Chan Zuckerberg Initiative

Essential frontiers: open data & software citations, an automated ML approach

Workshop on Open Citations And Openly Scholarly Metadata
October 5, 2022

Jennifer Lin, *Product Director, Indeed*
Ana-Maria Istrate, *Senior Research Scientist, Chan Zuckerberg Initiative*

Jennifer Lin
Indeed
Product Director

Ana-Maria Istrate
Chan Zuckerberg Initiative
Senior Research Scientist

Why extend the bibliographic record?

Software

Anders S (2010) **Babraham bioinformatics-fastqc a quality control tool for high throughput sequence data** Babraham Bioinformatics.

Attwood KM, Robichaud A, Westhaver LP, Castle EL, Brandman DM, Balgi AD, Roberge M, Colp P, Croul S, Kim I, McCormick C, Corcoran JA, Weeks A (2020) **Raloxifene prevents stress granule dissolution, impairs translational control and promotes cell death during hypoxia in glioblastoma cells** *Cell Death & Disease* **11**:989.
https://doi.org/10.1038/s41419-020-03159-5 | PubMed | Google Scholar

Bregman A, Avraham-Kelbert M, Barkai O, Duek L, Guterman A, Choder M (2011) **Promoter elements regulate cytoplasmic mRNA decay** *Cell* **147**:1473–1483.
https://doi.org/10.1016/j.cell.2011.12.005 | PubMed | Google Scholar

Book

Broad Institute (2009) **Picard Tools** By Broad Institute. Github.
Google Scholar

Choder M (2011) **Mrna imprinting** *Cellular Logistics* **1**:37–40.
https://doi.org/10.4161/cl.1.1.14465 | Google Scholar

Chowdhary S, Kainth AS, Gross DS (2017) **Heat shock protein genes undergo dynamic alteration in their three-dimensional structure and genome organization in response to thermal stress** *Molecular and Cellular Biology* **37**:1–22.

eLife journal article
https://doi.org/10.7554/eLife.76965

Data Citation Principles (2014):
https://doi.org/10.25490/a97f-egyk

Joint Declaration of Data Citation Principles

On this page:

Cite as: *Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014*
https://doi.org/10.25490/a97f-egyk

Translations

Japanese – https://doi.org/10.11502/rduf_rdc_jddcp_ja (added 31.01.2020).

>>> Endorsement List

Preamble

Sound, reproducible scholarship rests upon a foundation of robust, accessible data. For this to be so in practice as well as theory, data must be accorded due importance in the practice of scholarship and in the enduring scholarly record. In other words, data should be considered legitimate, citable products of research. Data citation, like the citation of other evidence and sources, is good research practice and is part of the scholarly ecosystem supporting data reuse.

In support of this assertion, and to encourage good practice, we offer a set of guiding principles for data within scholarly literature, another dataset, or any other research object.

These principles are the synthesis of work by a number of groups. As we move into the next phase, we welcome your participation and endorsement of these principles.

DC1 Data Citation Principles

Software Citation Principles (2016):
https://doi.org/10.7717/peerj-cs.86

PeerJ Computer Science

View 294 tweets

Related research

Share

‹ *PeerJ Computer Science*

Software citation principles

Research article | Digital Libraries | Software Engineering

Arfon M. Smith[1], Daniel S. Katz[2], Kyle E. Niemeyer[3], FORCE11 Software Citation Working Group

Tweet Authors

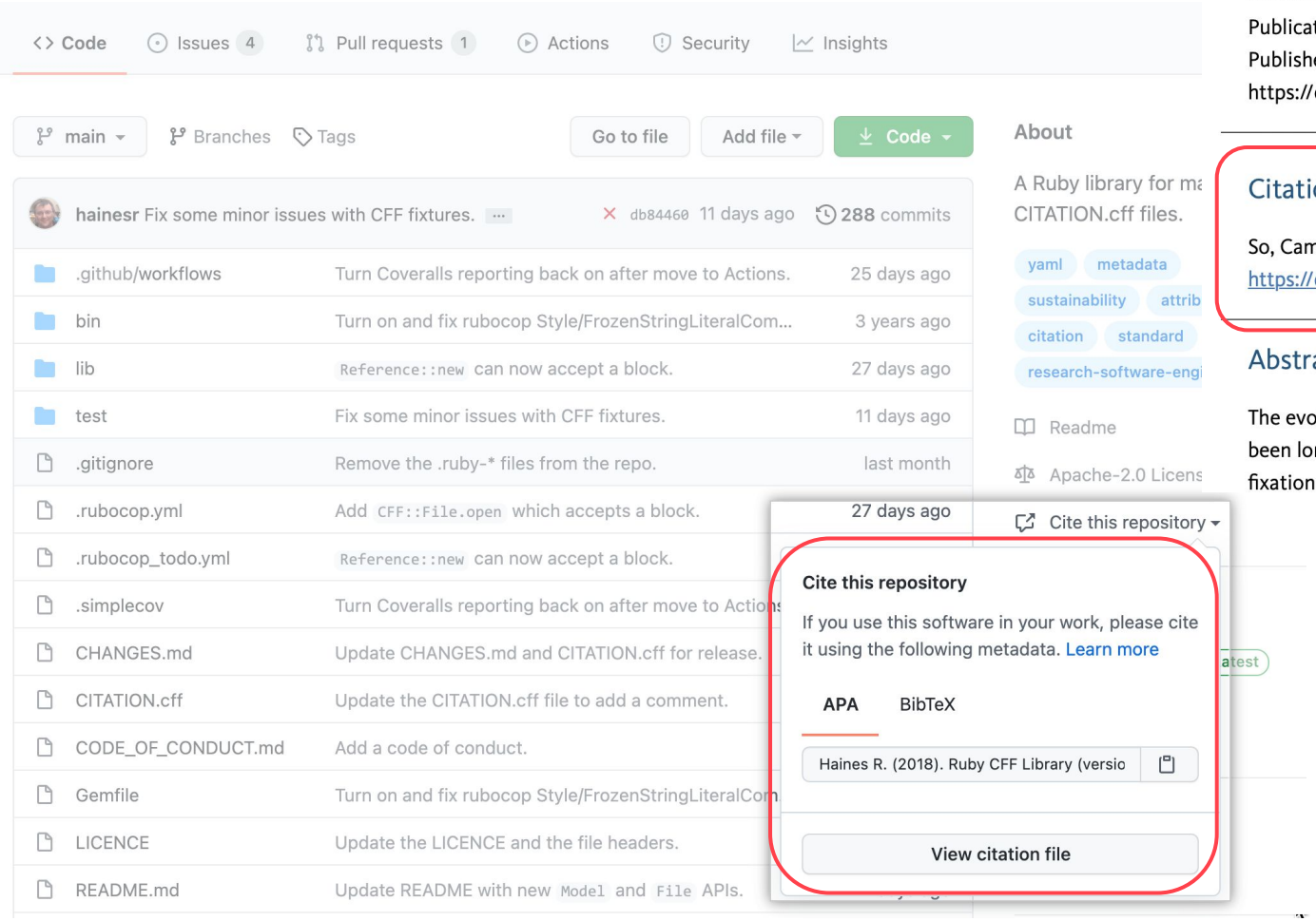
September 19, 2016

Note that a Preprint of this article also exists, first published June 27, 2016.

