreative) is a community-wide effort for evaluating text mining and information extraction systems applied to the biological domain.

Come to Seville

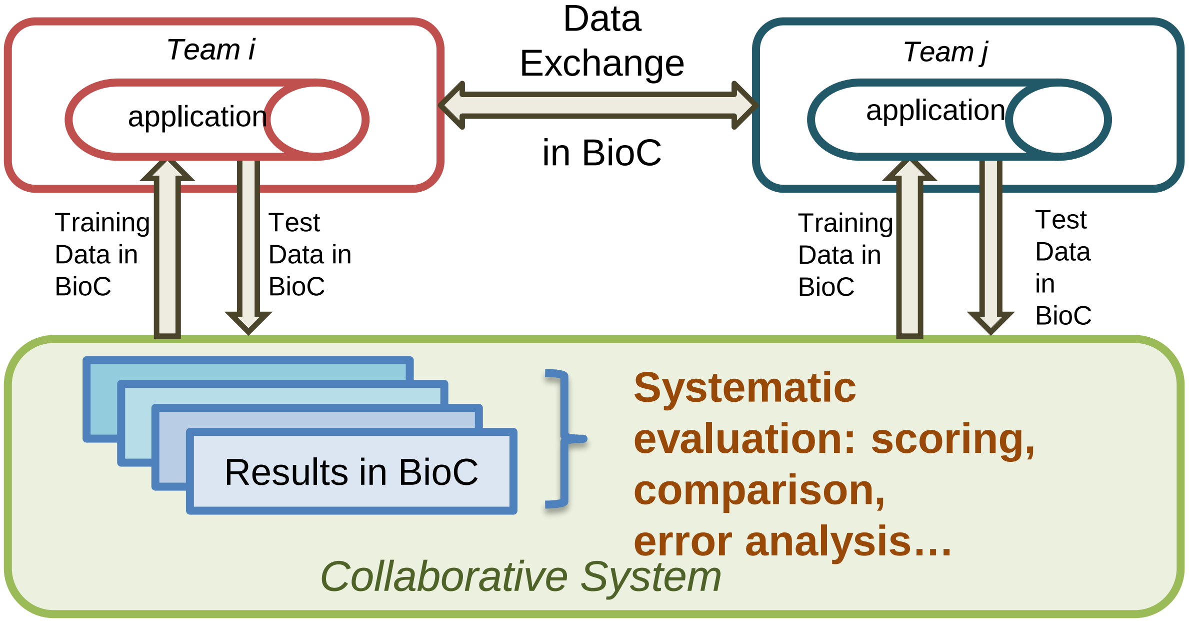
Track 1- Collaborative BioCurator Assistant Task (BioC)

Goal:
Build a complete system to assist BioGrid curation

Task Organizers:
Sun Kim, Rezarta Islamaj Doğan, Donald C. Comeau, W. John Wilbur (NCBI), Andrew Chatr-aryamontri (BioGrid)

Subtasks:

- Protein, organism
- Physical, genetic interaction
- Experimental method
- Visualization tool



Team i

application

Training Data in BioC

Test Data in BioC

Data Exchange

in BioC

Team j

application

Training Data in BioC

Test Data in BioC

Results in BioC

Systematic evaluation: scoring, comparison, error analysis…

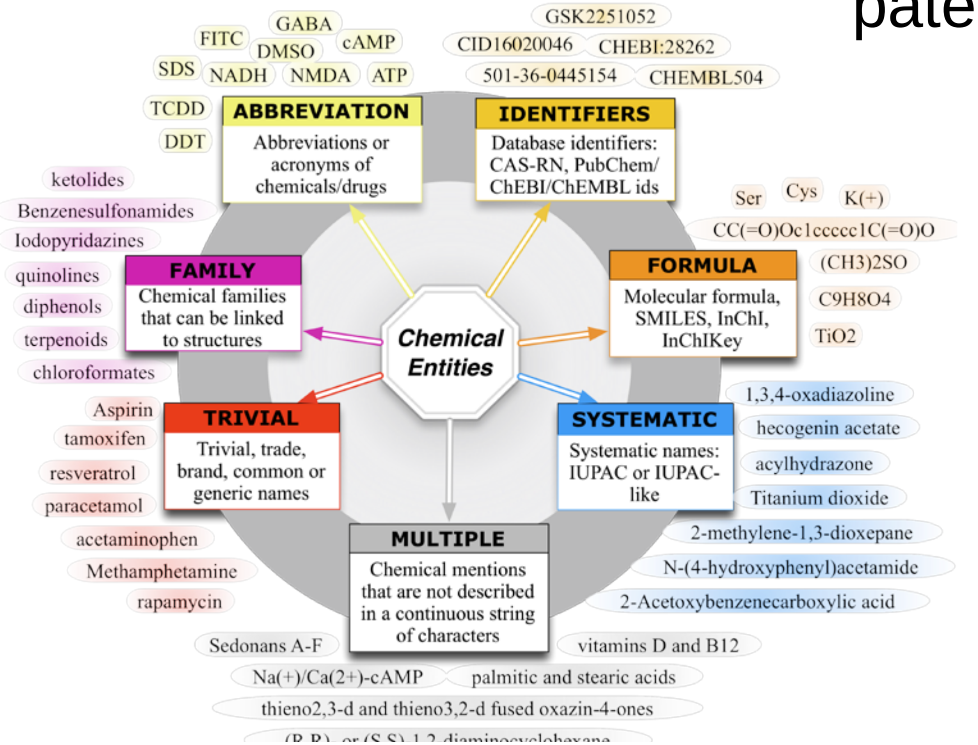
Collaborative System

Corpora:

- PubMed abstracts
- PubMed Central full text articles

Track 2- CHEMDNER Patents

Goal: Automatic extraction of chemical and biological data from patents



GABA
FITC
cAMP
DMSO
SDS
NADH
NMDA
ATP
GSK2251052
CID16020046
CHEBI:28262
501-36-0445154
CHEMBL504
TCDD
DDT
ABBREVIATION
Abbreviations or acronyms of chemicals/drugs
IDENTIFIERS
Database identifiers: CAS-RN, PubChem/ ChEBI/ChEMBL ids
Ser
Cys
K(+)
CC(=O)Oc1ccccc1C(=O)O
ketolides
Benzenesulfonamides
Iodopyridazines
quinolines
diphenols
terpenoids
chloroformates
FAMILY
Chemical families that can be linked to structures
Chemical Entities
FORMULA
Molecular formula, SMILES, InChI, InChIKey
(CH3)2SO
C9H8O4
TiO2
Aspirin
tamoxifen
resveratrol
paracetamol
acetaminophen
Methamphetamine
rapamycin
TRIVIAL
Trivial, trade, brand, common or generic names
SYSTEMATIC
Systematic names: IUPAC or IUPAC-like
1,3,4-oxadiazoline
hecogenin acetate
acylhydrazone
Titanium dioxide
2-methylene-1,3-dioxepane
N-(4-hydroxyphenyl)acetamide
2-Acetoxybenzenecarboxylic acid
MULTIPLE
Chemical mentions that are not described in a continuous string of characters
Sedonans A-F
vitamins D and B12
Na(+)/Ca(2+)-cAMP
palmitic and stearic acids
thieno2,3-d and thieno3,2-d fused oxazin-4-ones
(R,R)- or (S,S)-1,2-diaminocyclohexane

