



U.S. National Library of Medicine  
National Center for Biotechnology Information

# TeamTat

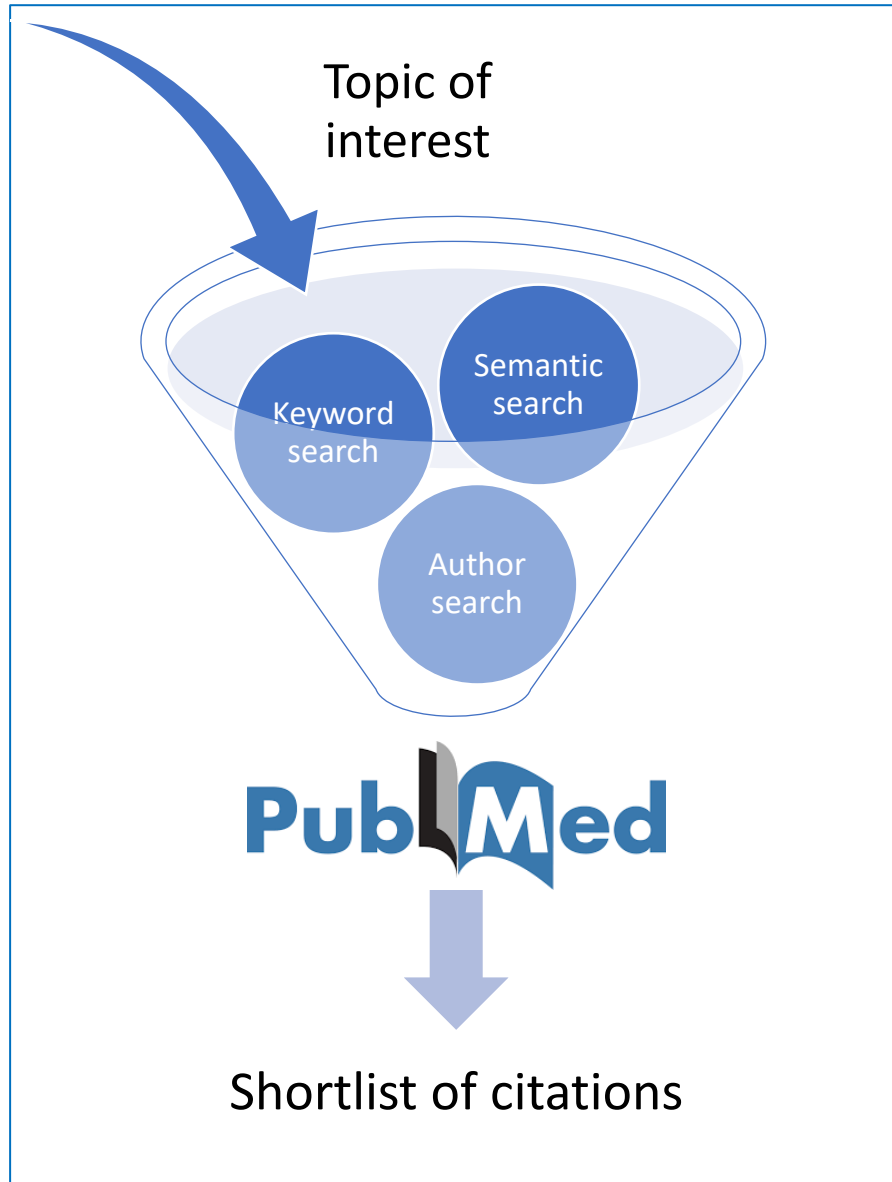
A collaborative text annotation tool

<https://www.teamtat.org/>

---

Rezarta Islamaj

# Literature Search



# PubMed query examples

- breast cancer
- stem cells
- t cell
- multiple sclerosis
- vitamin d
- lung cancer
- prostate cancer
- heart failure
- colorectal cancer
- rheumatoid arthritis
- atrial fibrillation
- back pain
- Alzheimer's
- gastric cancer
- Parkinson's

# BioNLP & Text Mining

PubTator<sub>Central</sub> 30826563 TUTORIAL API

group type sort freq

Search...