Author and article information

Abstract

Software is a critical part of modern research and yet there is little support across the scholarly ecosystem for its acknowledgement and citation



DRYAD

Genetic variance in contrasting environments

So, Cameron, University of Toronto, https://orcid.org/0000-0002-0663-195X
Sibolibane, Mia, University of Toronto
Weis, Arthur, University of Toronto
cameron.so@mail.mcgill.ca, micha.sibolibane@gmail.com, arthur.weis@utoronto.ca

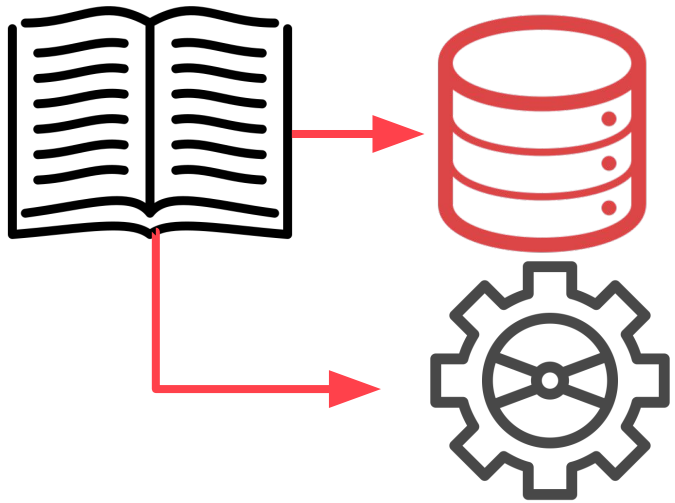
Publication date: October 4, 2023
Publisher: Dryad
https://doi.org/10.5061/dryad.dz08kprzx

Citation

So, Cameron; Sibolibane, Mia; Weis, Arthur (2023), Genetic variance in contrasting environments, Dryad, Dataset, https://doi.org/10.5061/dryad.dz08kprzx

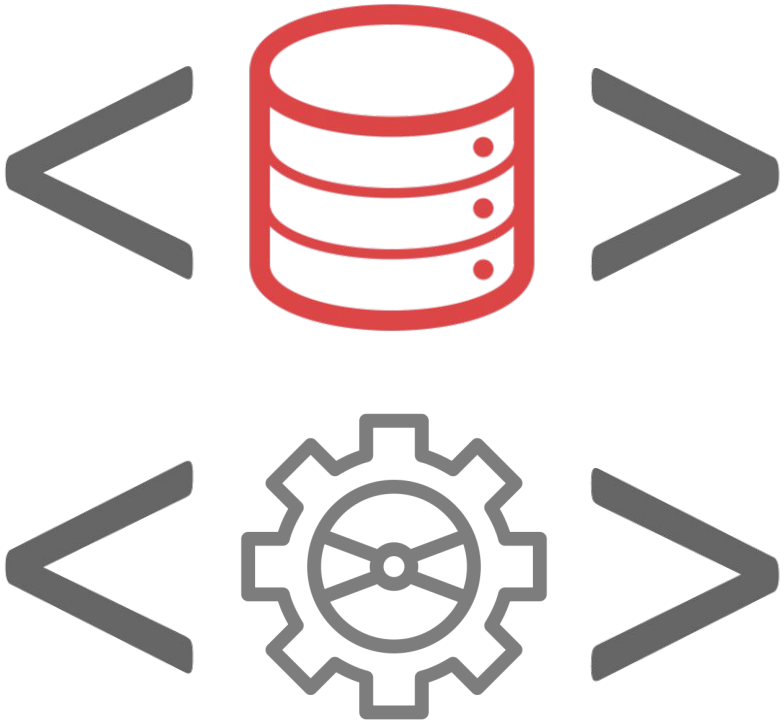
Abstract

The evolutionary response of a trait to directional selection depends upon the level of additive genetic variance. It h
been long argued that sustained selection will tend to deplete additive genetic variance as favoured alleles approac
fixation. Non-additive genetic variance, due to interactions among alleles within and between loci, does not immed



Github documentation

Publisher & Repository submission
Links as metadata



Documentation

Home > Documentation > Reference Linking > **Data and software citation deposit guide**

Search this section

- Documentation
- Setting up as a member
- The research nexus
- Metadata principles and practices
- How to register content
- Schema library
- Reports
- Reference Linking
  - How do I create reference links?
  - Data and software citation deposit guide
- Crossmark
- Cited-by
- Funder Registry

Data and software citation deposit guide

‹ How do I create reference links? Crossmark ›

As well as providing persistent links to scholarly content, we also provide community infrastructure by linking publications to associated content, making research easy to find, cite, link, and assess. Data citations are a core part of this service, linking publications to their supporting data, making both the research itself and the research process more transparent and reproducible.

