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

PubTator

Automated concept annotation for biomedical full text articles

Chih-Hsuan Wei* Alexis Allot* Robert Leaman Zhiyong Lu (PI)

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

INTRODUCTION

Biomedical Literature Textmining

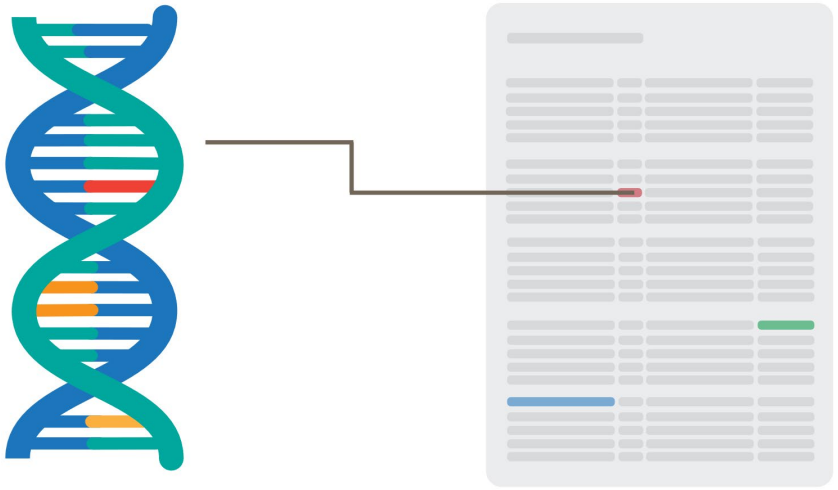
PROBLEM:

- Millions of biomedical publications in PubMed.
- Knowledge present in **unstructured** form as natural language.

Can not be easily :

- Processed by computers
- Integrated into curated databases

Biomedical Literature Textmining



SOLUTION:

Automated **text mining** allows to easily access and **extract knowledge within the biomedical literature**.

1. For downstream text mining applications
   - gene prioritization
   - genetic disease analysis
   - literature-based knowledge discovery

2. For faster biocuration

   Ex: curating a database (such as UniProt)

INTRODUCTION

Text Mining Tools

Our team has developed many dedicated tools.

GNormPlus

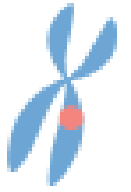
Chih-Hsuan Wei, Hung-Yu Kao, Zhiyong Lu. (2015) **GNormPlus: An integrative approach for tagging genes, gene families, and protein domains**. Biomed Res Int.

SR4GN

Chih-Hsuan Wei, Hung-Yu Kao, Zhiyong Lu. (2012) **SR4GN: a species recognition software tool for gene normalization.** PLoS One.

tmVar 2.0

Chih-Hsuan Wei, Lon Phan, Juliana Feltz, Rama Maiti, Tim Hefferon, Zhiyong Lu. (2017) **tmVar 2.0: integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine.** Bioinformatics.

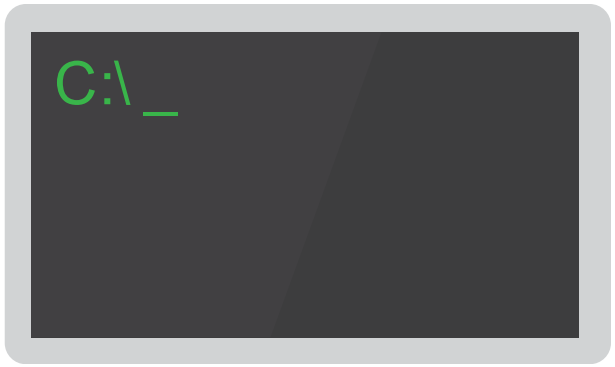


TaggerOne

Robert Leaman, Zhiyong Lu. (2016) **TaggerOne: joint named entity recognition and normalization with semi-Markov Model.** Bioinformatics.

INTRODUCTION

Biomedical Literature Textmining



PROBLEM: Text-mining command line tools require bioinformatics expertise to execute them.

SOLUTION: Web-based tools can simplify distributing results from text mining systems to a wide range of users:

- **No installation or maintenance**
- **No infrastructure requirement**

WEBSITE

What is PubTator ?