**GENE**  
ESR1 (16)

**DISEASE**  
BREAST CANCER (9)  
CANCER (2)  
PRIMARY DISEASE (2)  
METASTASIS (2)

**CHEMICAL**  
TAMOXIFEN (2)

**MUTATION**  
E380Q (1)  
p.(Y537C) (1)  
p.D538G (1)  
p.(L536R) (1)  
p.S463R (1)  
[more](#)

**SPECIES**  
PATIENTS (9)

**The prevalence of estrogen receptor-1 mutation in advanced breast cancer: The estrogen receptor one study (EROS1).**

PMID30826563  
NAJIM O, HUIZING M ... TJALMA W • CANCER TREAT RES COMMUN. 2019 • 2019

↓

BACKGROUND: Breast cancer has, due its high incidence, the highest mortality of cancer in women. The most common molecular variety of breast cancer is luminal subtype that expresses estrogen and progesterone receptors. Estrogen receptor alpha (ERalpha), encoded by the estrogen receptor1 (ESR1) gene, is expressed in approximately 70% of all breast cancers, and hormonal therapy represents a major treatment modality in all stages of ER positive breast cancers. Acquired mutations in the ligand-binding domain (LBD) of ERalpha, referred as ESR1 mutation, result in resistance to different endocrine therapies leading to disease progression or recurrence. Recent studies reviled that these ESR1 mutations lead to constitutive activity of the estrogen receptor ER, meaning that the receptor is active in absence of its ligand conferring resistance against endocrine therapy and tumor growth. Published studies have not yet been able to determine the exact prevalence rate of ESR1 mutations, but set the outer boundaries between 11-55%.  
PURPOSE: The goal of the present study is to determine the frequency rate of ESR1

BioConcepts  
 GENE  
 DISEASE  
 CHEMICAL  
 MUTATION  
 SPECIES  
 CELLLINE

## Extract Information from unstructured text

- Concepts
  - Diseases
  - Drugs
- Relations
  - Drug TREATS disease
  - Toxin CAUSES symptom
- Questions
  - Vaccine efficacy for Covid-19

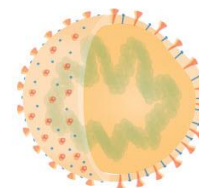
# Build automatic tools

Chemical entity recognition			Chemical entity normalization <sup>1</sup>		
Precision	Recall	F-measure	Precision	Recall	F-measure
<b>0.810</b>	<b>0.711</b>	<b>0.757</b>	<b>0.822</b>	<b>0.728</b>	<b>0.772</b>

GENE Entity Recognition			Gene Entity Normalization <sup>2</sup>		
Precision	Recall	F-measure	Precision	Recall	F-measure
<b>0.933</b>	<b>0.834</b>	<b>0.881</b>	<b>0.879</b>	<b>0.840</b>	<b>0.859</b>

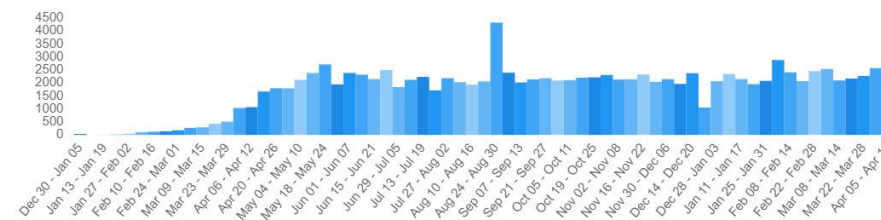
<sup>1</sup>Islamaj, R., Leaman, R., Kim, S. *et al.* NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature. *Sci Data* **8**, 91 (2021). <https://doi.org/10.1038/s41597-021-00875-1>