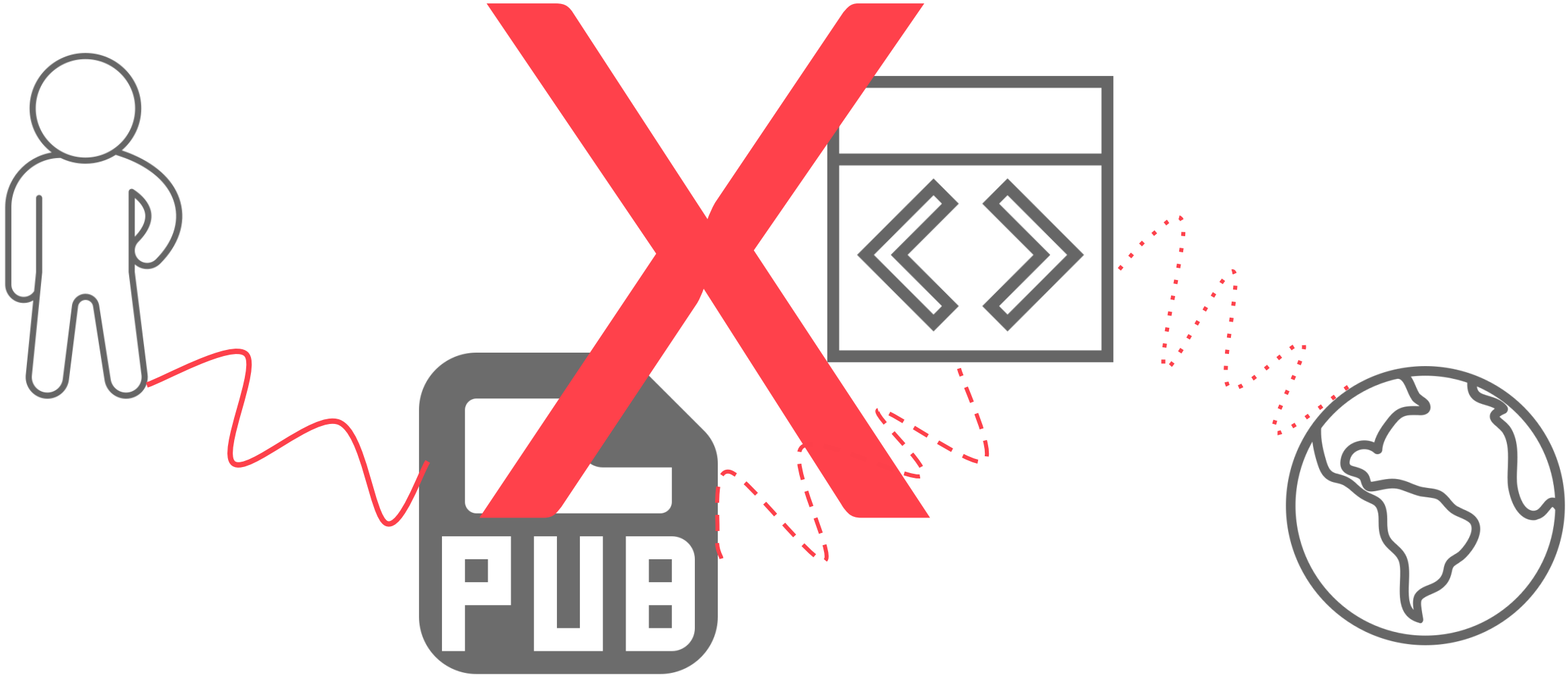
Data citations are references to data, just as bibliographic citations make reference to other scholarly sources.

Members deposit data citations by including them in their metadata as references and/or relationship types. Once deposited, data citations across journals (and publishers) are then aggregated and made freely available for the community to retrieve and reuse in a single, shared location.

There are two ways for members to deposit data citation links:

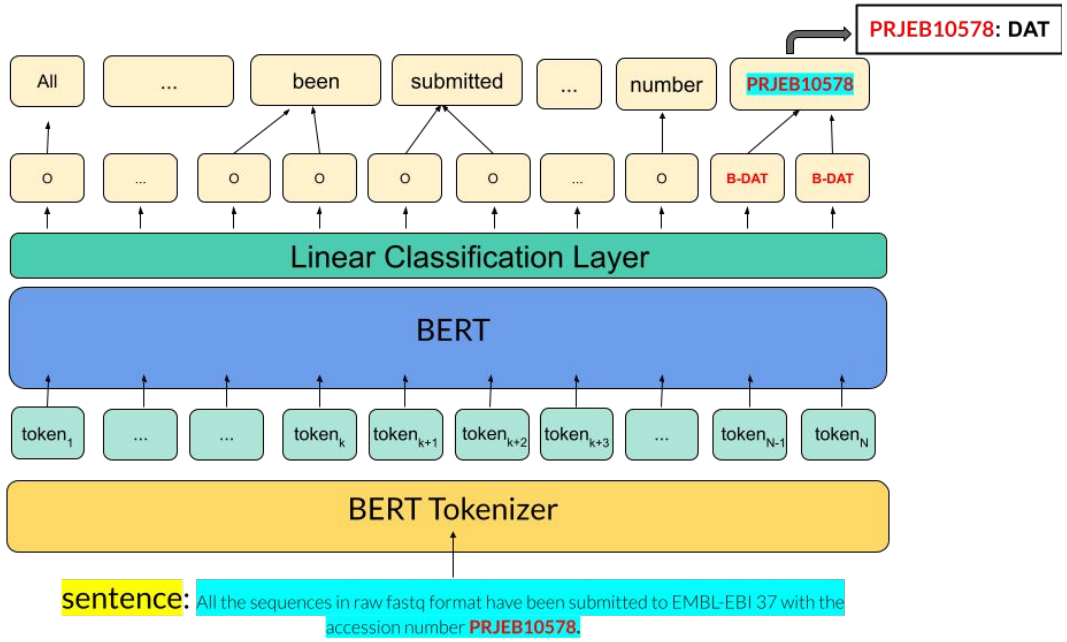
1. Bibliographic references: The main mechanism for depositing data and software citations is to insert them into an article's reference metadata. Data citations are included in the deposit of bibliographic references for each publication. Follow the general process for depositing references and apply tags as applicable.
2. Relationship type: data links are asserted in the relationship section of the metadata deposit,

