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

TeamTat

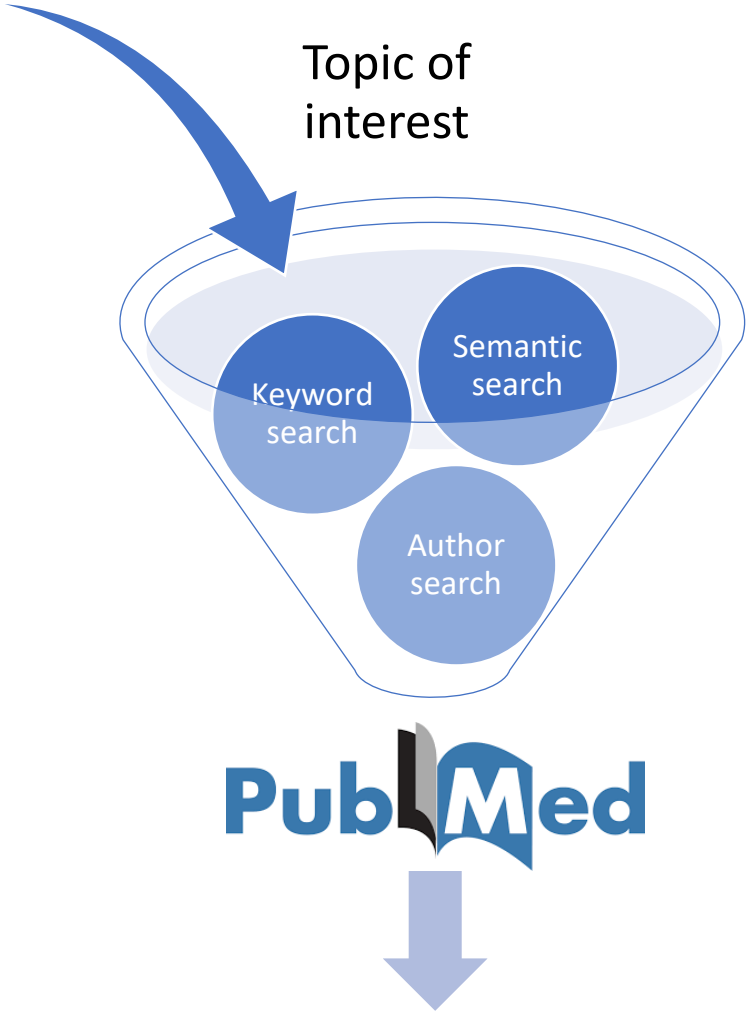
A collaborative text annotation tool

https://www.teamtat.org/

Rezarta Islamaj

MBE Summit/ Technical Language Processing Community of Interest, April 13, 2021