Task Organizers:

Martin Krallinger & Alfonso Valencia (CNIO), Florian Leitner (UPM), Obdulia Rabal & Julen Oyarzabal (CIMA)

Builds on successful task on abstracts 3,000 abstracts test set, 91% inter-annotator agreement. F-score 87.39%

Corpora: 30,000 manually annotated medicinal chemistry patent abstracts

Subtasks:

- **CEMP subtask** (chemical entity mention in patents)
- **CPD subtask** (chemical passage detection): the detection patent abstracts mentioning chemicals (text classification/triage)
- **GPRO subtask** (gene and protein related object task): gene/protein mentions

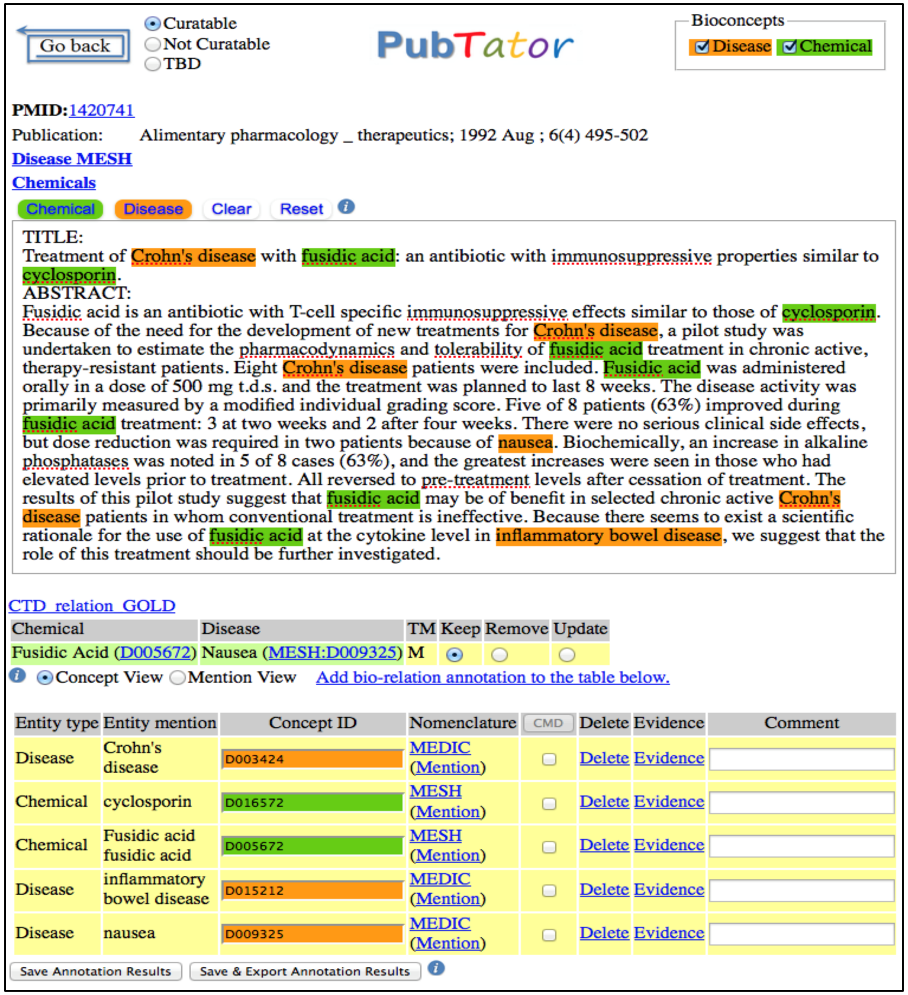
Track 3- Chemical-disease relation (CDR)

Goal: Advance the field in relation extraction from biomedical literature

Task Organizers:
Zhiyong Lu (NCBI),
Thomas Wiegers (CTD)

Subtasks:

- Disease Named Entity Recognition
- Chemical-induced disease relation extraction



● Curatable
○ Not Curatable
○ TBD

Bioconcepts
☑ Disease ☑ Chemical

PMID:1420741
Publication: Alimentary pharmacology _ therapeutics; 1992 Aug ; 6(4) 495-502
Disease MESH
Chemicals

TITLE:
Treatment of Crohn's disease with fusidic acid: an antibiotic with immunosuppressive properties similar to cyclosporin.
ABSTRACT:
Fusidic acid is an antibiotic with T-cell specific immunosuppressive effects similar to those of cyclosporin. Because of the need for the development of new treatments for Crohn's disease, a pilot study was undertaken to estimate the pharmacodynamics and tolerability of fusidic acid treatment in chronic active, therapy-resistant patients. Eight Crohn's disease patients were included. Fusidic acid was administered orally in a dose of 500 mg t.d.s. and the treatment was planned to last 8 weeks. The disease activity was primarily measured by a modified individual grading score. Five of 8 patients (63%) improved during fusidic acid treatment: 3 at two weeks and 2 after four weeks. There were no serious clinical side effects, but dose reduction was required in two patients because of nausea. Biochemically, an increase in alkaline phosphatases was noted in 5 of 8 cases (63%), and the greatest increases were seen in those who had elevated levels prior to treatment. All reversed to pre-treatment levels after cessation of treatment. The results of this pilot study suggest that fusidic acid may be of benefit in selected chronic active Crohn's disease patients in whom conventional treatment is ineffective. Because there seems to exist a scientific rationale for the use of fusidic acid at the cytokine level in inflammatory bowel disease, we suggest that the role of this treatment should be further investigated.

CTD_relation_GOLD

Chemical	Disease	TM	Keep	Remove	Update
Fusidic Acid (D005672)	Nausea (MESH:D009325)	M	●	○	○

○ Concept View ○ Mention View   Add bio-relation annotation to the table below.

Entity type	Entity mention	Concept ID	Nomenclature	CMD	Delete	Evidence	Comment
Disease	Crohn's disease	D003424	MEDIC (Mention)	□	Delete	Evidence	
Chemical	cyclosporin	D016572	MESH (Mention)	□	Delete	Evidence	
Chemical	Fusidic acid fusidic acid	D005672	MESH (Mention)	□	Delete	Evidence	
Disease	inflammatory bowel disease	D015212	MEDIC (Mention)	□	Delete	Evidence	
Disease	nausea	D009325	MEDIC (Mention)	□	Delete	Evidence	

Track 4- Causal Networks in BEL

What is BEL?