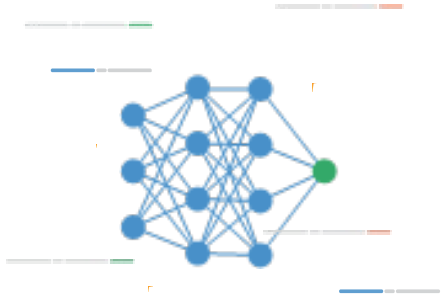
PubTator is a Web-based system providing **automatic annotations of biomedical concepts** such as genes and mutations in PubMed abstracts and PMC full-text articles.

Full-Text Articles

PubTator includes the full-text articles in the PMC Open Access subset (nearly 3 million) in addition to the 30+ million abstracts in PubMed.

Deep Learning

Cutting-edge machine learning and deep learning techniques are applied to concept disambiguation for improved accuracy.



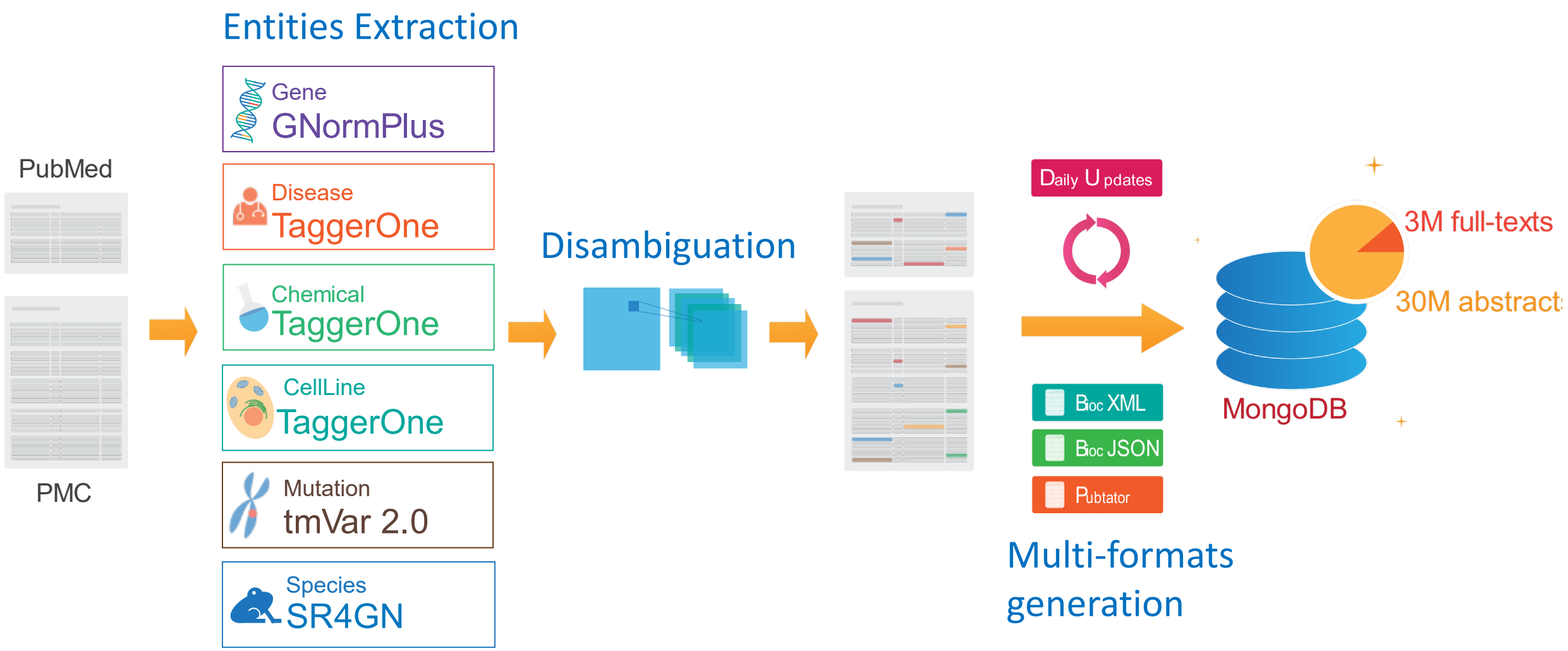
Always Up-to-date

PubTator adds new articles every day to always keep in sync with PubMed and PMC.

