Chan Zuckerberg Initiative

Building the Open Global Data Citation Corpus

Ana-Maria Istrate, Senior Research Scientist

Webinar • Feb 13, 2023

CZI Science

Supporting the science and technology that will make it possible to cure, prevent, or manage all diseases by the end of this century.

Open Science

Universal and immediate open sharing of all scientific knowledge, processes and outputs

We want to identify and democratize emerging and valuable **methods, tools**, and **datasets** and bring them to a broad and diverse set of scientists

so that they can come to meaningful conclusions faster

We create and share datasets on key research resources

GLOBAL BIODATA COALITION

CZI/GBC collaboration to surface biodata resources from full-text papers to build the Global Biodata Resource Inventory

DRYAD

CZ Software Mentions: A large dataset of software mentions in the biomedical literature

Dataset of software mentions from the biomedical literature (CC0)

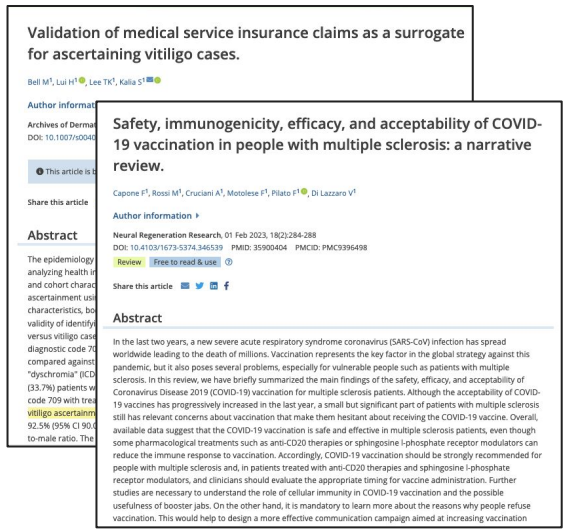
Chan Zuckerberg Initiative

DataCite

joining forces to **increase discoverability of datasets**

Dataset Discoverability

- Data aggregators (DataCite, Wikidata) have made it easier to discover datasets
- However, they don't have **comprehensive coverage**
  - Many domain specific repositories are not included
  - Not all datasets have DOIs
  - Majority of datasets are mentioned (not formally cited) in full-text of papers



extract datasets from the source

Datasets

GSE40279
https://identifiers.org/geo:GSE40279

GSE51032
https://identifiers.org/geo:GSE51032

https://doi.org/10.17632/RT6X6362YX.1

Definition

- A *dataset* is a collection of data that have been measured, collected, and/or analyzed as part of a research study.

- Datasets can be mentioned by:
  - **Accession Number IDs associated with a database** such as GEO, or BioProject, etc
  - **DOIs** associated with a repository such as Dryad, Zenodo, or Figshare, etc
  - **resources hosted on external URLs,** such as academic institutions or organizations

Data availability

All data generated or analysed during this study are included in the manuscript, supporting files and on https://github.com/perslab/timshel-2020 (copy archived at https://github.com/elifesciences-publications/timshel-2020).

The following previously published datasets were used:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Gloudemans M, Balliu B	2018	GWAS studies	https://github.com/mike-gloudemans/gwas-download	GitHub, gwas-download
Romanov RA, Zeisel A, Bakker J, Girach F, Hellysaz A, Tomer R, Alpár A, Mulder J, Clotman F, Keimpema E, Hsueh B, Crow AK, Martens H, Schwindling C, Calvigioni D, Bains	2017	Hypothalamus - HYPR	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE74672	NCBI Gene Expression Omnibus GSE74672
Kim D-W, Yao Z, Graybuck LT, Kim TK, Nguyen TN,	2019	Hypothalamus - VMH	https://doi.org/10.17632/ypx3sw2f7c.1	Mendeley Data, 10.17632/ypx3sw2f7c.1