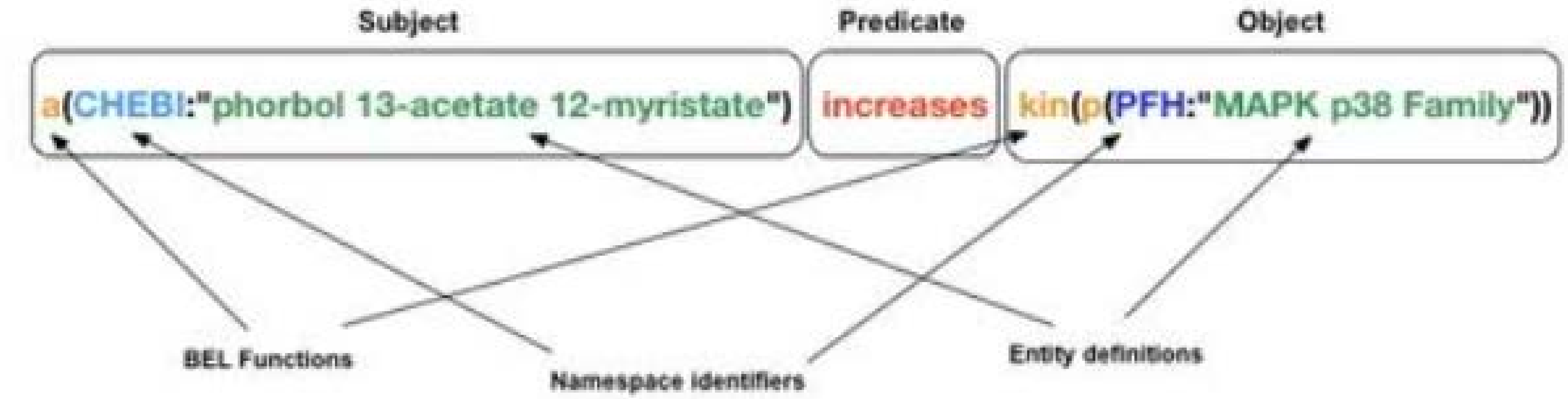
Biological Expression Language

Computable knowledge representation

Task Organizers:

Fabio Rinaldi (UZurich)

Juliane Fluck (Fraunhofer)



Corpus: 50 biological networks, 180,000 relationships

Subtasks:

- Generation of BEL statements given the evidence
- Find evidence for a given BEL statement

http://www.openbel.org/content/bel-lang-language-structure

Track 5- User Interactive Task

Task Organizers:
Cecilia Arighi, Qinghua Wang (PIR, UDel) and Lynette Hirschman (MITRE)

Evaluation of text mining tools by users

It is a *demonstration interactive* task

Need to involve users

↓

User Advisory Group (UAG)

A diverse sample of end users with multiple text mining needs

- Help to develop end user requirements for interactive text mining tools
- Serve as users for the interactive task
- Assist in corpora annotation for biocreative tasks
- Help in recruiting biocurators

UAG BioCreative V

Chairs: Cecilia Arighi and Zhiyong Lu

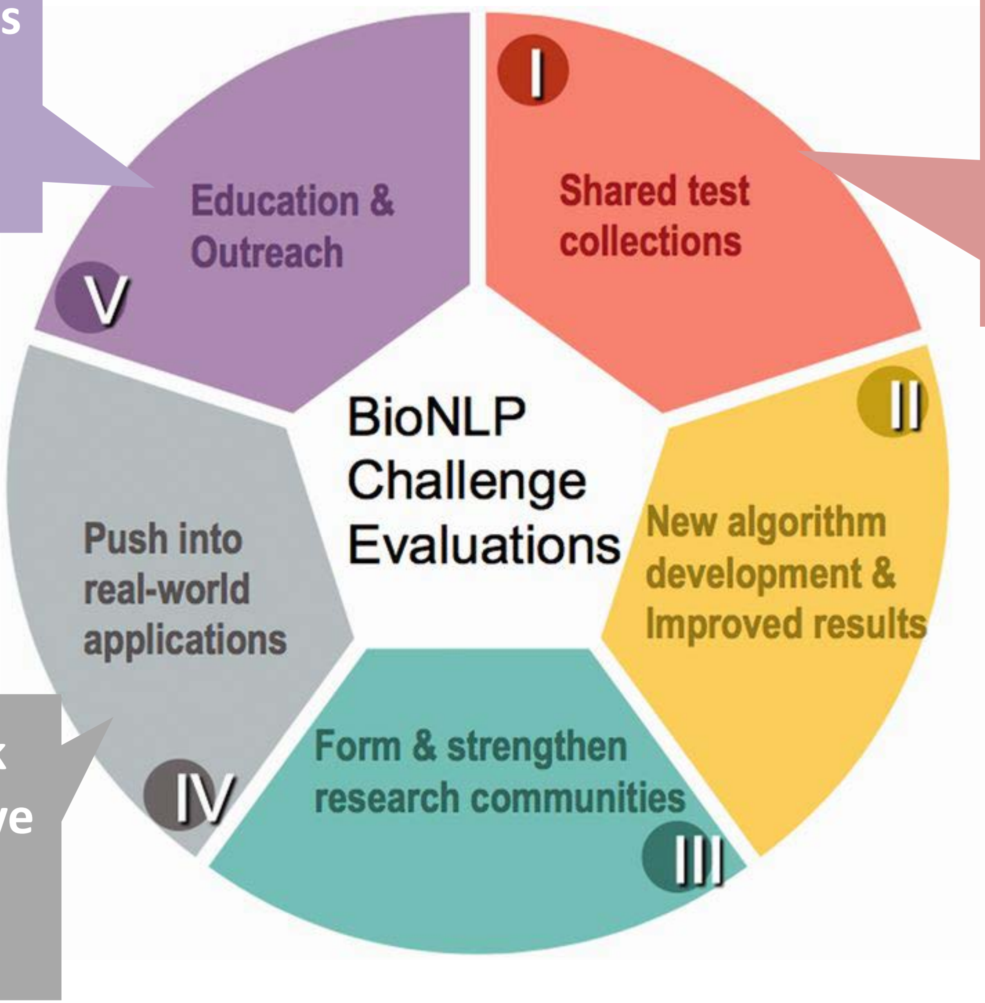
Andrew Chatr-aryamontri
Raul Rodriguez-Esteban
Stan Laulederkind
Sherri Matis-Mitchell
Johanna McEntyre
Peter McQuilton
Evangelos Pafilis
Sandra Orchard
Sangya Pundir
Mary Schaeffer
Kimberly Van Auken

BioGRID
Roche
RGD
ELSEVIER
Europe PubMed Central
FlyBase
hcmr
IntAct
EMBL-EBI
MaizeGDB Genome Browser
WormBase

Impact/contributions

Workshops at scientific meetings
Webinars (BioC)
Publications

BioC corpora
GN corpora
PPI corpora
CTD/CDR corpora
ChEMDNER corpora


BioNLP Challenge Evaluations
I
Shared test collections
II
New algorithm development & Improved results
III
Form & strengthen research communities
IV
Push into real-world applications
V
Education & Outreach

Interactive Task
BioC collaborative task

Modified from Chung-Chi Huang, and Zhiyong Lu Brief Bioinform 2015;bib.bbv024

Briefings in Bioinformatics

II- New Developments

BioCreative has promoted the development for state-of-the-art solutions

- Gene Mention: AIIA-GMT
- Gene Normalization: GNAT, GenNorm, ProMiner
- PPI triage: PIE
- GO categorization: GoCat

A variety of methods have been applied:

- Markov models
- Machine learning
- Rule-based
- Naïve Bayes classifiers
- Support Vector Machine

II- Experiment with New Metrics

Threshold Average Precision (TAP-k)

In Gene normalization BioCreative III

Derivative of Mean average precision (MAP) with a threshold determined by the first k errors in the ranked list.