PubTator supports six concept types: **genes/proteins**, **genetic variants**, **diseases**, **chemicals** and **cell lines**

PREPROCESSING

Preprocessing Pipeline



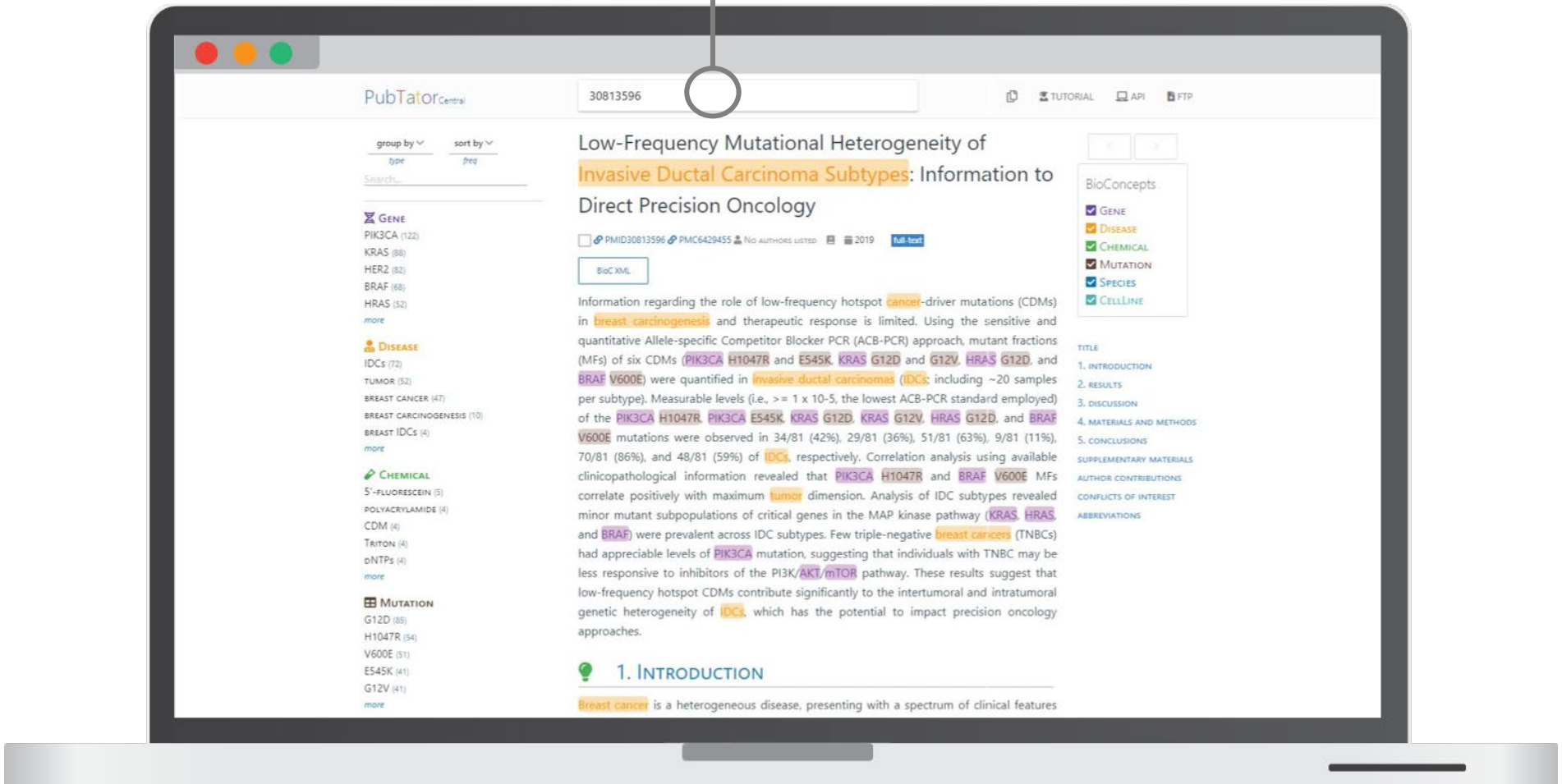
WEBSITE

Features

Semantic Search

Leverages PubTator annotations to find all publications mentioning an entity, regardless of which entity name the author uses

