COVID-19 DATA in HARVARD DATAVERSE

NIH Webinar on Sharing, Discovering, and Citing COVID-19 Data and Code in Generalist Repositories

April 24, 2020

Mercè Crosas, Ph.D.
University Research Data Management Officer
Chief Data Science and Technology Officer, IQSS
Harvard University

Picture of Barcelona public health professionals by Samuel Aranda for The New York Times

@mercecrosas @dataverseorg @HarvardDV

HARVARD
Dataverse

- Dataverse is an open-source software with a growing, active community
- Harvard Dataverse is one of the 56 Dataverse repositories world-wide
- Open to all research domains
- Researchers can deposit for free
- New data curation services (fee-based)

3,754 dataverses in Harvard Dataverse (collections of datasets)

96K searchable datasets

35K deposited datasets

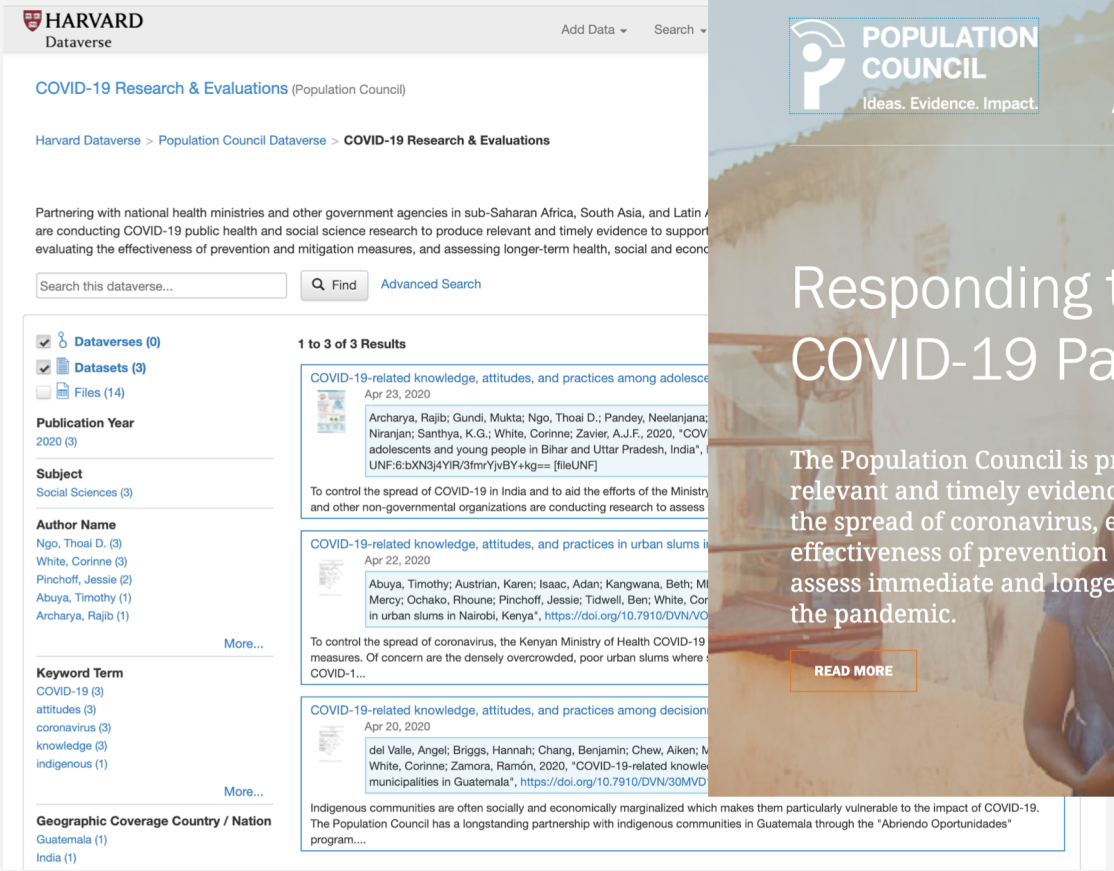
437K files

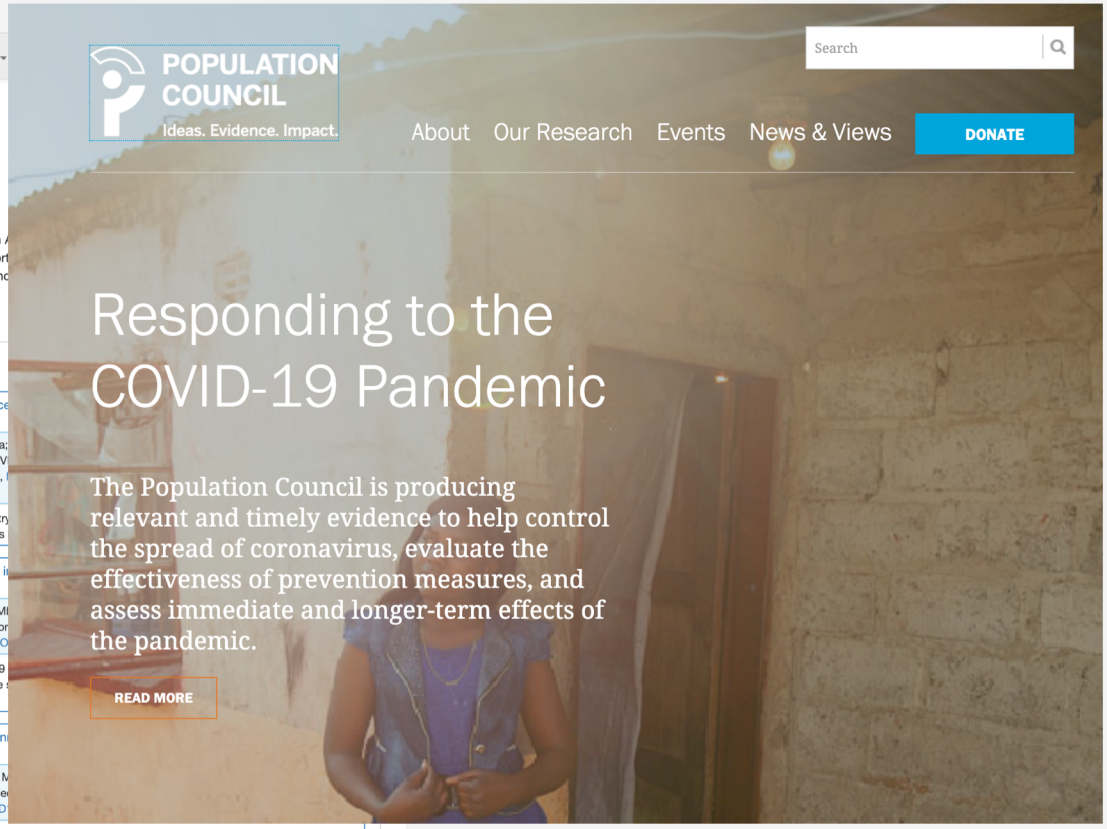
13M downloads

COVID-19 Datasets in Harvard Dataverse

- 50 COVID-19 datasets with 2,043 data files deposited since February 7
- Total of 48,471 downloads by April 23
- Include datasets on: COVID-19 statistics, social science studies to evaluate the effectiveness of government measures, survey data, Twitter data, gubernatorial responses, replication data for preprints