TAP-*k* is able to measure ranking, reflect the user tolerance of prediction errors (false positives), as well as make use of confidence scores.

Hierarchical Precision, Recall and F scores

In Gene Ontology task BioCreative IV

Given the hierarchical nature of GO, considers common parent terms in computer-predicted and human-annotated GO terms

IV-Form and streghten research communities

Successful BioNLP-user interactions through BioCreative

PubTator
Used by NLM for indexing. Currently being used by UniProt curators

TagTog
Gene indexing in Flybase (Cejuela et al., PMID:24715220)

ODIN
In PharmGKB workflow (Rinaldi et al., PMID: 22529178) now being tested on RegulonDB for BioCreative V

RLIMS-P
Phosphogrid curation (Torii et al., PMID:25122463)

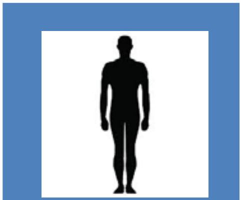
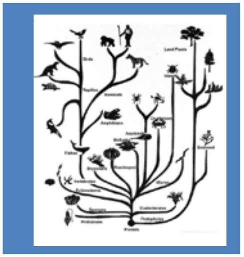
Impact/contributions

	# articles in ePMC
BioCreative Editorials	6
BioCreative mentioned in title or abstract	148
BioCreative is found in reference section	389

Top 10 MeSH terms in articles referencing BioCreative	**Frequency**
Humans	119
Natural Language Processing	98
Algorithms	81
Databases (Factual, genetic, protein)	79
Software	78
Computational Biology/methods	69
Vocabulary, Controlled	61
Artificial Intelligence	60
Information Storage and Retrieval/methods	51
Data Mining/methods	50

Gene Normalization in BioCreative

Convention for protein naming are different in different organisms. Differ in number of synonyms, in complexity of names,

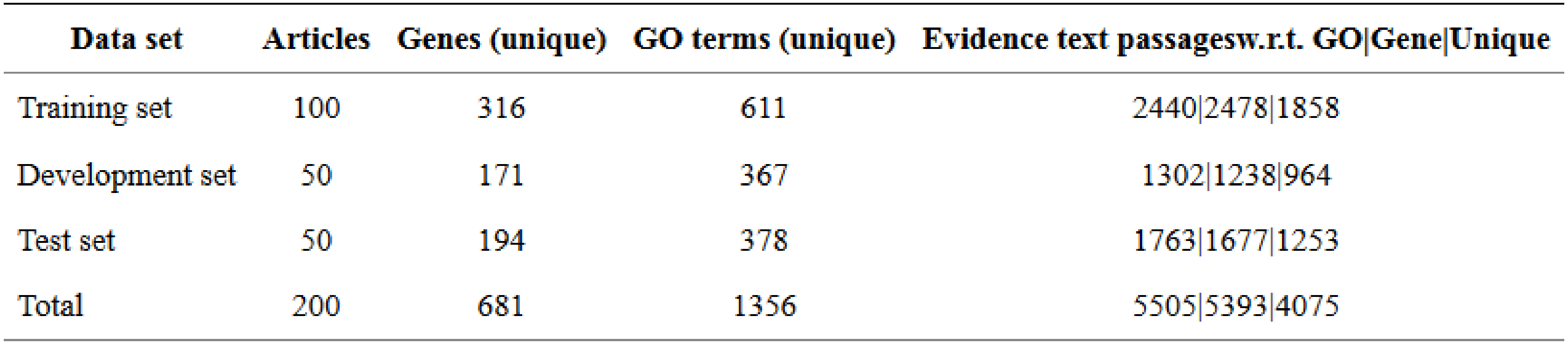
	BC I			BC II	BC III
					
F-Score	0.92	0.82	0.79	0.81	0.50
IAA %	87	91	69	91	-
Average synonyms per identifier	1.86	2.94	2.48	5.5	
Average synonym length in words	1	1.47	2.77	2.17	

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2559987/

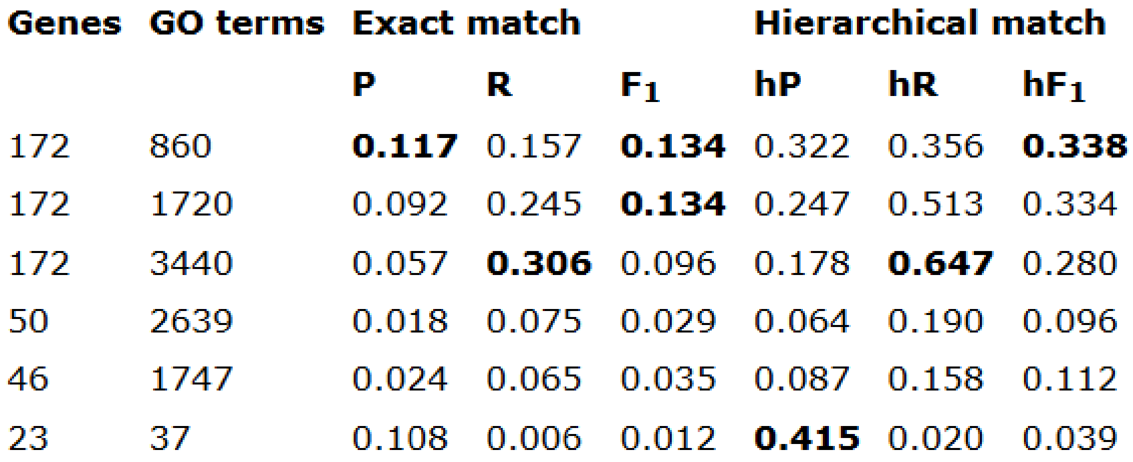
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3269937/

GO task

Overall statistics of the annotated corpus grouped by data sets



Data set	Articles	Genes (unique)	GO terms (unique)	Evidence text passagesw.r.t. GO|Gene|Unique
Training set	100	316	611	2440|2478|1858
Development set	50	171	367	1302|1238|964
Test set	50	194	378	1763|1677|1253
Total	200	681	1356	5505|5393|4075



Genes	GO terms	Exact match			Hierarchical match		
		P	R	F_1	hP	hR	hF_1
172	860	**0.117**	0.157	**0.134**	0.322	0.356	**0.338**
172	1720	0.092	0.245	**0.134**	0.247	0.513	0.334
172	3440	0.057	**0.306**	0.096	0.178	**0.647**	0.280
50	2639	0.018	0.075	0.029	0.064	0.190	0.096
46	1747	0.024	0.065	0.035	0.087	0.158	0.112
23	37	0.108	0.006	0.012	**0.415**	0.020	0.039