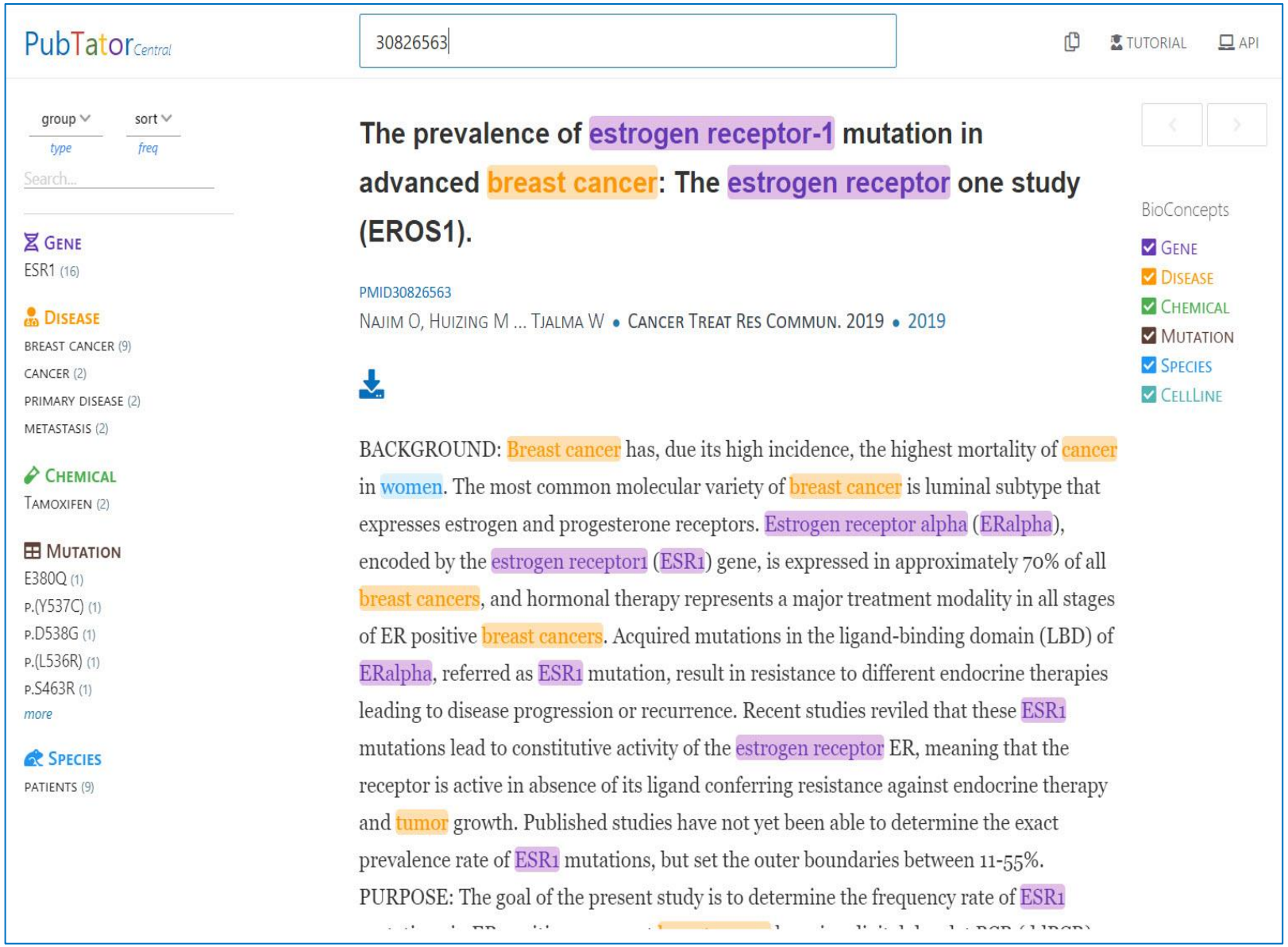
Literature Search



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



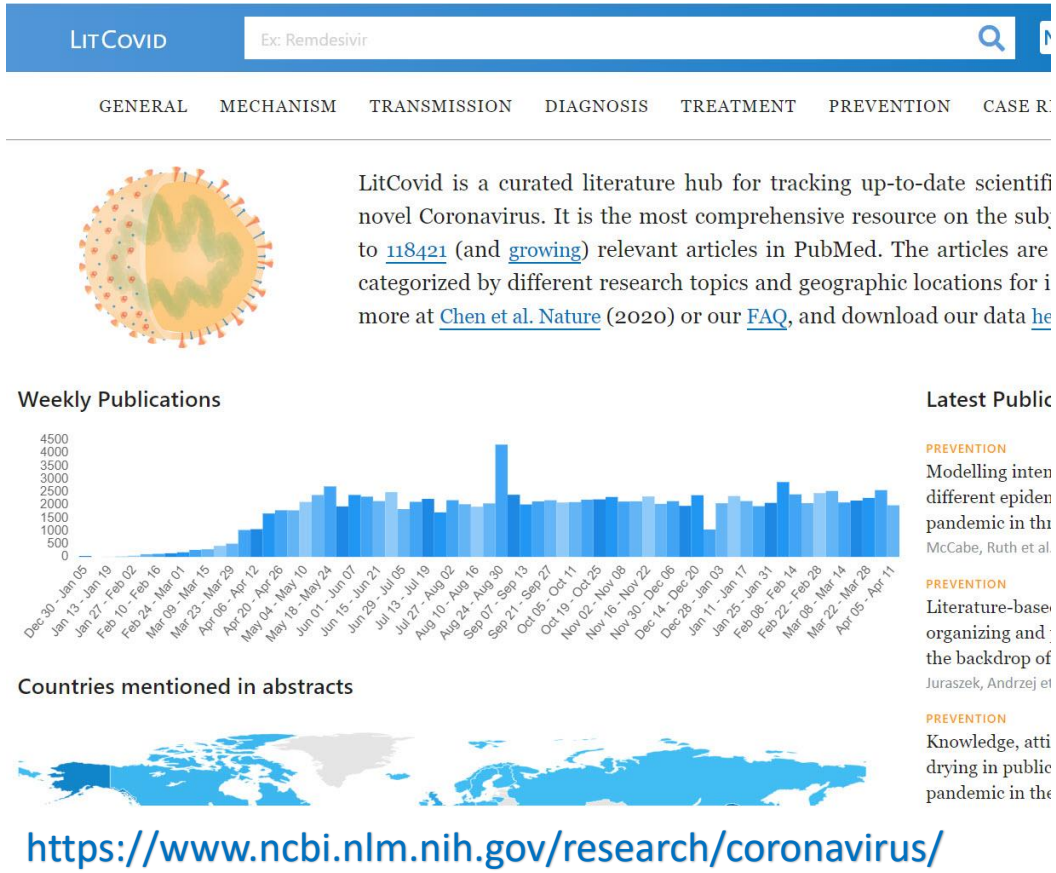
Extract Information from unstructured text