ESR1 vs **estrogen receptor**

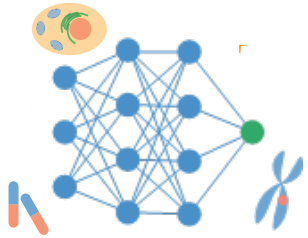


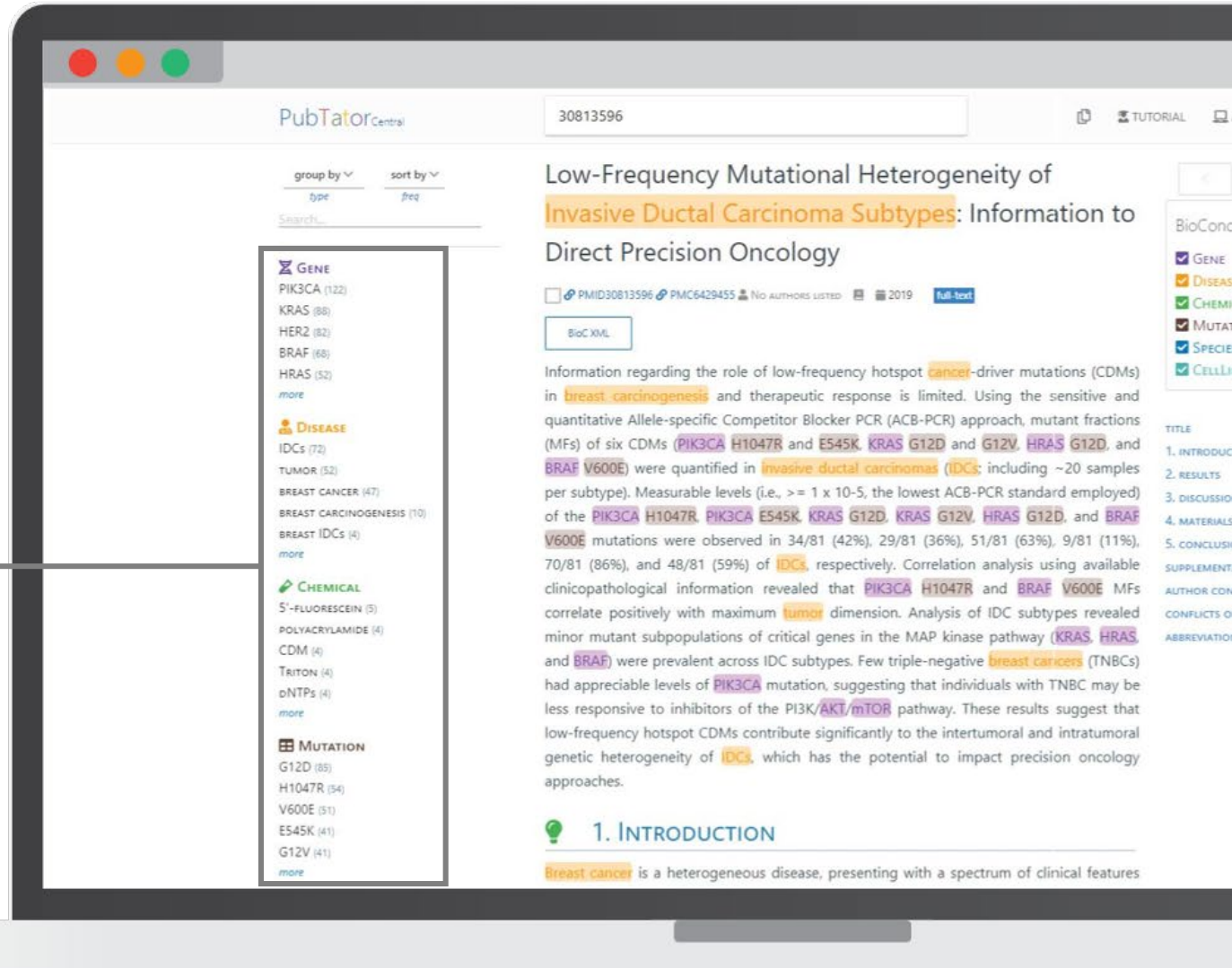
WEBSITE

Features

In-Document Search

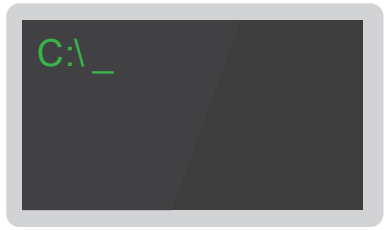
A menu displays a list of bioentities in a publication, allowing users to easily **navigate** to entities of interest.