IAA GO term selection
47% strict
62% hierarchical

Yuqing Mao et al. Database 2014;bau086
Van Auken et al., Database 2014, bau074

Database: The Journal of Biological Databases and Curation

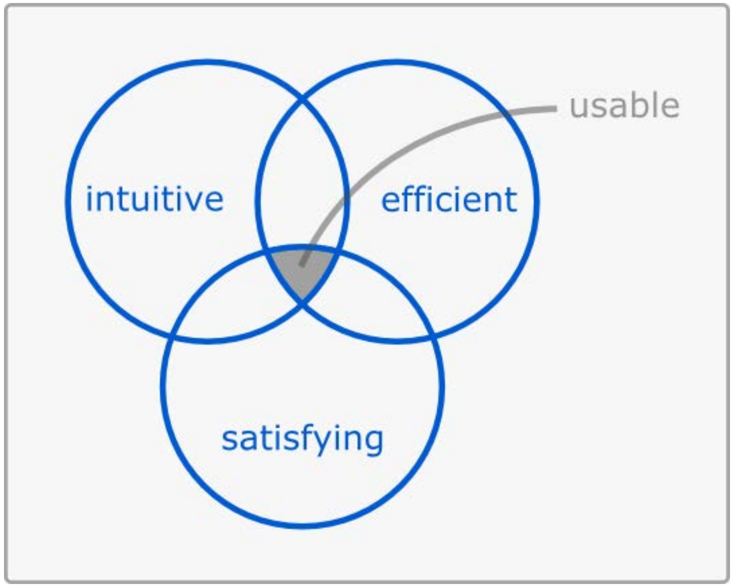
General Remarks

- The Tasks have evolved to resemble more the real scenario
- Improvements have been achieved in many tasks
- Combination of methods usually improves the performance
- Although results are not of sufficient quality to use as an entirely automated process, output from these tools can provide a head start for curators

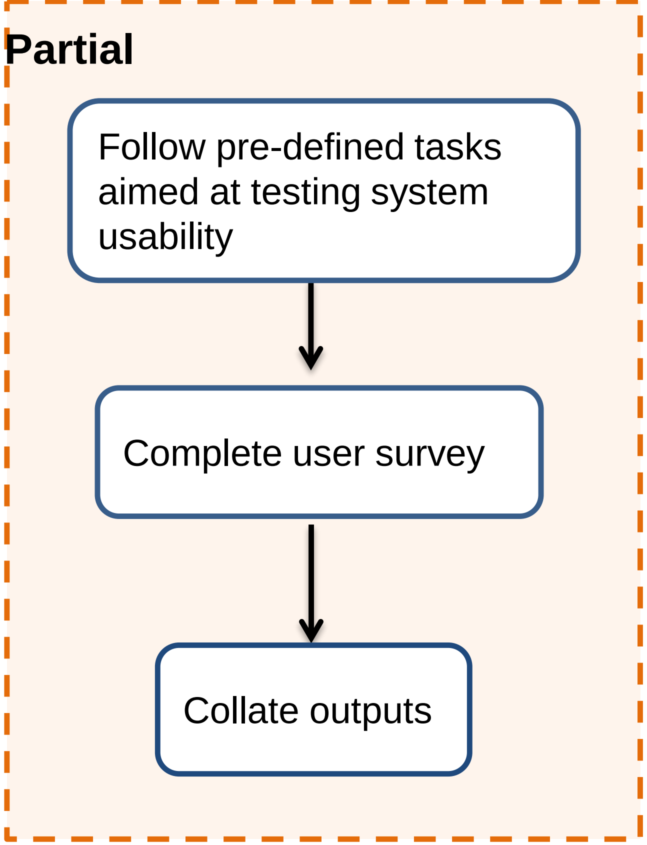
Interactive Task

Can text mining tools help in Biocuration

Our idea is to expose text mining systems to biocurators so they can provide feedback on the system and become adopters in the future



Two levels of user participation

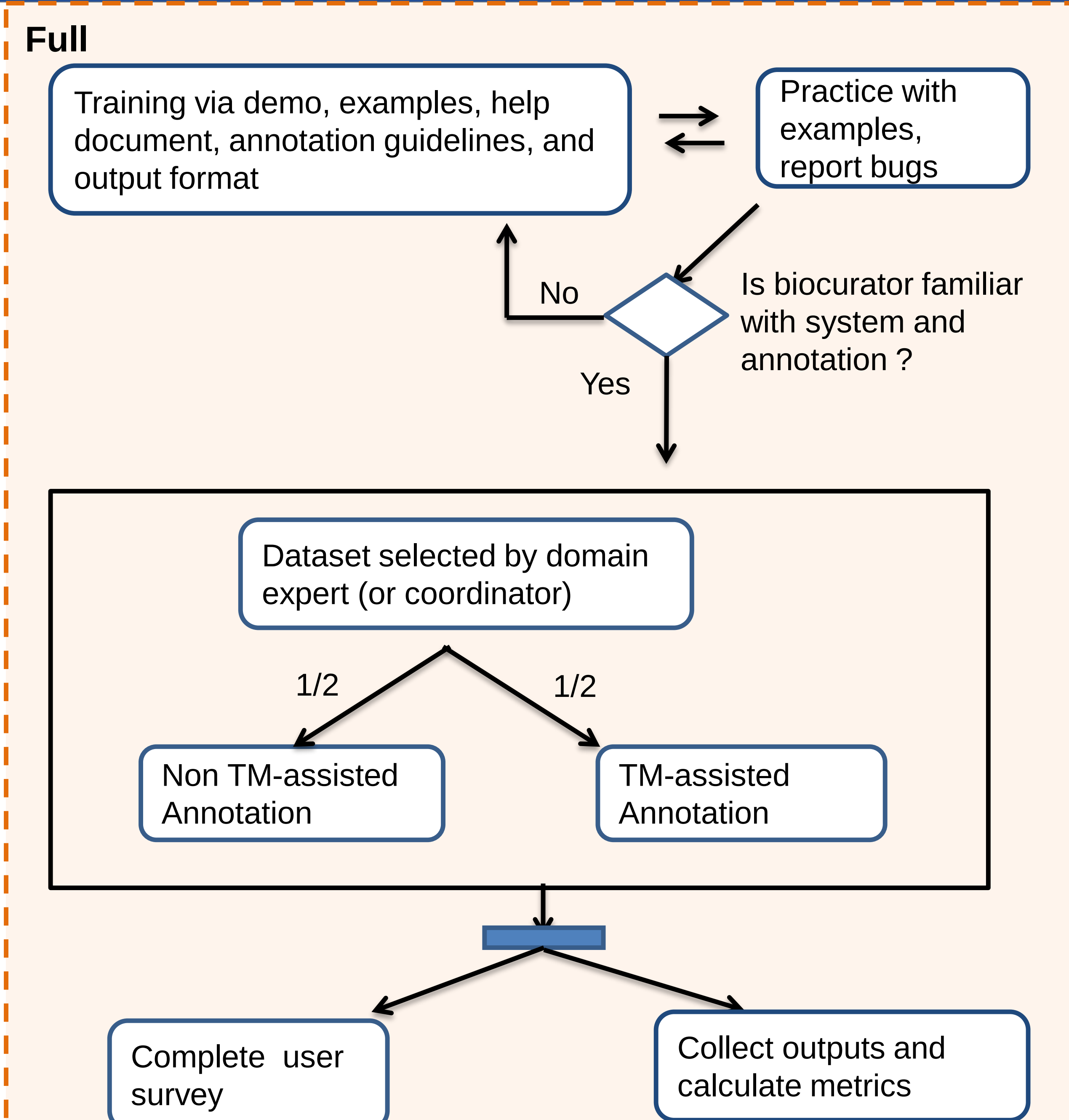


Partial

Follow pre-defined tasks aimed at testing system usability

Complete user survey

Collate outputs



Full

Training via demo, examples, help document, annotation guidelines, and output format

Practice with examples, report bugs

Is biocurator familiar with system and annotation ?