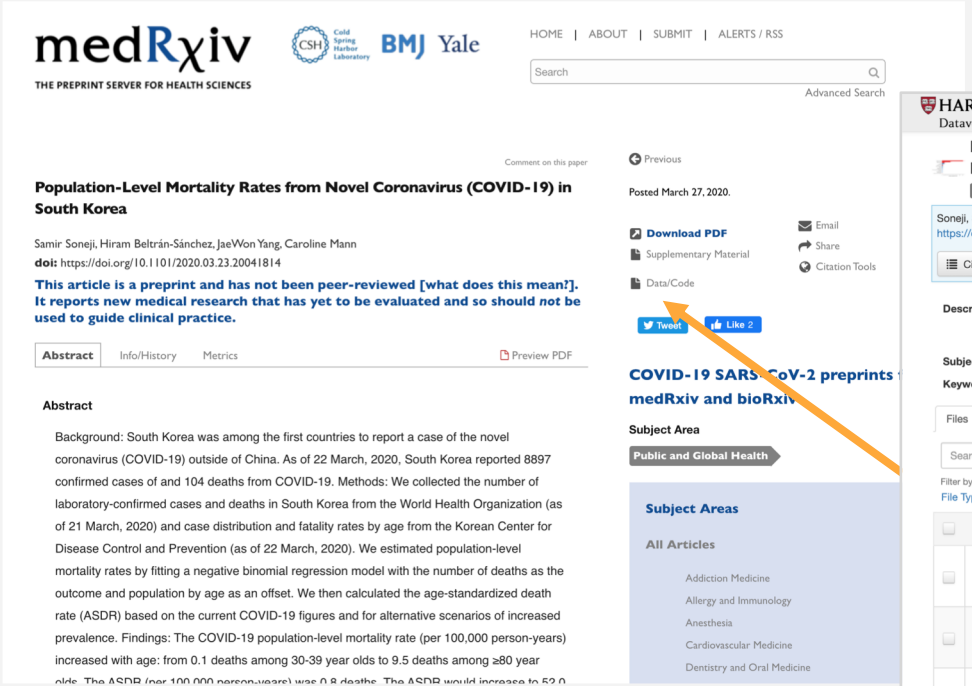
Example of COVID-19 datasets in Harvard Dataverse