<sup>2</sup>Islamaj, R., Wei, C-H., Cissel, D., *et al.* NLM-Gene, a richly annotated gold standard dataset for gene entities that addresses ambiguity and multi-species gene recognition. *J Biomed Informatics*, (2021). <https://doi.org/10.1016/j.jbi.2021.103779>



LitCovid is a curated literature hub for tracking up-to-date scientific novel Coronavirus. It is the most comprehensive resource on the subject to [118421](#) (and [growing](#)) relevant articles in PubMed. The articles are categorized by different research topics and geographic locations for more at [Chen et al. Nature](#) (2020) or our [FAQ](#), and download our data [here](#)

Weekly Publications



Latest Publications

- PREVENTION**  
Modelling international pandemic in the  
McCabe, Ruth et al.
- PREVENTION**  
Literature-based organizing and the backdrop of  
Juraszek, Andrzej et
- PREVENTION**  
Knowledge, attitudes in public pandemic in the

Countries mentioned in abstracts



<https://www.ncbi.nlm.nih.gov/research/coronavirus/>

# Gold-standard datasets

- AI algorithms learn from classified data
- Human annotated and verified datasets (created by experts in the field) can be used to develop, train and test automatic prediction algorithms
- **However,**  
**Human annotation requires considerable time, effort and expertise**

# Why TeamTat?

---

Easy to use annotation interface

---

Easy to interact and review interface  
(team annotation)

---

Easy to direct an annotation project  
(project management)

---

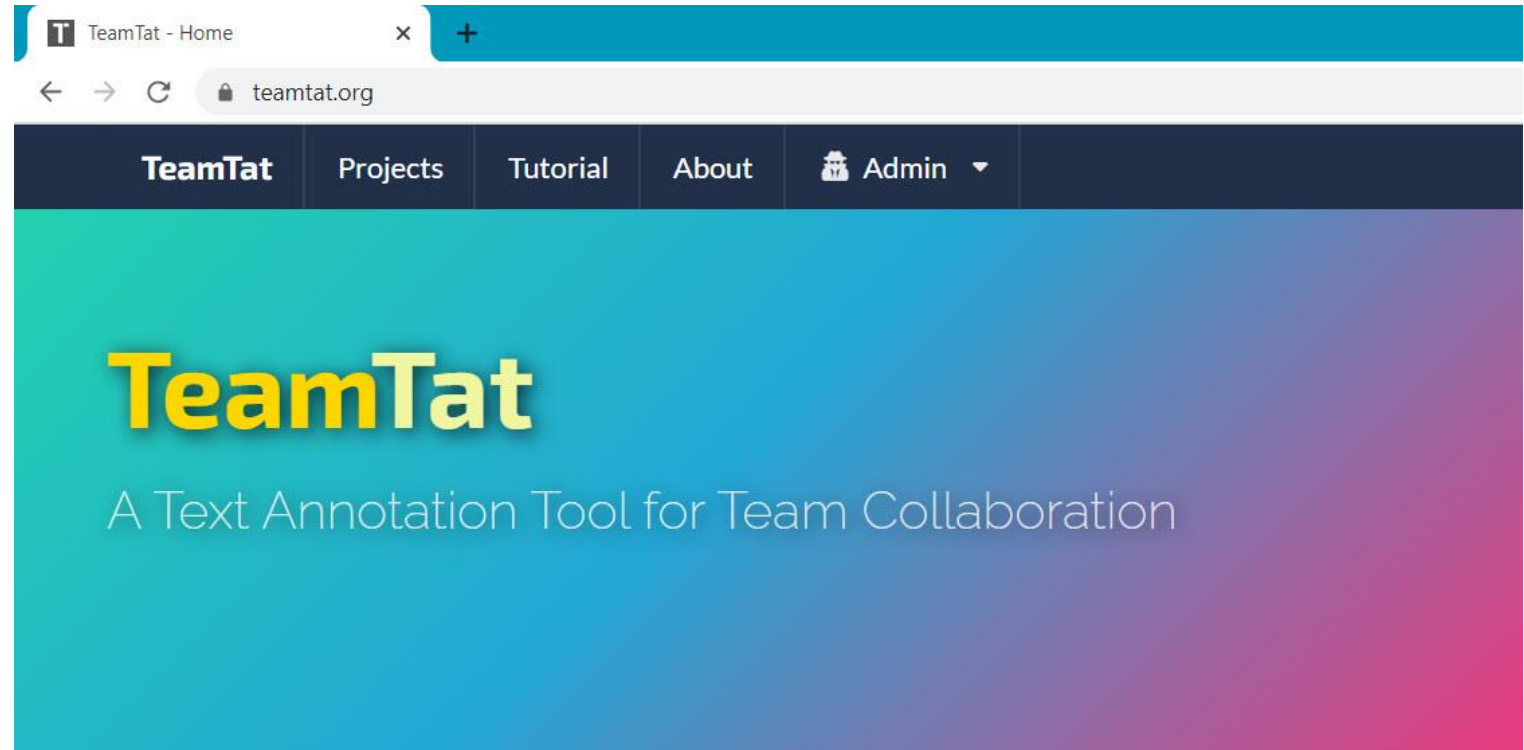
Easy to measure inter-annotator  
agreement (quality assessment)