- 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[1]		
Precision	Recall	F-measure	Precision	Recall	F-measure
0.810	**0.711**	**0.757**	**0.822**	**0.728**	**0.772**

GENE Entity Recognition			Gene Entity Normalization[2]		
Precision	Recall	F-measure	Precision	Recall	F-measure
0.933	**0.834**	**0.881**	**0.879**	**0.840**	**0.859**



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

[1]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

[2]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

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)

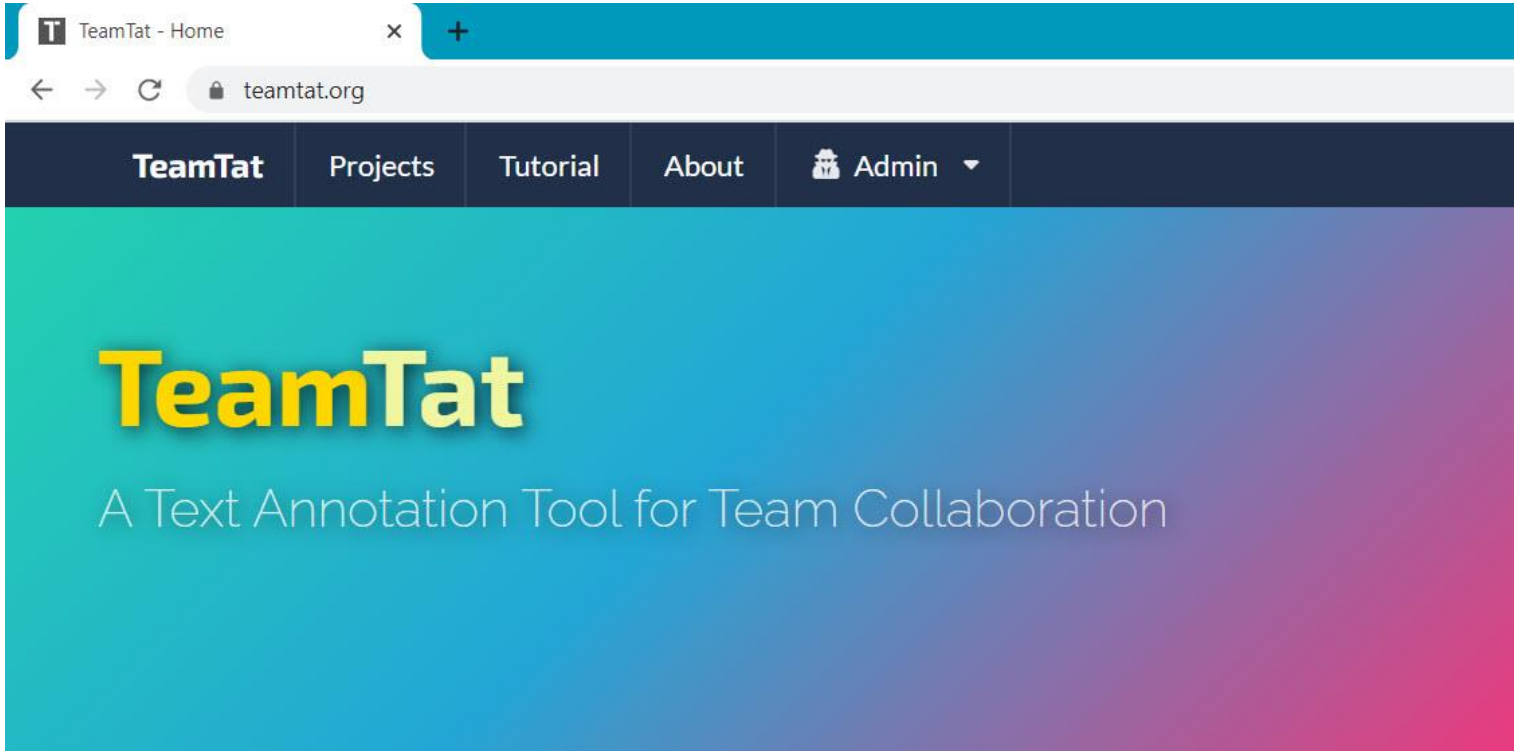
Easy to direct an annotation project (project management)