No

Yes

Dataset selected by domain expert (or coordinator)

1/2

1/2

Non TM-assisted Annotation

TM-assisted Annotation

Complete user survey

Collect outputs and calculate metrics

User Survey

Based on Questionnaire for User Interface Satisfaction (QUIS)

Five main categories:

1. Overall reaction
2. System’s ability to help complete tasks
3. Design of application
4. Learning to use the application
5. Usability

Goal: Try to find correlation of response to questions in survey with overall system satisfaction to learn what aspects are important to users

http://ir.cis.udel.edu/biocreative/survey.html

http://ir.cis.udel.edu/biocreative/survey2.html

Highlights of Previous Runs

1-Match between system and the real world (of Biocuration)

- The system should speak the users' language rather than system-oriented terms
- Systems should follow standards of its user community
- Sentence vs. Document level annotation

2-Testing the Systems NOT the Users

- Participants not being tested. But in the context of this activity we need to distinguish the participants into curation novice vs. expert because it has an impact on the performance

3-Documentation: Annotation guidelines and tutorials

- Provide detail annotation guidelines for the task
- Provide tutorial for system training with hands-on examples

Highlight of Previous Runs

4-System performance and functionalities in interface

- System performance is a key aspect to biocurators, but coupling results with functionalities that assist in easily correcting or finding additional information is very important for an interactive system

5-System Output

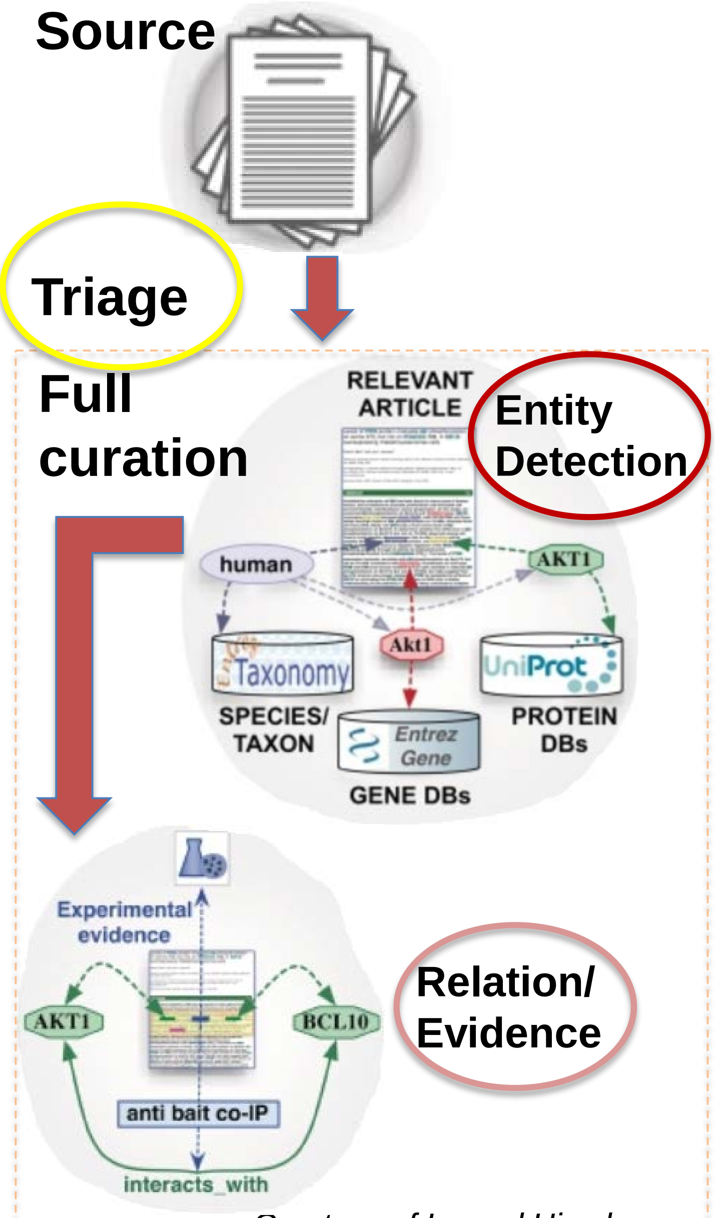
- To be useful for curation annotated results should be exported in standard formats that can be further utilized in the curation workflow. Tab-delimited and BioC formats were requested.

Metrics:

Time on task (objective)
Preference via survey (subjective)

Participating Systems in BioCreative IV

System	Description of the tool
Cell Finder	Annotation of gene, expression relation and cell type in text snippets from a set of articles
Ontogene	Detection of Gene/Chemical/Diseases and their interactions
MarkerRIF	Retrieval of articles about biomarkers, and extraction of disease and biomarker (gene) with normalization
SciKnowMine	Triage based on pre-trained categories of interest in full length articles
BioQRator	Retrieval based on relevance on protein-protein interaction information and annotation of protein pair
RLIMS-P	Triage on protein phosphorylation. Annotation of kinase, substrate and site with normalization.
Egas	Identification and extraction of protein-protein intearaction events described over PubMed abstracts related to neuropathological disorders
tagtog	Annotation of gene names within full-text documents especially machine-predicted documents
Argo	Annotation of metabolic process-related named entities, namely chemical entities and genes or gene products



Source

Triage

Full curation

Entity Detection

Relation/ Evidence

Courtesy of Lu and Hirschman

Recruitment

Recruitment of Biocurators

Call for participation via International Society for Biocuration (ISB) mailing list, and the ISB meeting and BioCreative websites

Personal invitation

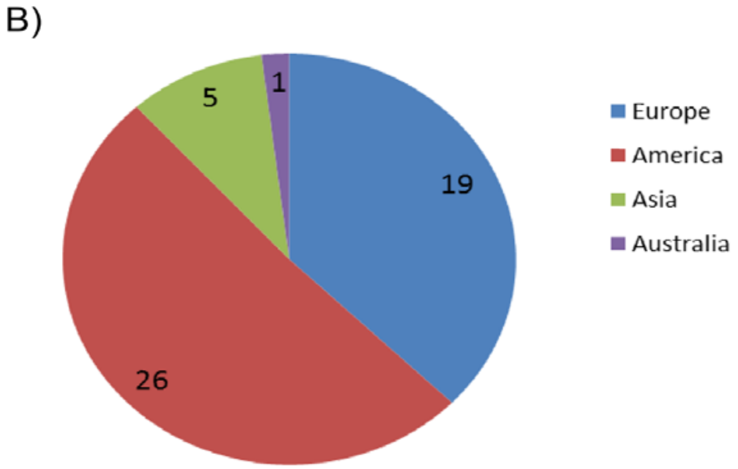
What's in it for Biocurators?