Current citation production process (workflow) is Long (tedious) & Lossy

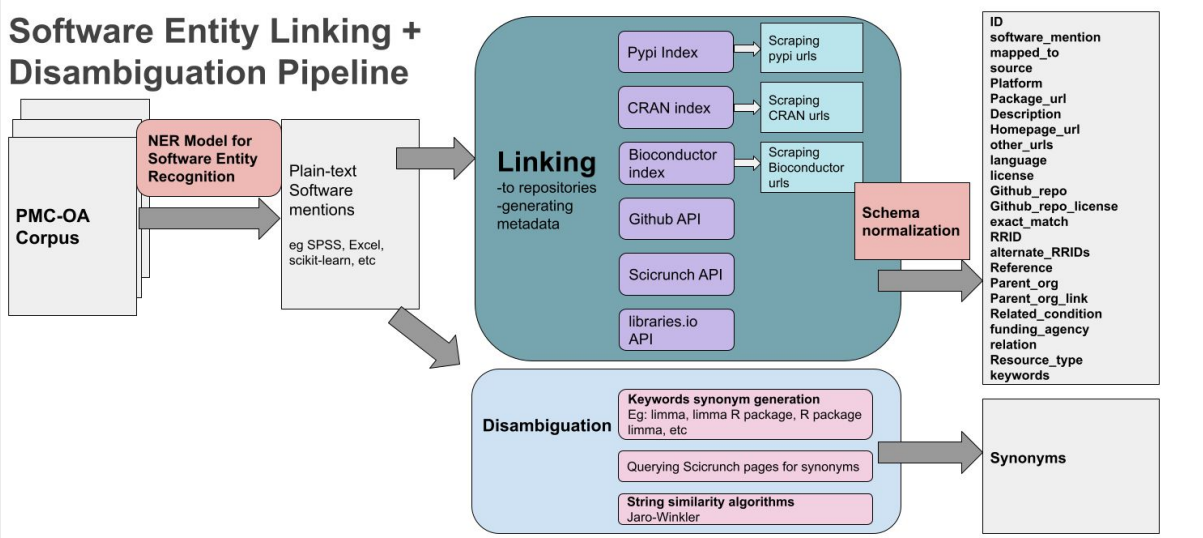


Exploring new frontiers of open citations

1. Extracting Dataset Mentions from the literature using ML



2. Extracting Software mentions from the literature using ML



Definitions

Type of Entity	Datasets	Software
Definition	A collection of data that have been measured, collected, and/or analyzed as part of a research study. Includes: - **Accession Number IDs associated with a database such as GEO, GenBank , or BioProject** - DOIs associated with a repository such as Dryad, Zenodo, or Figshare - resources hosted on external URLs, (academic institutions, organizations)	The set of computer programs, procedures, codes and routines, including those used for research data collection, processing and analysis. - Includes obvious software ***(Image J)*** as well as algorithms and programs ***(bowtie, BLAST)*** - not databases (ArrayExpress, Github), web platforms (Facebook, Google Earth), hardware (Kinect)
Examples	Metagenomes from the Gulf of Mexico are available under the NCBI BioProject PRJNA291283. (*BioProject database)* The microarray data had been previously deposited at Gene Expression Omnibus (GEO) under accession number GSE2603. (*GEO)* The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD001585. (*ProteomeXchange)*	A heatmap was drawn using the R package, gplots [24], and a principle component analysis performed to identify the highest contributing factors All statistical comparisons (nonparametric t tests) were performed with usage of Graph Pad Prism 6 software (AMU licence) Computer programming for this task was done using LabView, version 8.5

Summary

Datasets

	CZI publishers
# full-text papers	~ 16M
# papers with mentions	~ 315k
# total dataset mentions	~ 914k
# unique datasets	~ 400k
# paper-dataset links	~ 700k

Software

	CZI publishers	PMC-OA
# full-text papers	~ 16M	~ 3.8M
# papers with mentions	~ 2.8M	~ 2.4M
# total software mentions	~ 48M	~ 19.2M
# unique software mentions	~ 900k	~ 1.12M

Examples of top extracted datasets

Extracted Dataset	# papers	Database	Dataset Name
GPL570	417	GEO	GPL570: [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array
MN908947.3	386	GenBank	MN908947.3: Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome
MN908947	318	ENA	ENA - MN908947: Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome.
GSE14520	288	GEO	GSE14520: Gene expression data of human hepatocellular carcinoma (HCC)
GSE39582	270	GEO	GSE39582: Gene expression Classification of Colon Cancer defines six molecular subtypes with distinct clinical, molecular and survival characteristics [Expression]
K03455	267	GenBank	K03455.1: Human immunodeficiency virus type 1 (HXB2), complete genome; HIV1/HTLV-III/LAV reference genome
GSE31210	212	GEO	GSE31210: Gene expression data for pathological stage I-II lung adenocarcinomas
GSE16011	176	GEO	GSE16011: Intrinsic Gene Expression Profiles of Gliomas are a Better Predictor of Survival than Histology
GSE2034	174	GEO	GSE2034: Breast cancer relapse free survival
GSE62254	174	GEO	GSE62254: Molecular analysis of gastric cancer identifies discrete subtypes associated with distinct clinical characteristics and survival outcomes: the ACRG (Asian Cancer Research Group) study [gastric tumors]

Examples of top extracted software

Extracted Software	# papers
SPSS	285,279
R	191,817
GraphPad Prism	119,357
ImageJ	94,428
Excel	79,826
GraphPad	75,438
SAS	74,919
BLAST	54,870
Stata	46,279
MATLAB	46,265

Methodology

1. Mining for Datasets from Full-Text Papers

Task: extract mentions of **dataset** entity types from text (NER problem)

Transcriptome and Proteome Exploration to Model Translation Efficiency and Protein Stability in *Lactococcus lactis*

Clémentine Dressaire,[1,2,3] Christophe Gitton,[4] Pascal Loubière,[1,2,3] Véronique Monnet,[4] Isabelle Queinnec,[5] and Muriel Cocaign-Bousquet[1,2,3,*]

Mark Stitt, Editor

Author information ▸ Article notes ▸ Copyright and License information ▸

Associated Data

Supplementary Materials ▸

Abstract

This genome-scale study analysed the various parameters influencing protein levels in cells. To achieve this goal, the model bacterium *Lactococcus lactis* was grown at steady state in continuous cultures at different growth rates, and proteomic and transcriptomic data were thoroughly compared. Ratios of mRNA to protein were highly variable among proteins but also, for a given gene, between the different growth conditions. The modeling of cellular processes combined with a data fitting modeling approach allowed both translation efficiencies and degradation rates to be estimated for each protein in each growth condition. Estimated translational efficiencies and degradation rates strongly differed between proteins and were tested for their biological significance through statistical correlations with relevant parameters such as codon or amino acid bias. These efficiencies and degradation rates were not constant in all growth conditions and were inversely proportional to the growth rate, indicating a more efficient translation at low growth rate but an antagonistic higher rate of protein degradation. Estimated protein median half-lives ranged from 23 to 224 min, underlying the importance of protein degradation notably at low growth rates. The regulation of intracellular protein level was analysed through regulatory coefficient calculations, revealing a complex control depending on protein and growth conditions. The modeling approach enabled translational efficiencies and protein degradation rates to be estimated, two biological parameters extremely difficult to determine experimentally and generally lacking in bacteria. This method is generic and can now be extended to other environments and/or other micro-organisms.