URL

Accession number

DOI

Adapted from https://elifesciences.org/articles/55851

Dataset Accession Number IDs

Methylome data were downloaded from Hannum et al[5] and EPIC[26] (Gene Expression Omnibus, ***GSE40279*** and ***GSE51032***) and were processed alongside the methylation data generated from our sample.

https://doi.org/10.1001/jamanetworkopen.2020.15428

GSE40279
https://identifiers.org/geo:GSE40279

GSE51032
https://identifiers.org/geo:GSE51032

DOIs

Data associated with this study has been deposited at Mendeley Data under the accession number ***https://doi.org/10.17632/RT6X6362YX.1***.

https://doi.org/10.1016/j.heliyon.2020.e05507

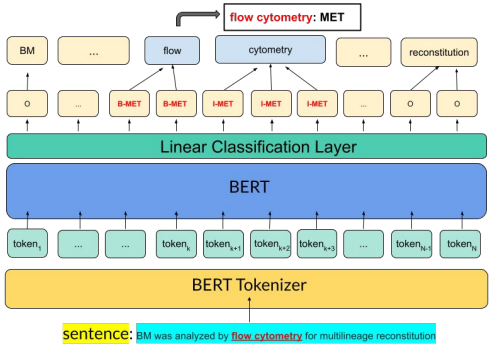
https://doi.org/10.17632/RT6X6362YX.1

Methods

How We Did It

Machine Learning Model

SciBERT-based Named Entity Recognition



Corpus

CZI Full-Text, Europe PMC Full-Text

Extraction

Retrieve dataset mentions

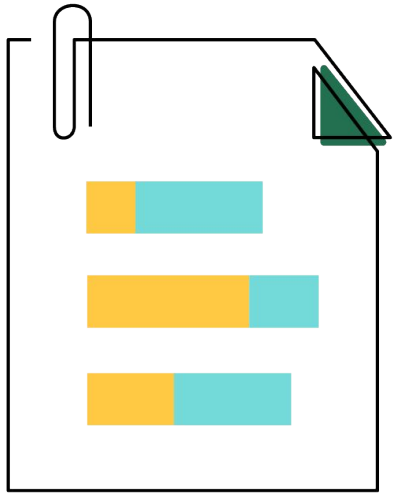
The microarray data had been previously deposited at Gene Expression Omnibus (GEO) under accession number **GSE2603.**

Linking

Link to a repository

https://identifiers.org/geo:GSE2603

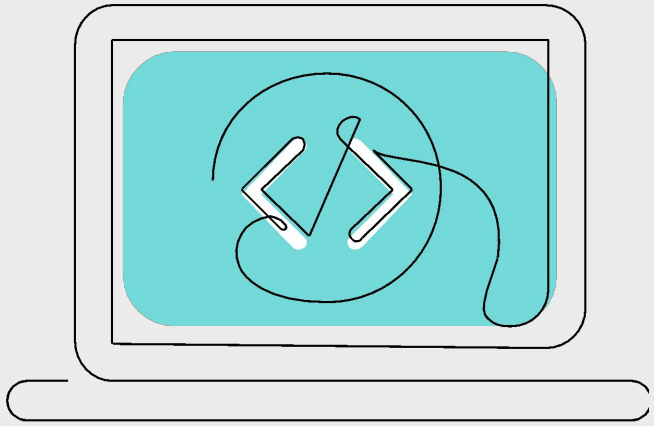
CZI Contribution to the Open Global Data Citation Corpus



seed datafile

dataset-paper links extracted with ML models from Europe PMC Full-Text Open-Access Corpus

working with DataCite and other partners to keep the corpus refreshed



algorithms

new ML methodology in mining datasets from full-text papers

will be open-sourced

Thank you!

CZI Science

@cziscience

https://medium.com/@cziscience

CZI Open Science

https://czi.co/OpenScience

CZI Science Tech

https://tech.chanzuckerberg.com/scitech/