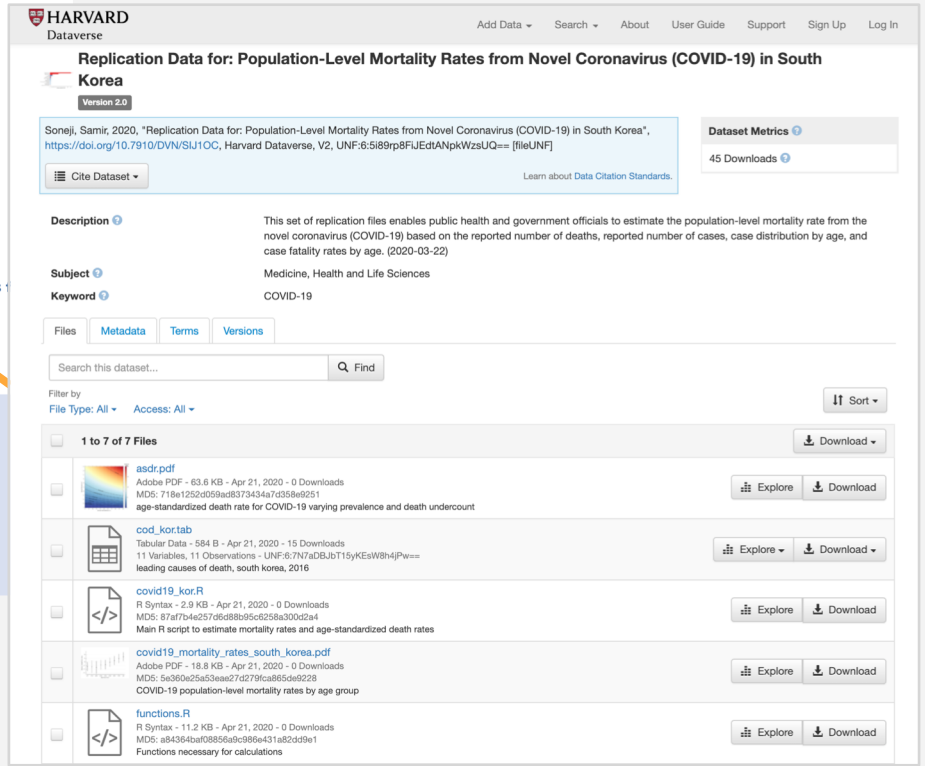
@mercecrosas @dataverseorg @HarvardDV

Example of COVID-19 replication data

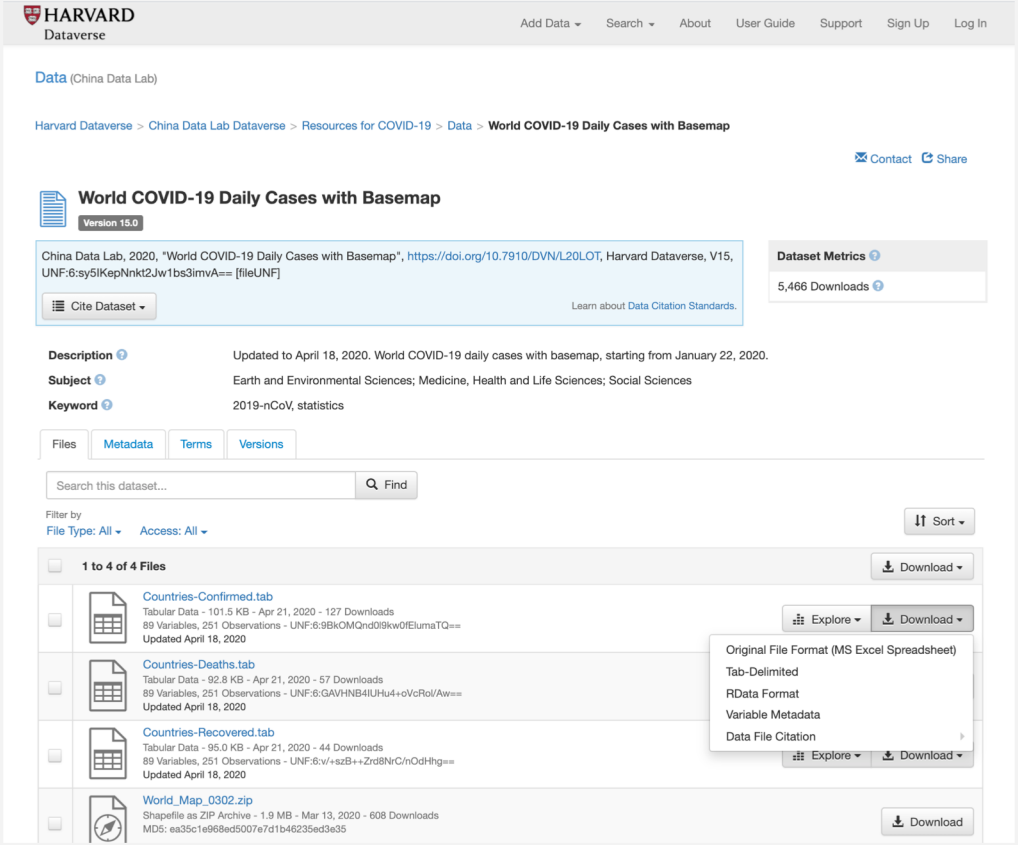


Preprint

Replication Data + Code



Most popular dataset in Harvard Dataverse in 2020



Published on February 20

5,466 Downloads

8,293 Page Views

From China Data Lab

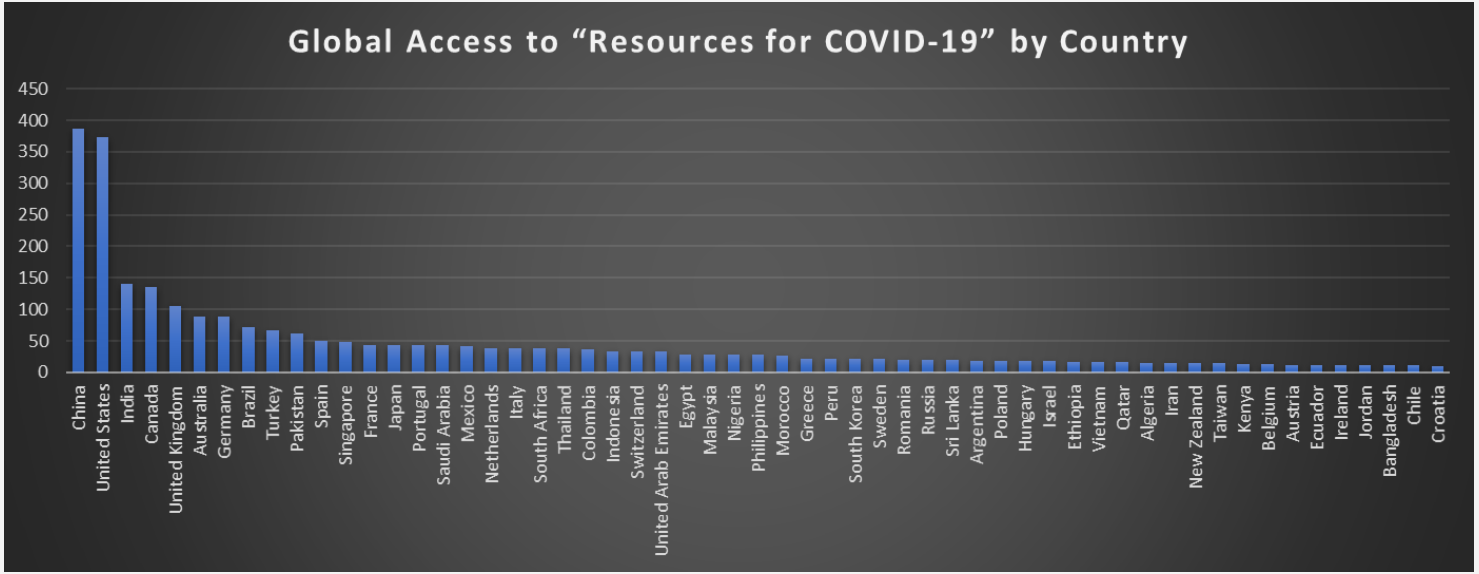
Resources for Novel Coronavirus Global Research

- Part of a larger NSF IUCRC “Spatial Data Lab” project
- Focuses on COVID-19 data and analytical tools (GIS, Statistics)
- Enables access to cloud computing resources
- Facilitates creating, sharing analytical workflows