journal article/preprint

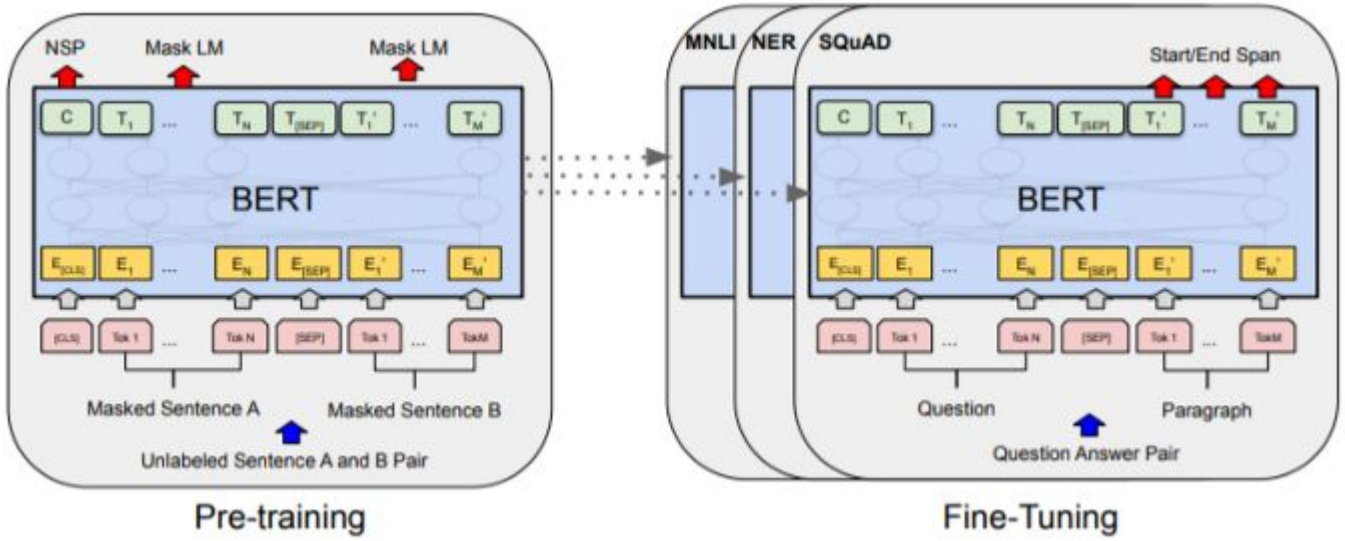
Datasets

GSE29272 *(GEO)*
PRJNA428721 *(BioProject)*
KU182910 *(GenBank)*
PXD015758 *(PRIDE)*
...

Named Entity Recognition (NER) problem

Pre-trained Language Models

- Models pre-trained on huge amounts of text
- They learn to capture complexities in natural language from the training corpus
- BERT, ERNIE, StructBERT, DeBERTa, T5, RoBERTa, ELECTRA, AlBERT, GPT-3, etc



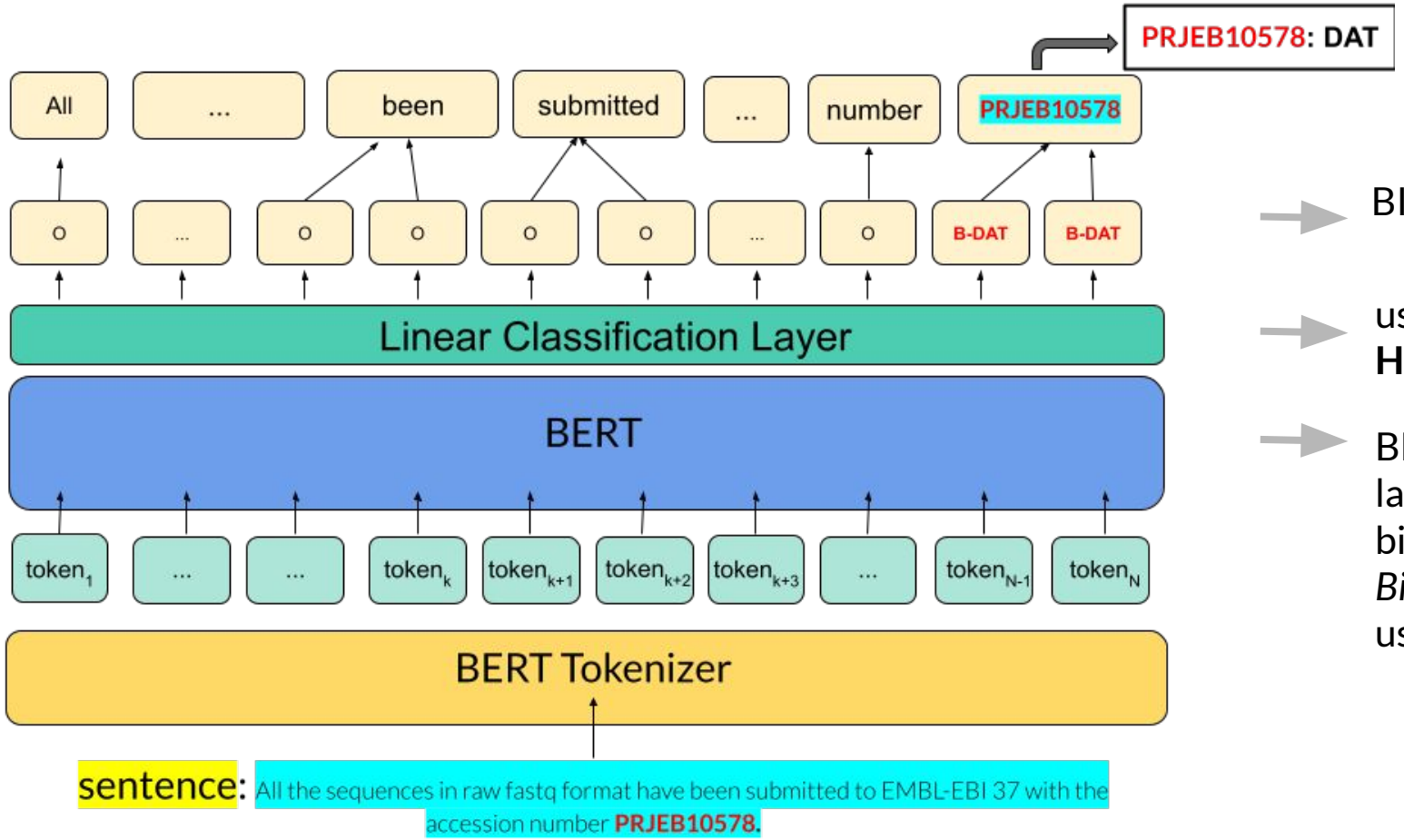
BERT: BooksCorpus + English Wikipedia

NER
Question Answering
Sentence Classification
...

