El Data Commons
o com facilitar la col·laboració i accés a les dades de recerca; una visió

CSUC

Cicle de Conferències sobre la gestió de dades de recerca
Consorci de Serveis Universitaris de Catalunya
Novembre 18, 2020

Mercè Crosas, Ph.D., Harvard University
scholar.harvard.edu/mercecrosas @mercecrosas

IQSS
The Institute for Quantitative Social Science

HARVARD UNIVERSITY

Photo: Boston Common
Creator: coleong | Credit: Getty Images/iStockphoto

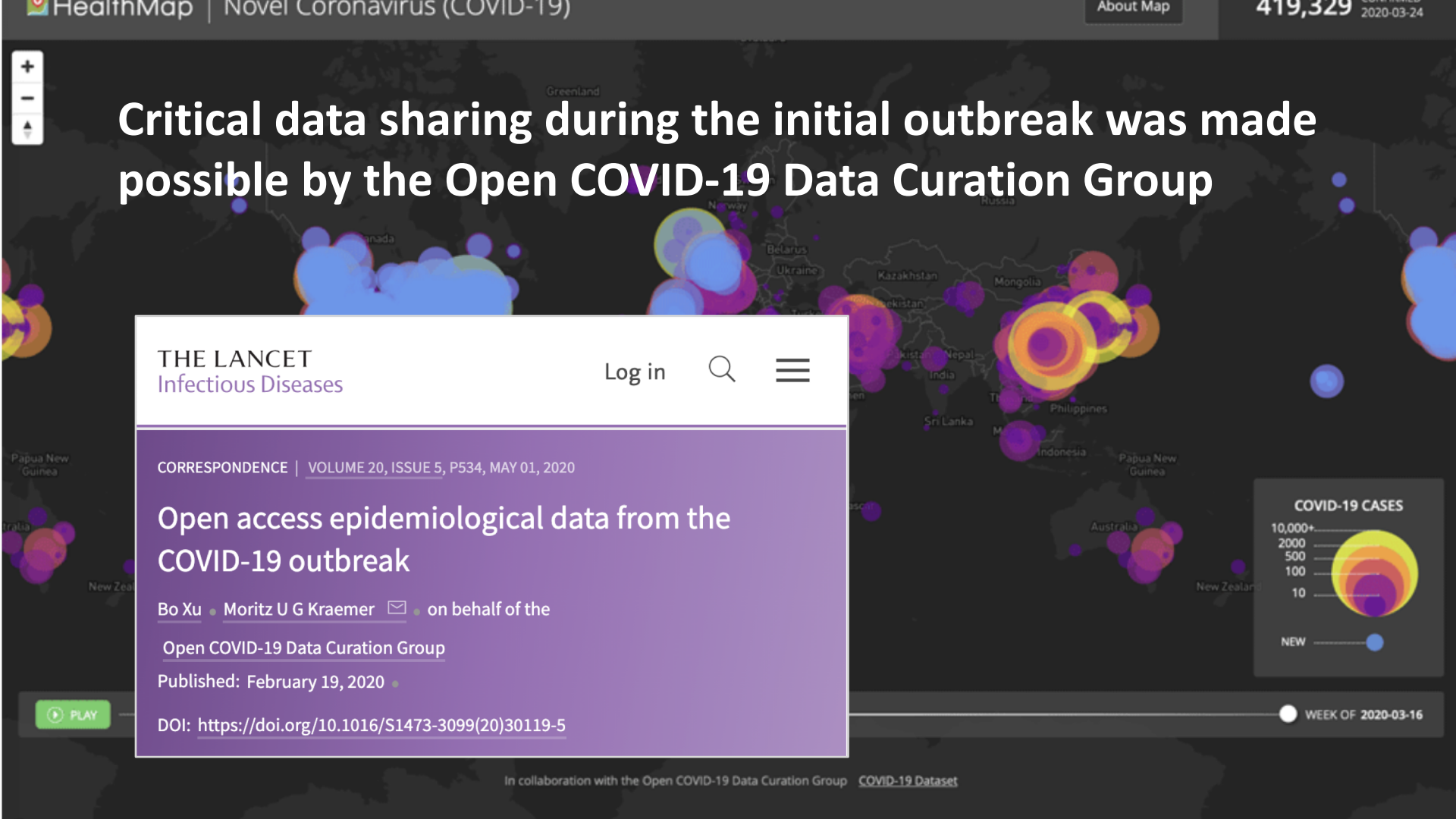
How to facilitate collaboration and access to research data