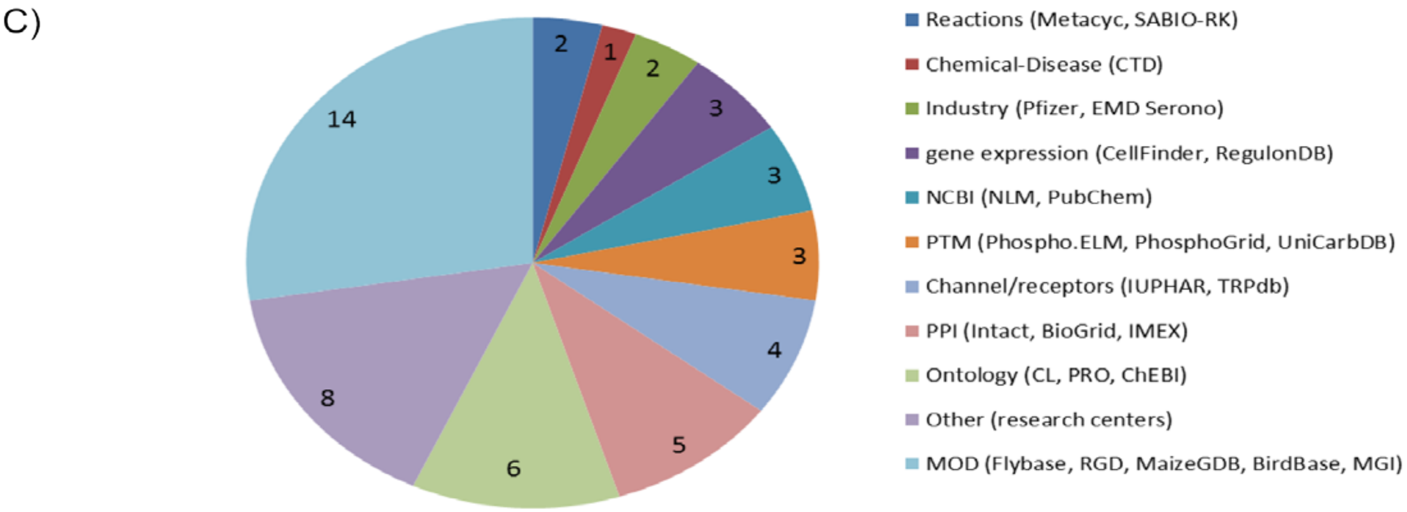
- Exposure to state-of-the-art text mining systems
- Direct communication and interaction with developers
- Contribution to tools that meet the needs of biocurators
- Adoption of text mining tool
- Potential publication in peer reviewed journal
- Focus on a set of articles that will add to their curation effort

Recruitment in BioCreative IV

A)

System	# Curators	
	Full	Partial
CellFinder	4	2
Ontogene	2	3
MarkerRIF	3	3
SciKnowMine	1	5
BioQRator	6	6
RLIMS-P	3	5
EGAS	4	4
Tagtog	7	2
Argo	3	1
Total	33	31

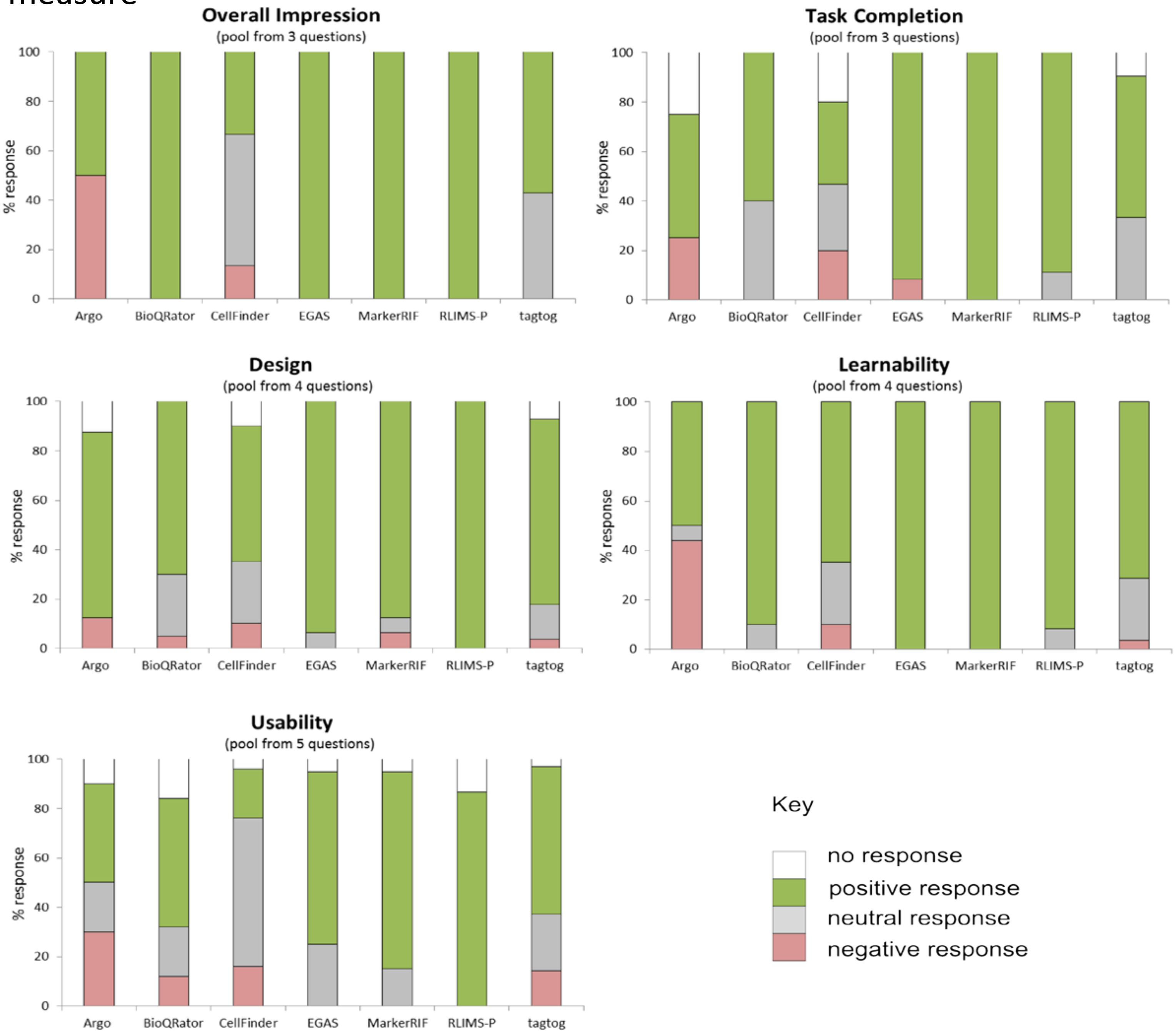




Time on task in full level participation curation task and curator experience level