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

ML Model Architecture

Fine-tuning pre-trained language models on the NER task of identifying dataset mentions



BIO-scheme: [B-DAT, O]

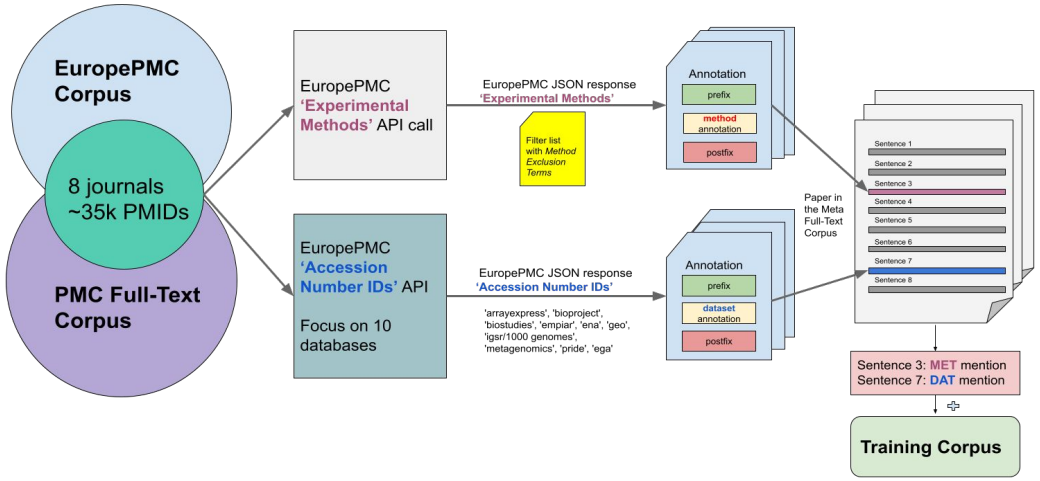
use BertForTokenClassification **Huggingface** module

BERT architecture initialized from language models pre-trained on the biomedical field *BioBERT, SciBERT, PubmedBERT, etc ..* using **Huggingface**

Training Dataset

Built on top of existing work

1. Used openly available annotations from Europe PMC*
2. **Curated the output** with the help of our bio-curation team (eg excluded terms that did not fit our definitions)
3. Mapped the terms to sentences in our full-text corpus



* Europe PMC is an open science platform for articles and preprints
http://europepmc.org/

	Train	Val	Test	Total
# sentences	52206	6526	6526	**65258**
# dataset mentions	862	141	116	**1119**

Training Dataset Composition

Datasets	
Mention	**Frequency**
GM12878	32
GM06990	10
GM12878.	6
PRJNA512236.	5
R10000	4
NA06991.	4
GDS534	3
GSE87339.	3
E-GEOD-40710).	3
NA19238	3

Examples of terms in our training corpus

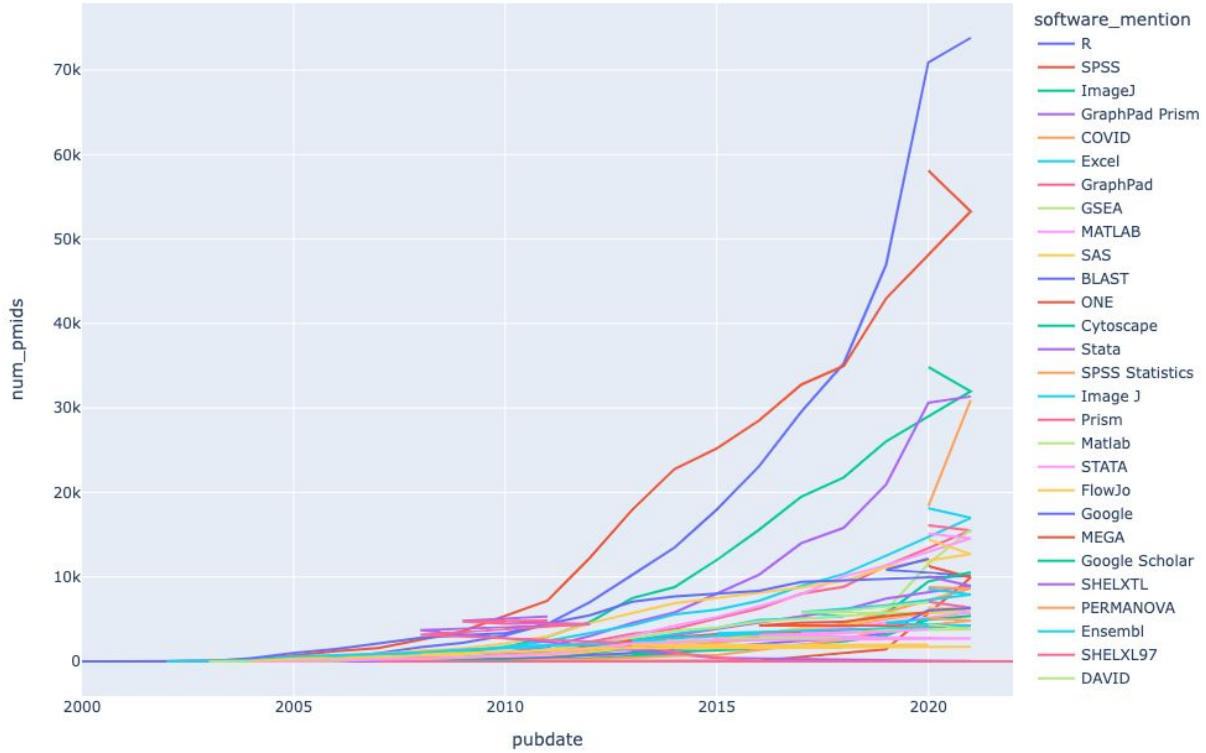
Evaluation

Model	P	R	F1
BioBERT	0.909	0.826	0.865
SciBERT	0.93	0.962	0.946
PubmedBERT	0.857	0.946	0.9
PubmedBERT - FullText	0.819	0.94	0.875
BlueBERT	0.926	0.939	0.932
BlueBERT (MIMIC III)	0.938	0.911	0.924
SapBERT	0.929	0.938	0.933
SapBERT (mean token)	0.948	0.861	0.903
BioELECTRA	0.881	0.937	0.908
BioELECTRA (PMC)	0.878	0.939	0.908
ELECTRAMed	0.907	0.952	0.929
BiomedRoberta (base)	0.875	0.954	0.913
BiomedRoberta (ChemProt)	0.948	0.84	0.891
BiomedRoberta (RCT 500)	0.889	0.947	0.917

Metrics on the test split

2. Software mentions in the literature

Task: Extract, link & disambiguate software mentions from the literature

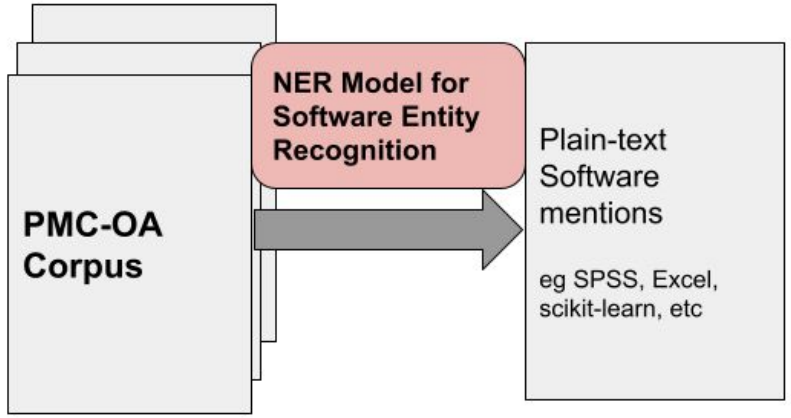


Evolutions of software mentions over time for the top 30 most frequent mentions in the PMC-OA since 2000