- All datasets published in Harvard Dataverse
- Datasets include: China census, socio-economic data, China, US, and Global daily COVID-19 data, health facilities, policies and regulations, climate, population mobilities.
- Datasets updated weekly

Credit to Wendy Guan, Tao Hu (IQSS, CGA, Harvard University)

Resources for Novel Coronavirus Global Research

- “Resource for COVID-19” dataverse has had more than 30,000 downloads
- Over 130 countries have accessed the datasets



Credit to Wendy Guan, Tao Hu (IQSS, CGA, Harvard University)

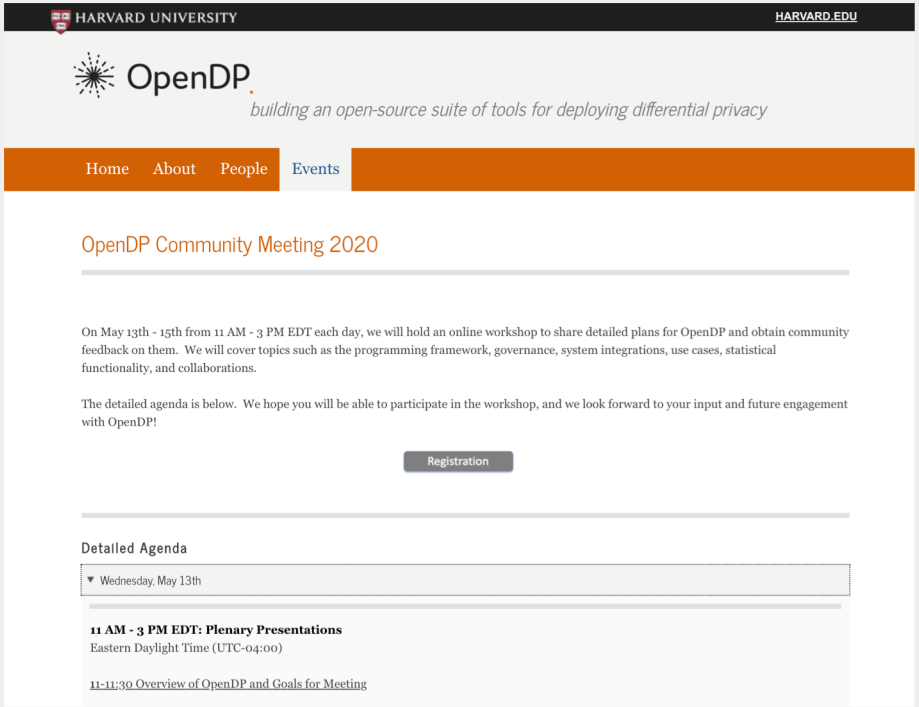
Making European Union COVID-19 Datasets FAIR