---

You own your data

# Goals

- Tool:
  - Publicly available
  - Web-based, open-source (local installation for sensitive data)
- Data
  - Integration with PubMed/PMC
  - Unicode Support, full text support, image view
- Functionality:
  - Annotation
  - Team Project Management
  - Quality assessment



Rezarta Islamaj, Dongseop Kwon, Sun Kim, Zhiyong Lu, *TeamTat: a collaborative text annotation tool*, *Nucleic Acids Research*, Volume 48, Issue W1, 02 July 2020, Pages W5–W11, <https://doi.org/10.1093/nar/gkaa333>

# TeamTat Usage

- 1,685 annotator accounts
- 425 project managers
- 917 projects
- 32,425 documents
- 1,109,555 annotations
- 353 annotation rounds

- \*Languages: English, Portuguese, German
- \*Types of documents
  - Medical Literature (journal articles, abstracts, or full text)
  - Clinical documents and doctor notes
  - Non-medical articles
- University research teams, research labs, and other institutions

\*that we know because of teams that have reached out



# TeamTat: How to define annotation schema

The image shows two screenshots of the TeamTat web application interface. The top screenshot displays the 'PMC-article' project page with the 'Entity Types' tab selected. A message indicates that annotators need to be assigned before starting a round. A 'New Entity Type' form is open, with 'Gene' entered in the 'Name' field and 'GENE:' in the 'Prefix' field. The bottom screenshot shows the 'Relation Types' tab, which contains a table with one entry: 'GeneDisease' with 2 nodes. A 'Pick Color' button is visible next to the 'GeneDisease' entry. A color palette is also shown on the right side of the interface.

**Top Screenshot: Entity Types**

TeamTat Projects Tutorial d725614a05c5

Projects / PMC-article / Entity Types

PMC-article **Version #0** **Preparing**

Give PMCID and automatically retrieve article from PubMed Central Open Access

⚠ You need to assign annotators first before starting a round.

Download AI Tool Start Round

Documents (2) Members (1) Assignments (0) **Types** Lexicons (0) Tasks (0)

**Entity Types (0)** Relation Types (0)

No entity types. Create an entity type (i.e. concept) using the button below or during manual annotation.

**New Entity Type**

**New Entity Type**

Name: Gene

Prefix: GENE:

**Bottom Screenshot: Relation Types**

TeamTat Projects Tutorial d725614a05c5

Projects / PMC-article / Relation Types

PMC-article **Version #0** **Preparing**

Give PMCID and automatically retrieve article from PubMed Central Open Access

⚠ You need to assign annotators first before starting a round.