Steps

1. NER model to extract plain-text software mentions
2. Linking & Disambiguation to cluster variations of the same software entry together

STEP 1: Extract software mentions from PMC-OA, publishers_collection



NER Model:

- built by our co-worker Ivana Williams
- Scibert model fine-tuned on the SoftCite Dataset

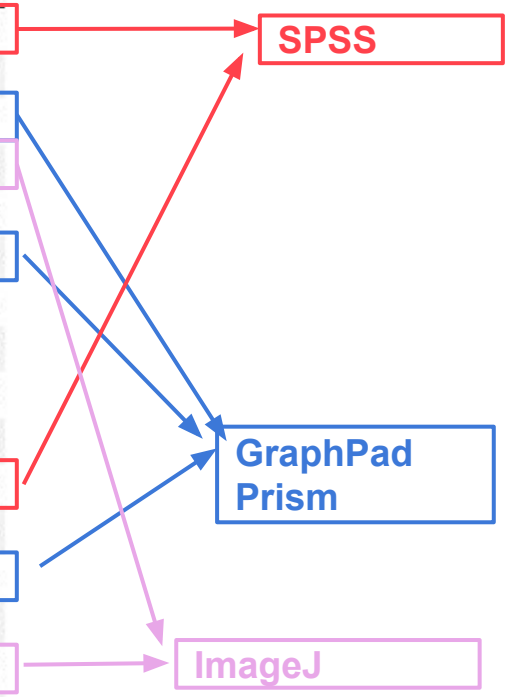
```
f1 socre: 0.922123
Accuracy score: 0.995906
           precision    recall  f1-score   support

 software     0.9014    0.9343    0.9176       959
  version     0.9216    0.9515    0.9363       309

micro avg     0.9063    0.9385    0.9221      1268
macro avg     0.9063    0.9385    0.9221      1268
```

Most frequent plain-text software mentions on the PMC-OA corpus extracted by the NER model

	software_mention	num_pmids
0	SPSS	285279
1	R	191817
2	GraphPad Prism	119357
3	ImageJ	94428
4	Excel	79826
5	GraphPad	75438
6	SAS	74919
7	BLAST	54870
8	Stata	46279
9	MATLAB	46265
10	SPSS Statistics	37875
11	STATA	34149
12	Prism	32155
13	Matlab	30967
14	Image J	30835
15	FlowJo	29008
16	MEGA	24059
17	COVID	20163
18	SHELXL97	19265
19	Cytoscape	18072
20	SHELXS97	17499
21	Ensembl	16915
22	Google Scholar	15330
23	-	15189
24	Google	14604



Challenge: to assess Image J software impact, we need to account for all possible variations in which software appears

STEP 2: Disambiguation

	software_mention	num_pmids
0	SPSS	285279
1	R	191817
2	GraphPad Prism	119357
3	ImageJ	94428
4	Excel	79826
5	GraphPad	75438
6	SAS	74919
7	BLAST	54870
8	Stata	46279
9	MATLAB	46265
10	SPSS Statistics	37875
11	STATA	34149
12	Prism	32155
13	Matlab	30967
14	Image J	30835
15	FlowJo	29008
16	MEGA	24059
17	COVID	20163
18	SHELXL97	19265
19	Cytoscape	18072
20	SHELXS97	17499
21	Ensembl	16915
22	Google Scholar	15330
23	-	15189
24	Google	14604

output from the NER model

Disambiguation

= synonym generation

Challenge: software can be mentioned in a paper under many different variations

CLUSTER 1: **SPSS**

```
spss) statistics
spss)
spss software
spss
Statistical Package for the Social Sciences
SpSS
IBM SPSS Statistics:International Business Machines
SPSS Statistics
IBM SPSS
(spss)
(spss statistics
(SPSS)
(SPSS Statistics
```

CLUSTER 2:**ImageJ**

```
imagej)
imagej - image processing and analysis in java
image j
Imagej
ImageJ2
ImageJ)
ImageJ - Image Processing and Analysis in Java
Image J
ImageJrun
ImageJ-Fiji
ImageJ:Fiji
ImageJ/fiji
ImageJ/JAVA
```

CLUSTER 3: **GraphPad Prism**

```
Graph-PadPrism
Graph-Pad Prim
GhraphPad Prism7
Grappad Prism
GraphPad6 (Prism)
GraphPad prism 7®
Graph pad Prism5®
GraphPad Prism for MacOs X
GraphPad Prism - Graph Pad
GraphPad Prism (Graph Pad)
GraphPrism.5
GraphPadPrim
GraphPad Pad
Graph Prism®
Graph Prisms
Graph Prisma
```

Synonym Generation

Disambiguation

= synonym generation

Keywords synonym generation
Eg: limma, limma R package, R package limma, etc

Querying Scicrunch pages for synonyms

String similarity algorithms
Jaro-Winkler

software_mention	synonym	synonym_conf
scikit-learn	scikit-learn python package	0.99
scikit-learn	scikit-learn python library	0.99
scikit-learn	scikit-learn python	0.99
scikit-learn	scikit-learn library for Python	0.99
scikit-learn	scikit-learn Python package2223	0.99
scikit-learn	scikit-learn Python package for	0.99
scikit-learn	scikit-learn Python package	0.99

software_mention	synonym	synonym_conf
BLASTN	Standard Nucleotide BLAST	1
BLASTN	Nucleotide Blast	1
BLASTN	NCBI BLASTN	1
BLASTN	BLASTn	1
SPSS	Statistical Package for the Social Sciences	1
SPSS	SPSS	1
SPSS	IBM SPSS Statistics:International Business Machines SPSS Statistics	1
SPSS	IBM SPSS	1

software_mention	synonym	synonym_conf	synonym_source
SPSS Statistics	SPSS Statistics	1.0	string_similarity
SPSS Statistics	SPSS Statistics©	0.986667	string_similarity
SPSS Statistics	SPSS Statisticsm	0.986667	string_similarity
SPSS Statistics	SPSS Statistics)	0.986667	string_similarity
SPSS Statistics	SPSS Statistics®	0.986667	string_similarity
SPSS Statistics	SPSS Statistic	0.985714	string_similarity
SPSS Statistics	SPSSS Statistics	0.984444	string_similarity
SPSS Statistics	SPS Statistics	0.980952	string_similarity
SPSS Statistics	SPPSS Statistics	0.98	string_similarity

similarity_matrix

clustering (DBSCAN)