WEBSITE

Features



Free Access

PubTator Central data is **free** and can be accessed :

- **interactively** through a web browser
- **programmatically** via **RESTful API**
- **downloaded** in bulk via **FTP**

WEBSITE

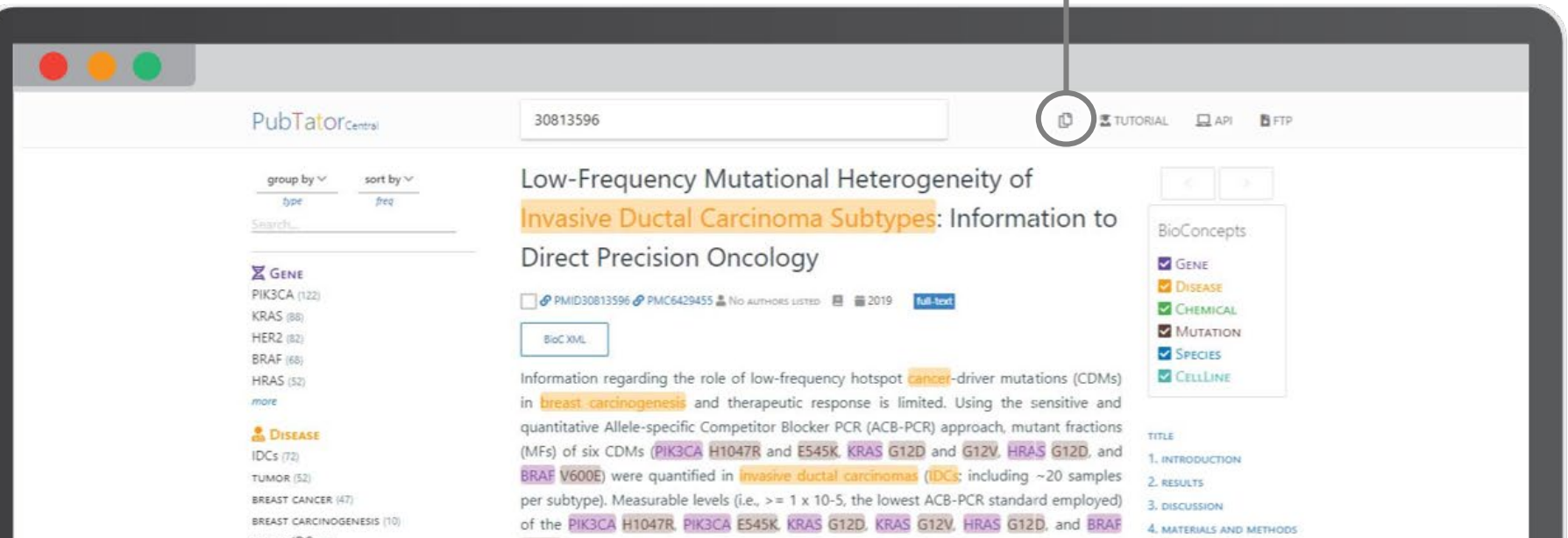
Features

Collections

Articles can be **organized** into collections, and then be **viewed** or **downloaded** together.

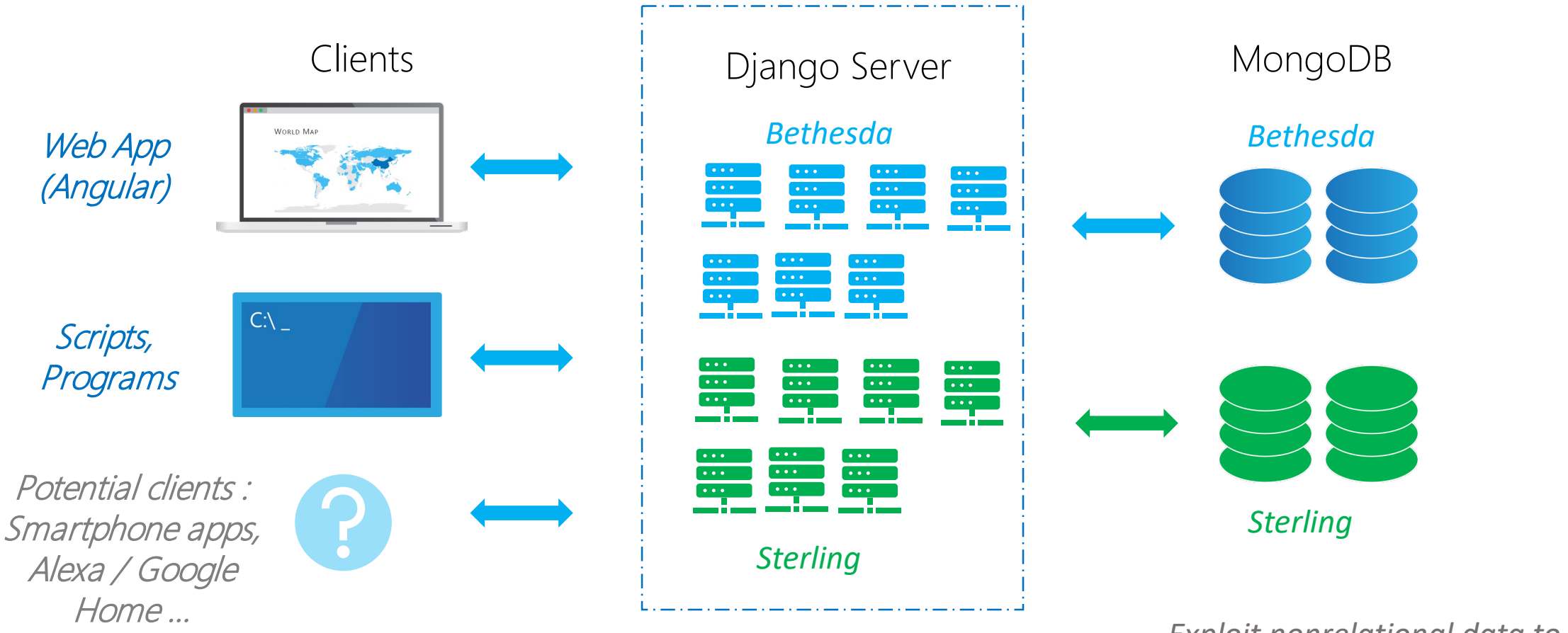
Articles may be added by:

- **Selecting** specific publications
- Entering a **query**
- Entering a list of **PMIDs**



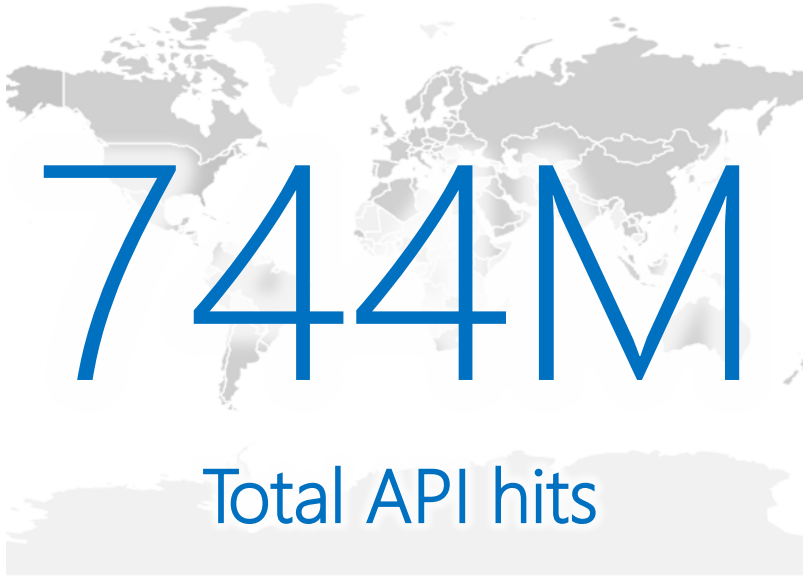
WEBSITE

(Future proof) Architecture



USAGE

Usage



USE CASE 1 : PRIORITIZING PROTEINS ASSOCIATED WITH GENETIC MUTATIONS IN CANCERS

"The phrase "mutation cancer" was used as the search term to retrieve the commonly mutated genes tagged by PubTator. The proteins co-published with each of the identified genetic mutations were retrieved, respectively."

Yu KH, Lee TM, Wang CS, et al. **Systematic Protein Prioritization for Targeted Proteomics Studies through Literature Mining**. J Proteome Res. 2018;

USE CASE 2 : UNIPROT DATABASE CURATION

"With the assistance of the PubTator text-mining tool, we tagged more than 10000 articles to assess the ratio of papers relevant for curation."

Poux S, Arighi CN, Magrane M, et al. **On expert curation and scalability: UniProtKB/Swiss-Prot as a case study**. Bioinformatics. 2017;

Conclusions

PubTator is a web-based system for automated concept annotations in PubMed abstracts and PMC-TM full text articles

FTP (Updated monthly)
ftp://ftp.ncbi.nlm.nih.gov/pub/lu/PubTatorCentral/

API (Updated daily)
https://www.ncbi.nlm.nih.gov/research/pubtator-api/publications/export/biocxml?full=true&pmids=30375428

Raw Text annotation service
https://www.ncbi.nlm.nih.gov/research/pubtator/api.html

Sample client codes in multiple languages : python, java and perl
https://www.ncbi.nlm.nih.gov/research/pubtator/api.html

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

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