System	no TM (min)	TM (min)	TM (min)	t_{noTM}/ t_{TM}	t_{noTM}/ t_{TM}	Curation experience (years)
BioQRator	275	195		1.4		<1
	70	100		0.7		>3
	160	180		0.9		>3
	150	150		1.0		1-3
Egas	93.71	60.13		1.6		<1
	184	120		1.5		1-3
	104.91	26.21		4.0		<1
	64.48	60.86		1.1		>3
MarkerRIF	212	90	145	2.4	1.5	1-3
	115	84	70	1.4	1.6	<1
	170	95	103	1.8	1.7	<1
RLIMS-P	585	560		1.0		>3
	301	186		1.6		1-3
	164	161		1.0		<1

Subjective measure



Overall Impression
(pool from 3 questions)

Task Completion
(pool from 3 questions)

Design
(pool from 4 questions)

Learnability
(pool from 4 questions)

Usability
(pool from 5 questions)

Some Observations:

The curation time does not always go hand by hand with user overall system satisfaction

With BioQRator and RLIMS-P curators are satisfied with system even the time required in the no-TM-assisted versus TM-assisted curation was comparable for each.

Some reasons:

-system provides a nice interface with functionalities that in the long run makes the monotonous curation work more enjoyable

-some systems have both retrieval and extraction steps, the curators appreciated the retrieval step because it saves a lot of time in article selection. However, the task was measured on the extraction step and most time was spent on normalization.

Areas of improvement detected via usability

System accessibility: due to one of the following; firewalls, system temporarily down, or inability to log in.

Error messages: either no error message displayed or the error message did not satisfactorily explain the problem.

Hidden functionality: key functionality for executing the TM task not apparent to curators.

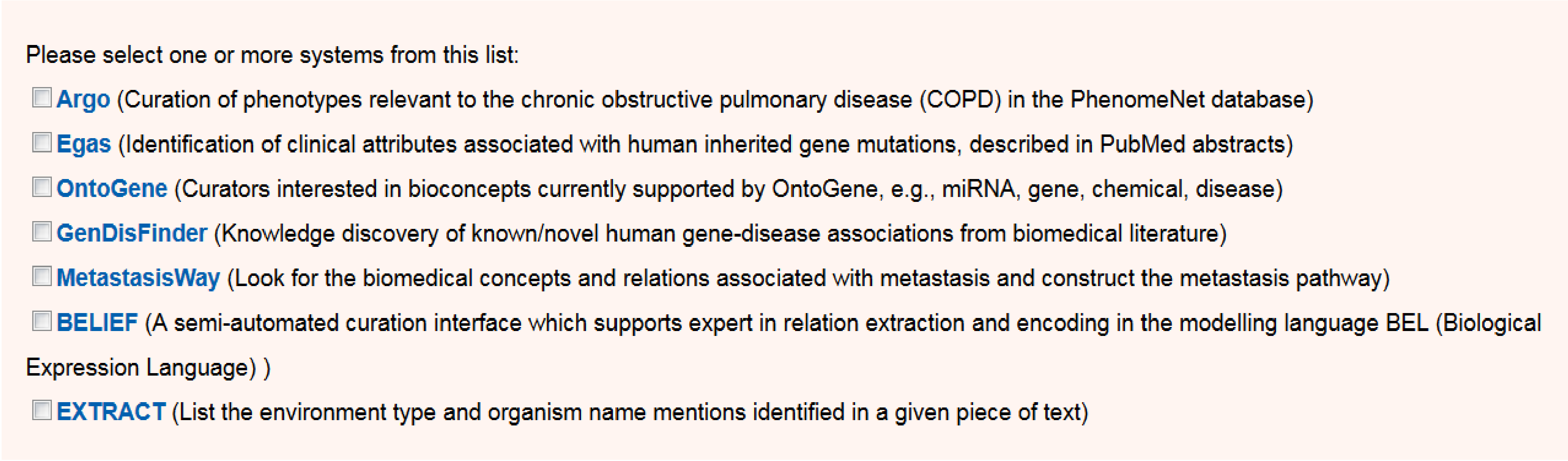
Language and icons: icons and names of sections/functionalities non-intuitive or used TM jargon.

Look and feel: Color choice for entity highlighting was not optimal for color blinded users for some of the systems

Recruiting users for IAT now!

http://www.biocreative.org/tasks/biocreative-v/iat-task-biocurators/

Seven Systems for different tasks



Please select one or more systems from this list:

☐ **Argo** (Curation of phenotypes relevant to the chronic obstructive pulmonary disease (COPD) in the PhenomeNet database)

☐ **Egas** (Identification of clinical attributes associated with human inherited gene mutations, described in PubMed abstracts)

☐ **OntoGene** (Curators interested in bioconcepts currently supported by OntoGene, e.g., miRNA, gene, chemical, disease)

☐ **GenDisFinder** (Knowledge discovery of known/novel human gene-disease associations from biomedical literature)

☐ **MetastasisWay** (Look for the biomedical concepts and relations associated with metastasis and construct the metastasis pathway)

☐ **BELIEF** (A semi-automated curation interface which supports expert in relation extraction and encoding in the modelling language BEL (Biological Expression Language))

☐ **EXTRACT** (List the environment type and organism name mentions identified in a given piece of text)

Evaluation Period: period June 22 to July 31
Flexible and remotely conducted
Total time commitment estimation over that period:
Full participation: 12h
Partial participation: 30min-1h

Sign In!

Concluding Remarks

Manual curation is accurate, but does not scale. Text mining scales, but is not accurate

Interactive systems like those presented in IAT can provide curators with decision support:

- suggesting important papers to curate
- highlighting entities of relevance in text
- offering controlled vocabularies and ontologies
- on-the-fly error-correction
- removal of redundancy

Concluding Remarks

- Tools developed in past BioCreative challenges have been integrated as modules in a subset of the participating systems, such as GenNorm in RLIMS-P; and PIE for protein-protein interaction article ranking and retrieval BioQRator.

- This demonstrates the importance of the traditional shared tasks to promote development of state-of-the-art text mining tasks that when mature these offer text mining solutions that can be integrated in a system framework.

Acknowledgements

BioCreative Organizers

UAG Members

Text mining teams and biocurators that participated in the BioCreative tasks

Funding
NIH/NIGMS R13-GM109648-01A1; DOE: DE-SC0010838; NIH Intramural Research Program, National Library of Medicine; IMI-eTOX, under Grant Agreement nr. 115002; Biomolecular and Bioinformatics Resources Platform of Institute of Health Carlos III Grant: PT13/0001/0001