C1 C2 C2

Jaro-Winkler String Similarity Metric
= metric for measuring edit distance between two sequences

- favors strings that match on longer prefixes
- Jaro–Winkler distance - Wikipedia

STEP 3: Link to repositories

CLUSTER 1
scikit-learn: scikit-learn API
Python toolbox scikit-learn Scikit-Learn
library scikits-learn
scikit-learn (sklearn)
Sklearn.svm
Sklearn
Sklearn API
…

CLUSTER 2
limma: R package limma,
r package limma
package "limma
package limma
Bioconductor R-package limma
limma package (Linear Model for Microarray
Data}
…

CLUSTER 3
BLAST: BLAST Genome search
BLAST SEARCH
BLAST Search
BLAST Search tool
BLAST Searching
BLAST Sequence Similarity Search
BLAST TEXT
Blast searches
…

exact match search on cluster name

Linking
-to repositories
-generating metadata

- Pypi Index
- CRAN index
- Bioconductor index
- Github API
- Scicrunch API

Schema normalization

CLUSTER 1
Scikit-learn:
https://pypi.org/project/scikit-learn
…

CLUSTER 2
limma:
https://www.bioconductor.org/packages/release/bioc/html/limma.html

CLUSTER 3
BLAST:
https://scicrunch.org/browse/resources/SCR_008419
…

cluster name: mention with highest frequency on the PMC-OA corpus

High-level Statistics

Software - Disambiguation + Linking on PMC-OA comm

	# mentions	%	paper-mention links	%	notes
non-disambiguated	~ 393k	35%	~ 700k	8.95%	no sig synonyms
non-disambiguated	~ 400k	36.11%	~ 1M	12.85%	no output from clustering
disambiguated	~ 323k	28.88%	~ 6.3M	78.18%	~ 97k unique software entities
disambiguated + linked	~ 185k	16.55%	~ 4.5M	55.78%	
# unique mentions	~ 1.12M		~ 8M		

Resources

Datasets

- Github Repo: https://github.com/chanzuckerberg/full-text-mining-ner
- Extracted data-paper links: https://github.com/chanzuckerberg/full-text-mining-ner/tree/main/extracted_data(CC0)

Software

- Dataset available on Dryad: CZ Software Mentions Dataset: A large dataset of software mentions in the biomedical literature (CC0)
- ArXiv preprint: https://arxiv.org/abs/2209.00693
- Github Repo: GitHub - chanzuckerberg/software-mentions
- Blog Post: https://medium.com/czi-technology/new-data-reveals-the-hidden-impact-of-open-source-in-science-11cc4a16fea2

Thank you!

CZI-wide

https://twitter.com/ChanZuckerberg
https://www.facebook.com/chanzuckerberginitiative/
https://www.instagram.com/chanzuckerberginitiative
www.linkedin.com/company/chan-zuckerberg-initiative
https://www.youtube.com/channel/UCZioJ6fb9SuRdLIO7DIE09w
https://medium.com/czi-technology

CZI Science

https://twitter.com/cziscience
https://medium.com/@cziscience

Ana-Maria Istrate

https://twitter.com/_aistrate
www.linkedin.com/in/amistrate

References

Models

1. **BioBERT:** J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4):1234–1240, 2020.
2. **SciBERT:** I. Beltagy, K. Lo, and A. Cohan. Scibert: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676, 2019.
3. **PubmedBERT:** Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon. Domain-specific language model pretraining for biomedical natural language processing. arXiv preprint arXiv:2007.15779, 2020.
4. **BlueBERT:** Y. Peng, S. Yan, and Z. Lu. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. arXiv preprint arXiv:1906.05474, 2019.
5. **ClinicalBERT:** E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott. Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323, 2019.
6. **BioELECTRA:** K. raj Kanakarajan, B. Kundumani, and M. Sankarasubbu. Bioelectra: Pretrained biomedical text encoder using discriminators. In Proceedings of the 20th Workshop on Biomedical Language Processing, pages 143–154, 2021.
7. **ELECTRAMed:** G. Miolo, G. Mantoan, and C. Orsenigo. Electramed: a new pre-trained language representation model for biomedical nlp. arXiv preprint arXiv:2104.09585, 2021.
8. **SapBERT** F. Liu, E. Shareghi, Z. Meng, M. Basaldella, and N. Collier. Self-alignment pretraining for biomedical entity representations. arXiv preprint arXiv:2010.11784, 2020.
9. **BiomedRoberta:** S. Gururangan, A. Marasovic, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. ´ Smith. Don't stop pretraining: adapt language models to domains and tasks. arXiv preprint arXiv:2004.10964, 2020.

Biomedical Papers

1. Bonet, Jaume, et al. "Rosetta FunFolDes–A general framework for the computational design of functional proteins." *PLoS computational biology* 14.11 (2018): e1006623.
2. Lawrence, Michael, et al. "Software for computing and annotating genomic ranges." *PLoS computational biology* 9.8 (2013): e1003118.
3. Han, Ju, et al. "Molecular predictors of 3D morphogenesis by breast cancer cell lines in 3D culture." *PLoS computational biology* 6.2 (2010): e1000684.
4. Rex, Julia, et al. "Model-based characterization of inflammatory gene expression patterns of activated macrophages." *PLoS computational biology* 12.7 (2016): e1005018.
5. Stallard, Paul, et al. "Classroom based cognitive behavioural therapy in reducing symptoms of depression in high risk adolescents: pragmatic cluster randomised controlled trial." *Bmj* 345 (2012).
6. Rajagopal, Nisha, et al. "RFECS: a random-forest based algorithm for enhancer identification from chromatin state." *PLoS computational biology* 9.3 (2013): e1002968.
7. Glover, Clive H., et al. "Meta-analysis of differentiating mouse embryonic stem cell gene expression kinetics reveals early change of a small gene set." *PLoS computational biology* 2.11 (2006): e158.
8. Dubourg, Grégory, et al. "Gut microbiota associated with HIV infection is significantly enriched in bacteria tolerant to oxygen." *BMJ open gastroenterology* 3.1 (2016): e000080.
9. Newberg, Justin Y., et al. "SBCDDB: Sleeping Beauty Cancer Driver Database for gene discovery in mouse models of human cancers." *Nucleic acids research* 46.D1 (2018): D1011-D1017.
10. Fabregat, Antonio, et al. "The reactome pathway knowledgebase." *Nucleic acids research* 46.D1 (2018): D649-D655.