- In the European Union coronavirus datasets mostly not shared as FAIR data
- Some countries (Italy, Austria) publish their data on GitHub, without persistent identifiers
- Most countries share the official statistics as PDFs, sometimes with incorrect or missing data points
- In Spain and the Netherlands all data collected and processed by volunteers and shared on GitHub
- It's a big challenge to force official institutions to translate all their variables from the national language to English and use the same codebook for all countries to make data interoperable
- Slava Tykhonov from DANS-KNAW created an European COVID-19 data hub and archive with FAIR datasets in the Harvard Dataverse, updated daily: https://dataverse.harvard.edu/dataverse/covid-19-eu

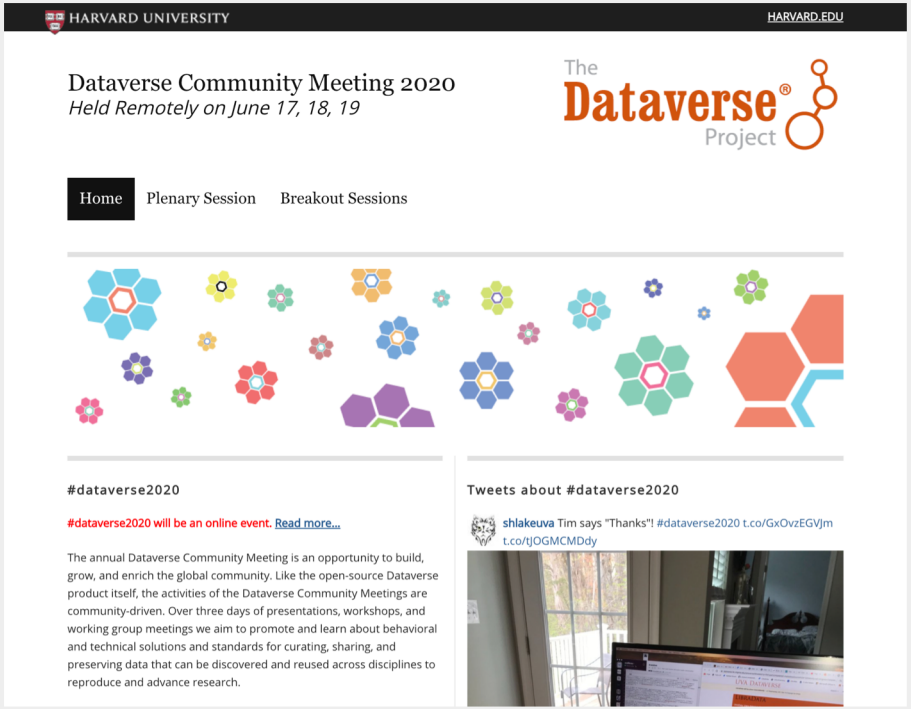
Another interesting initiative that started collecting and processing coronavirus data through Harvard Dataverse is CoronaWhy, a global distributed organization, mobilizing volunteers and community partners to address the current challenge through data science, AI, and knowledge sharing.

Thanks and Upcoming Events

OpenDP Community Meeting, May 13-15, 2020:
Session on Differential Privacy to Explore and Analyze COVID-19 sensitive data



Dataverse Community Meeting, June 17-19, 2020:
Session on Global COVID-19 Data Sharing



@mercecrosas @dataverseorg @HarvardDV