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

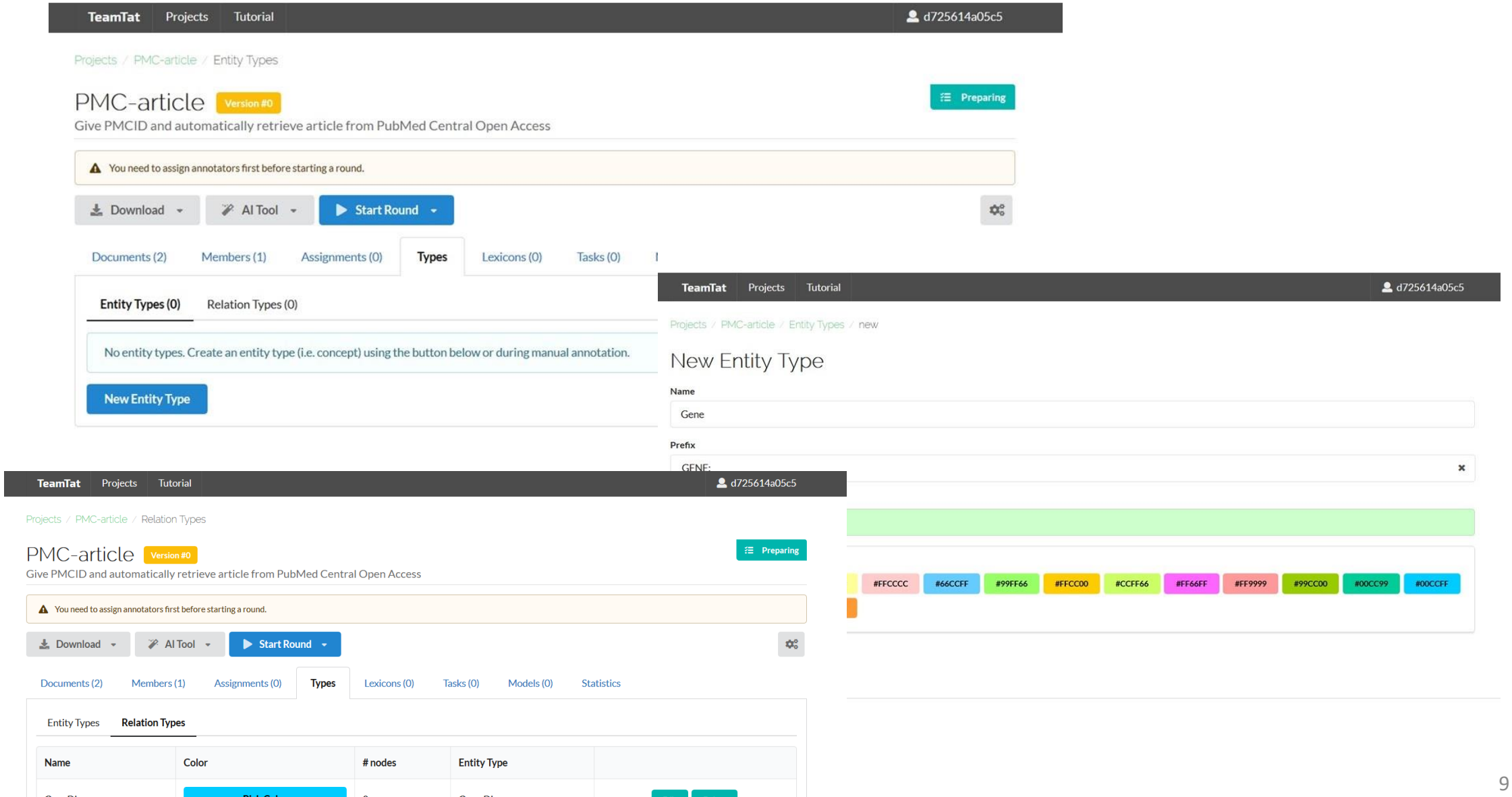
TeamTat Usage

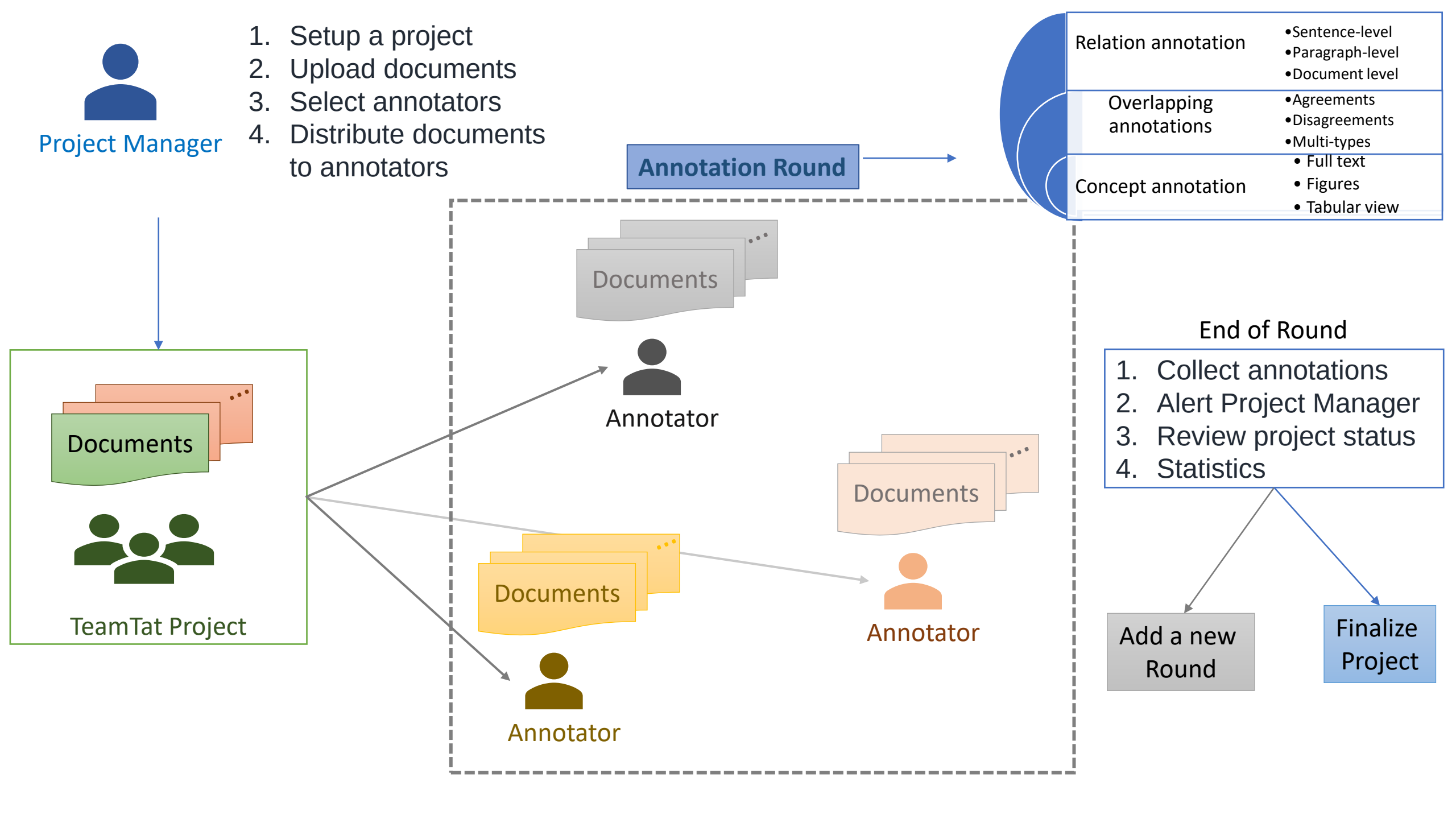
- 1,685 annotator accounts
- 425 project managers
- 917 projects
- 32,425 documents
- 1,109,555 annotations
- 353 annotation rounds

