# THE DATA CITATION CORPUS

## Enabling easier access to data citations & evaluation of data usage at scale

Iratxe Puebla, DataCite iratxe.puebla@datacite.org 0000-0003-1258-0746

SCAN ME

MAKE DATA COUNT

### PROBLEM: CHALLENGES UNDERSTANDING DATA USAGE

With millions of datasets now openly available, there is a pressing need to understand the reach and impact of open data. This is necessary to create incentives that make data sharing **rewarding,** and to inform **best practices for data stewardship.**

**Data citations** are a measure of data usage, but finding and analyzing them is a challenge for various reasons:

- Insufficient adoption of workflows and best practices
- Researchers refer to data in a variety of ways
- Citations stored in different, often closed locations

### SOLUTION: AN OPEN, COMPREHENSIVE VIEW OF DATA CITATIONS

We need easier and more comprehensive ways to get access to data citations.

## THE DATA CITATION CORPUS

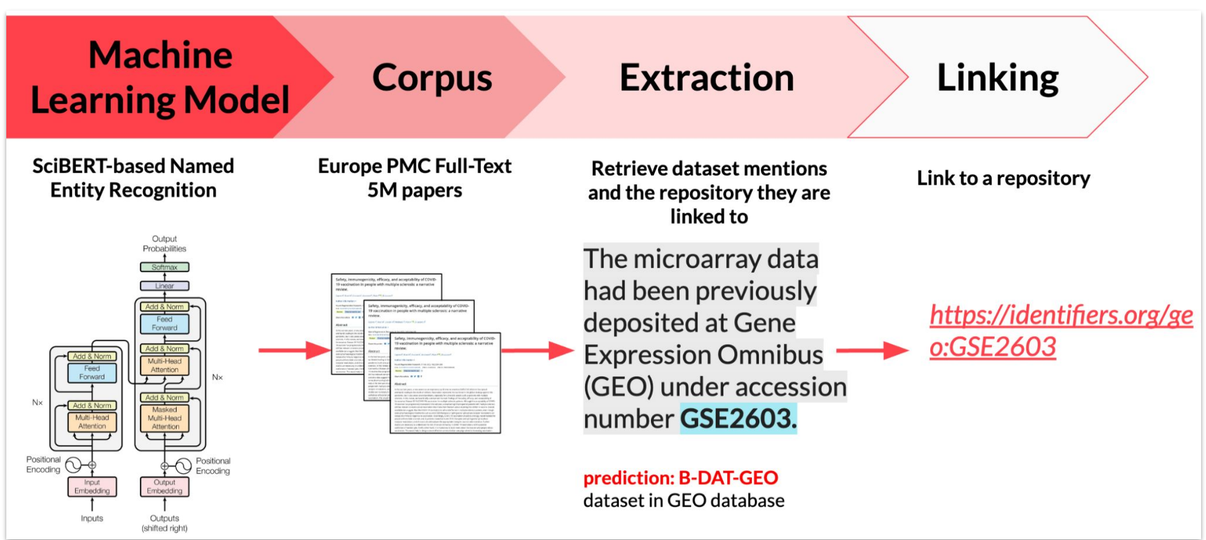
A comprehensive corpus that aggregates data citations from different sources into a centralized, publicly accessible community resource.

Data citations collected via persistent identifier metadata + Data citations identified via additional methods, e.g. curation, machine learning


### BUILDING THE DATA CITATION CORPUS

Collaboration between DataCite and Chan Zuckerberg Initiative (CZI) to bring together citations to data with DOIs and accession numbers.

- ✓ Data citations from DOI metadata, via DataCite Event Data
- ✓ Data citations identified by mining article full text using a SciBERT-base Named Entity Recognition model
  - Mined 5.3 million articles in Europe PMC
  - Searched for dataset DOIs and accession numbers for 44 repositories in the life sciences




### FIRST RELEASE OF THE DATA CITATION CORPUS

The first release of the Data Citation Corpus includes:

- **1.3 million** data citations via DataCite Event Data
- **8.5 million** data citations from CZI

Available via a data file and the Data Citation Corpus dashboard: http://corpus.datacite.org/dashboard

### NEXT STEPS

- Refine model to address false positives
- Enrich metadata for existing data citations: affiliation, funder, subject
- Ingest citations from additional sources to increase coverage across disciplines

### CONTRIBUTE TO THE DATA CITATION CORPUS

- ✓ Request the data file and provide feedback
- ✓ Submit citations to the Data Citation Corpus

**tinyurl.com/datacitationcorpus**

DataCite

Chan Zuckerberg Initiative

wellcome

Wellcome Trust grant 226453/Z/22/Z