Download AI Tool Start Round

Documents (2) Members (1) Assignments (0) **Types** Lexicons (0) Tasks (0) Models (0) Statistics

**Entity Types** **Relation Types**

Name	Color	# nodes	Entity Type	
GeneDisease	<b>Pick Color</b>	2	Gene.Disease	<b>Edit</b> <b>Delete</b>

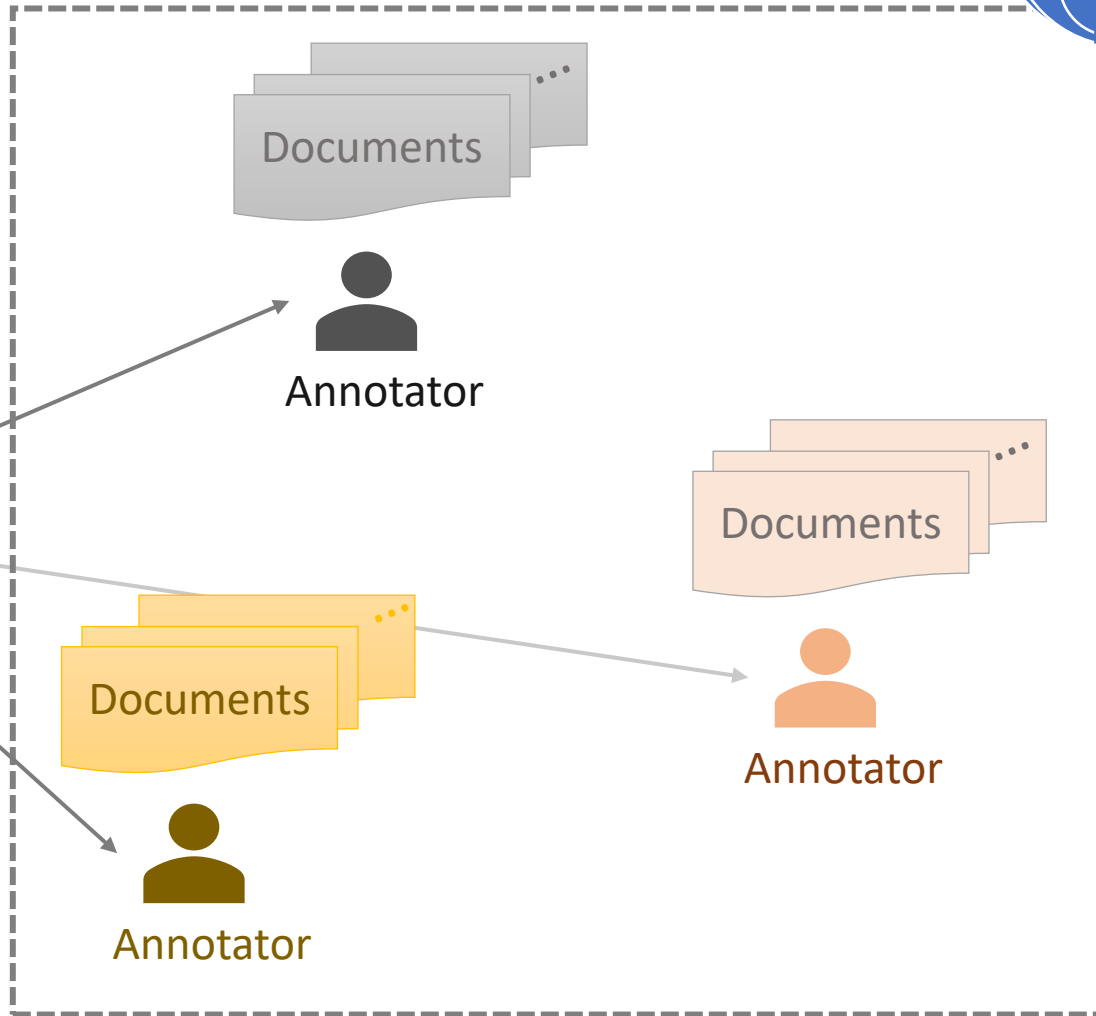
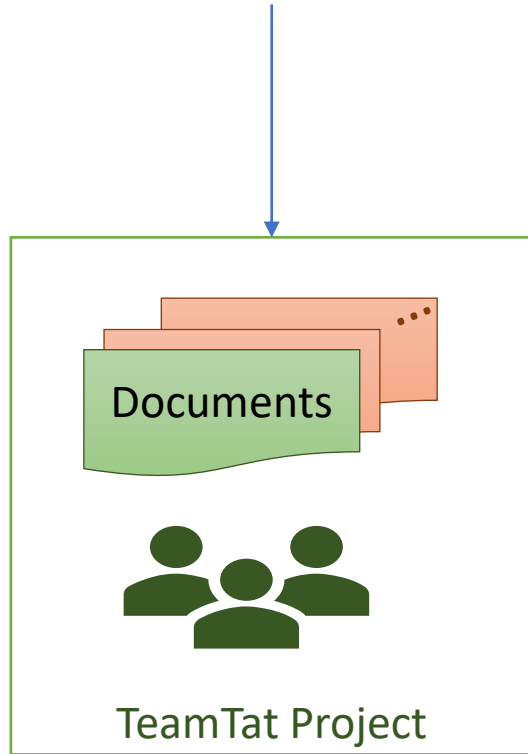
Color palette: #FFCCCC, #66CCFF, #99FF66, #FFCC00, #CCFF66, #FF66FF, #FF9999, #99CC00, #00CC99, #00CCFF