- *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

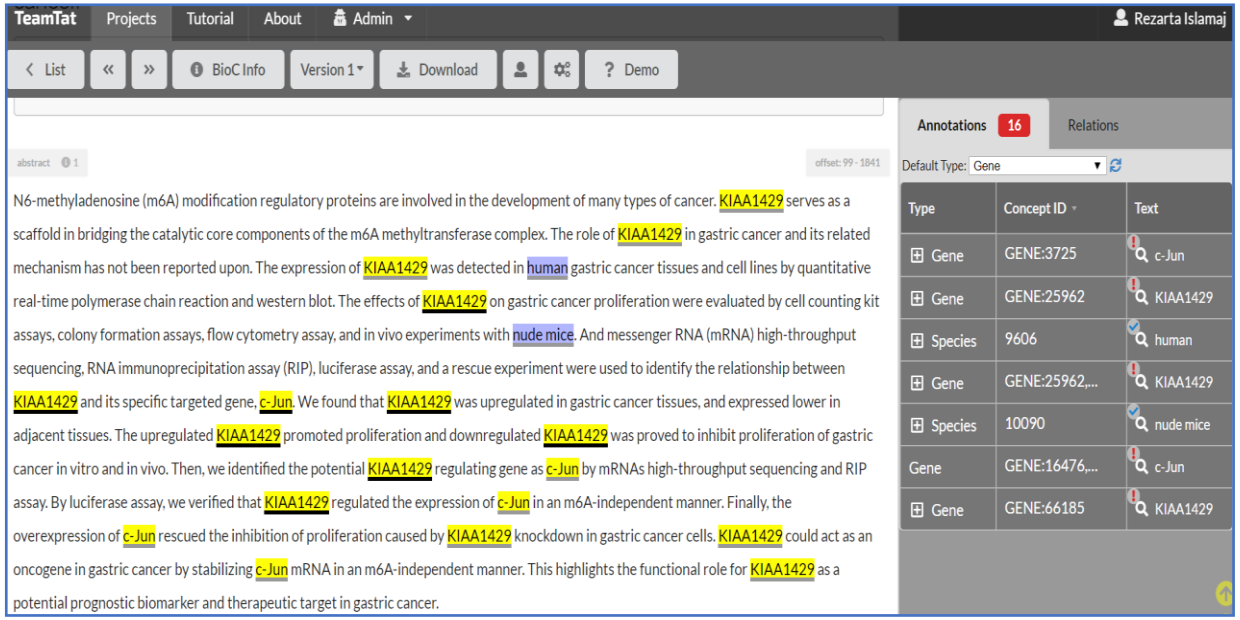
TeamTat: How to define annotation schema