1. Setup a project
2. Upload documents
3. Select annotators
4. Distribute documents to annotators

Annotation Round

Relation annotation	<ul style="list-style-type: none"><li>• Sentence-level</li><li>• Paragraph-level</li><li>• Document level</li></ul>
Overlapping annotations	<ul style="list-style-type: none"><li>• Agreements</li><li>• Disagreements</li><li>• Multi-types</li></ul>
Concept annotation	<ul style="list-style-type: none"><li>• Full text</li><li>• Figures</li><li>• Tabular view</li></ul>



End of Round

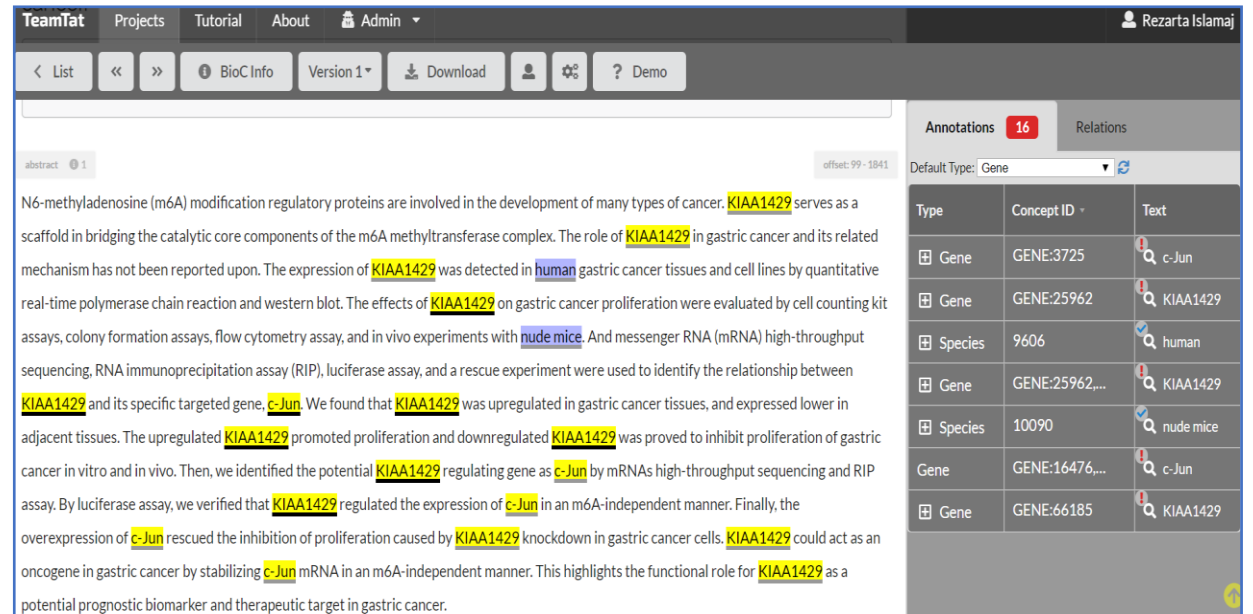
1. Collect annotations
2. Alert Project Manager
3. Review project status
4. Statistics

Add a new Round

Finalize Project

# TeamTat: Streamline annotation

- Easy user-interface
- Specify text boundaries
- Normalize to controlled Vocabularies
- Multiple entity types
- Multiple relation types
- Automatic annotation of repeat occurrences



The screenshot displays the TeamTat web application interface. The top navigation bar includes 'TeamTat', 'Projects', 'Tutorial', 'About', and 'Admin'. Below this is a secondary bar with 'List', 'BioC Info', 'Version 1', 'Download', and 'Demo' buttons. The main content area shows an abstract text with several entities highlighted in yellow: 'KIAA1429', 'human', 'nude mice', and 'c-Jun'. To the right of the text is a 'Relations' panel with a table of annotations.

Type	Concept ID	Text
Gene	GENE:3725	c-Jun
Gene	GENE:25962	KIAA1429
Species	9606	human
Gene	GENE:25962,...	KIAA1429
Species	10090	nude mice
Gene	GENE:16476,...	c-Jun
Gene	GENE:66185	KIAA1429



# TeamTat: Project Management

- Analyze each round of annotations
- Produce inter-annotator agreement statistics
- Visual clues alert to annotator disagreements
- Start a new annotation round
  - Individual
  - Collaborative
- Finalize a project
- Download data

# Conclusions

- WEB: <https://www.teamtat.org/>
- Source Code: <https://github.com/ncbi-nlp/TeamTat>

## TeamTat:

- Intuitive entity/relation annotation, adapting to different annotation guidelines
- Multi-role support (i.e., project manager, annotator)
- Improves annotation efficiency
- Individual and collaborative annotation
- Corpus quality assessment support

## Authors:

Rezarta Islamaj, Dongseop Kwon, Sun Kim, and Zhiyong Lu

biocreative.bioinformatics.udel.edu

Login | Register

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

News About Events Tasks Resources

## BioCreative VII

### BioCreative VII challenge and workshop (Events) [2020-01-22]

#### BioCreative VII Challenge and Workshop CFP

The workshop will take place sometime during the first two weeks of November 2021. The dates and modality are not set yet, however there will be a virtual component.

**BioCreative:** Critical Assessment of Information Extraction in Biology is a community-wide effort for evaluating text mining and information extraction systems applied to the biology domain. BioCreative has been an invaluable source for advancing state-of-the-art text mining methods by providing reference datasets and a collegial environment to develop and evaluate these methods in both shared and interactive modes. The sudden spread of COVID-19 has triggered an unexpected pressure on the biomedical community to quickly identify potential treatments by repurposing existing drugs or identifying new chemicals with anti-Sars-CoV-2 activity. Thus, BioCreative VII will focus around detection of chemicals, drug related substances with three tracks: Track 1 (DrugProt) focuses on the detection of interactions between chemicals/drugs/substances and genes/proteins in abstracts, Track 2: Chem track) focuses on detecting chemical names and their MeSH encoding in full-length articles and Track 3: Medications in Tweets focuses on extracting medication mentions from social media.