Project Manager
1. Setup a project
2. Upload documents
3. Select annotators
4. Distribute documents to annotators
Annotation Round
Relation annotation
•Sentence-level
•Paragraph-level
•Document level
Overlapping annotations
•Agreements
•Disagreements
•Multi-types
Concept annotation
• Full text
• Figures
• Tabular view
Documents
TeamTat Project
Documents
Annotator
Documents
Annotator
Documents
Annotator
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

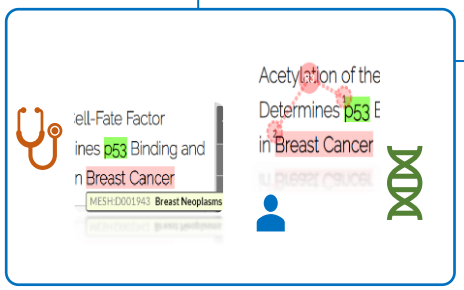
- Easy user-interface
- Specify text boundaries
- Normalize to controlled Vocabularies
- Multiple entity types
- Multiple relation types
- Automatic annotation of repeat occurrences



TeamTat: Annotation

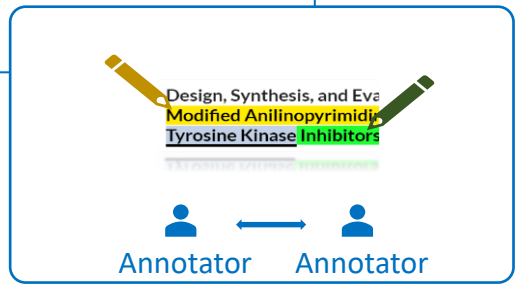
Independent Annotation Round

- Individual Workspace
- User-friendly interface
- Link to Ontologies
- Personalizable workspace
- Work at your own pace
- Visual clues to see dis-agreements
- Annotation partners identities are hidden
- All annotators' entries are saved and used for reference



Collaborative Annotation Round

- Collaborative Workspace
- User-friendly interface
- Link to Ontologies
- Personalizable workspace
- Work together.
- Assumes real-time discussion and disagreement resolution
- Visual clues to see dis-agreements
- Annotation partners identities are open
- Annotators' entries are final



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

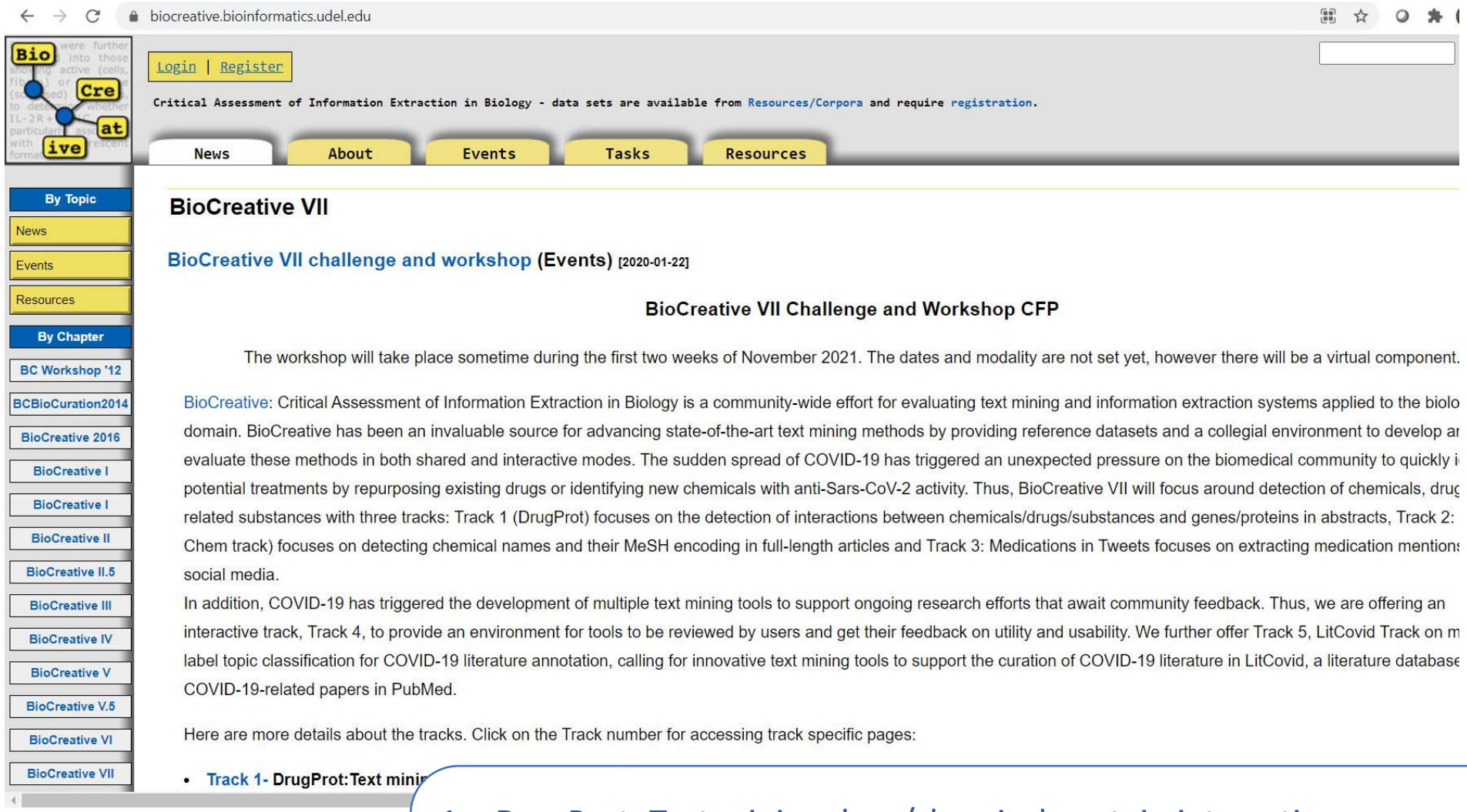
TeamTat:

- 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:

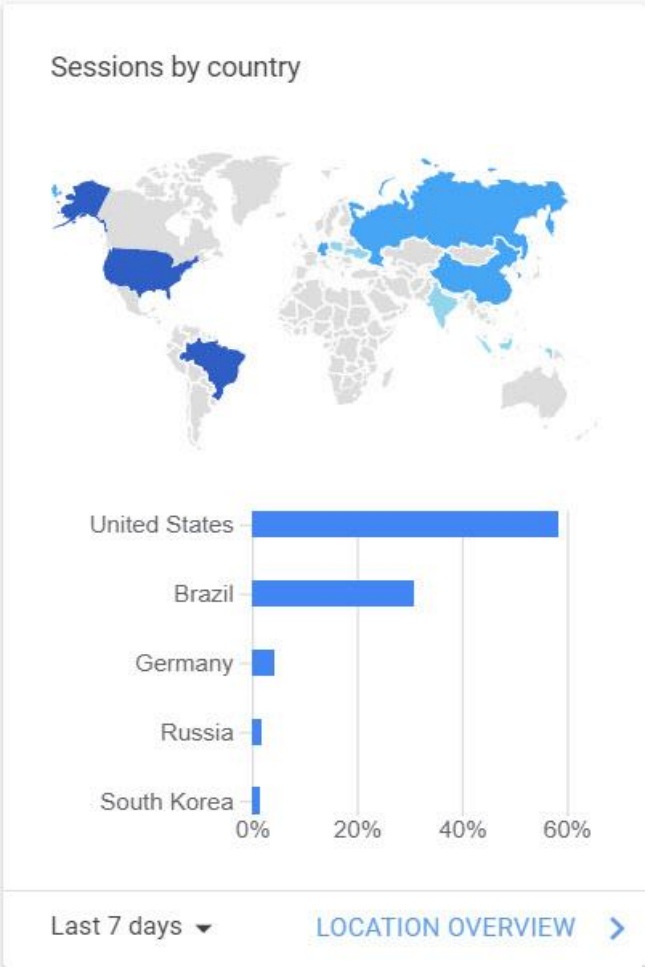
Rezarta Islamaj, Dongseop Kwon, Sun Kim, and Zhiyong Lu

This research was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health.



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



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