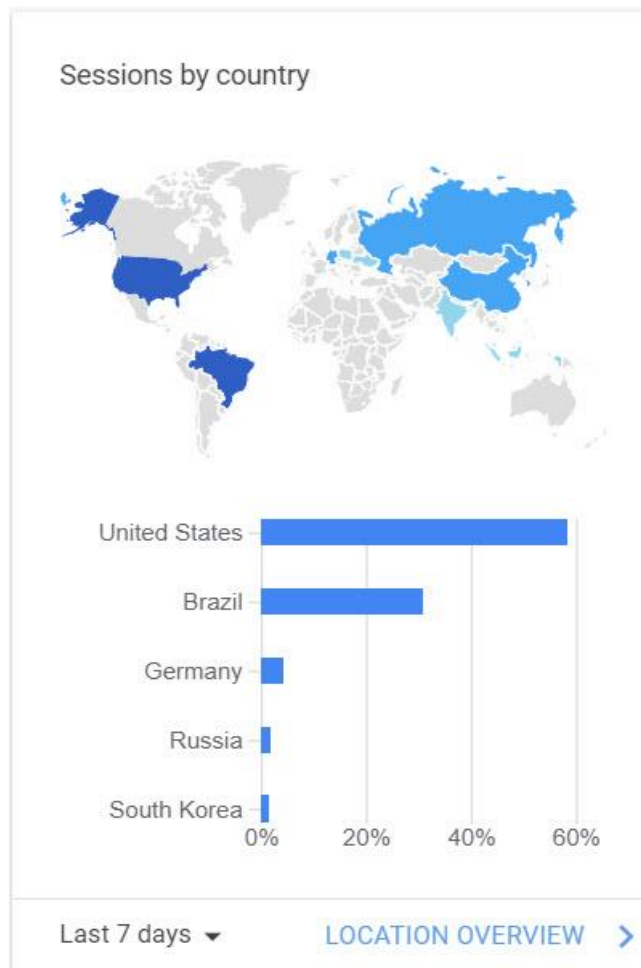
In addition, COVID-19 has triggered the development of multiple text mining tools to support ongoing research efforts that await community feedback. Thus, we are offering an interactive track, Track 4, to provide an environment for tools to be reviewed by users and get their feedback on utility and usability. We further offer Track 5, LitCovid Track on multi-label topic classification for COVID-19 literature annotation, calling for innovative text mining tools to support the curation of COVID-19 literature in LitCovid, a literature database of COVID-19-related papers in PubMed.

Here are more details about the tracks. Click on the Track number for accessing track specific pages:

- [Track 1- DrugProt: Text mining](#)

1. [DrugProt: Text mining drug/chemical-protein interactions](#)
2. [NLM-Chem Track: Full-text Chemical Identification PubMed](#)
3. [Automatic extraction of medication names in tweets](#)
4. [COVID-19 text mining tool interactive demo](#)
5. [LitCovid track: Multi-label topic classification for COVID-19 literature](#)

# TeamTat Users



		Acquisition
Country ?		Users ? ↓
		<b>2,365</b> % of Total: 100.00% (2,365)
1.	United States	<b>804</b> (33.91%)
2.	Germany	<b>511</b> (21.55%)
3.	Brazil	<b>340</b> (14.34%)
4.	China	<b>145</b> (6.12%)
5.	India	<b>73</b> (3.08%)
6.	South Korea	<b>58</b> (2.45%)
7.	United Kingdom	<b>43</b> (1.81%)
8.	Japan	<b>41</b> (1.73%)
9.	France	<b>37</b> (1.56%)
10.	Spain	<b>34</b> (1.43%)