- Progress
- Challenges
- Vision

Progress

Data sharing is defined as “making data available to people other than those who have generated them”.

https://www.energy.gov/eere/funding/eere-digital-data-management-glossary

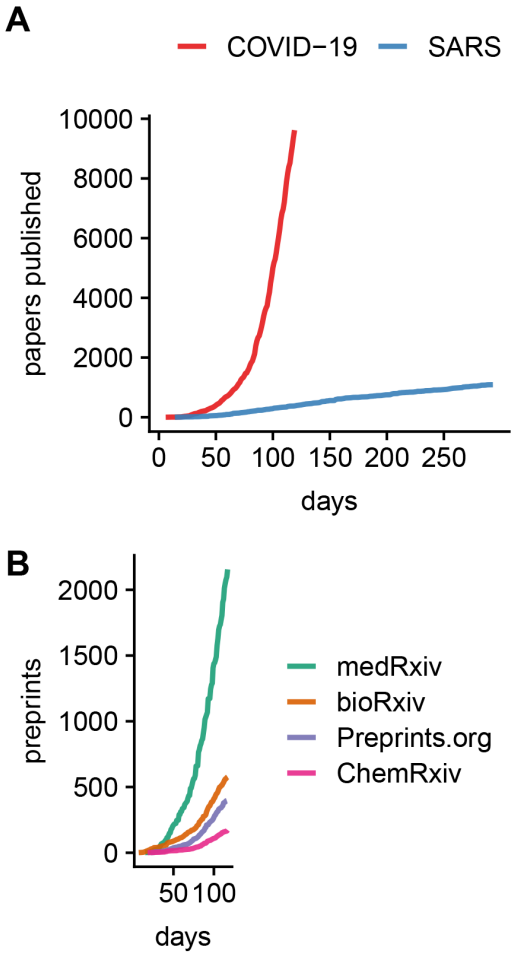

HealthMap | Novel Coronavirus (COVID-19)
About Map
419,329
2020-03-24
Critical data sharing during the initial outbreak was made possible by the Open COVID-19 Data Curation Group
THE LANCET
Infectious Diseases
Log in
CORRESPONDENCE | VOLUME 20, ISSUE 5, P534, MAY 01, 2020
Open access epidemiological data from the COVID-19 outbreak
Bo Xu • Moritz U G Kraemer • on behalf of the
Open COVID-19 Data Curation Group
Published: February 19, 2020 •
DOI: https://doi.org/10.1016/S1473-3099(20)30119-5
COVID-19 CASES
10,000+
2000
500
100
10
NEW
PLAY
WEEK OF 2020-03-16
In collaboration with the Open COVID-19 Data Curation Group COVID-19 Dataset

SIGNIFICANCE

Science in the face of Covid-19: faster, better, stronger?

Written by Simon Schwab and Leonhard Held on 08 May 2020.

“During the pandemic, conflicting information can undermine trust in science. Openness and the **sharing of data** enable researchers to **collaborate, review and reproduce findings**, and such activities strengthen trust in science.”



Rapid and responsible data sharing was also key in previous outbreaks

nature
International weekly journal of science

NATURE | COMMENT

Data sharing: Make outbreak research open access

Nathan L. Yozwiak, Stephen F. Schaffner & Pardis C. Sabeti

25 February 2015

Establish principles for rapid and responsible data sharing in epidemics, urge Nathan L. Yozwiak, Stephen F. Schaffner and Pardis C. Sabeti.

The benefits of responsible and transparent data sharing are not unique to outbreaks

- Enables collaboration
- Advances research
- Helps confirm published results
- Strengthens trust in science

Clear progress in data sharing and access

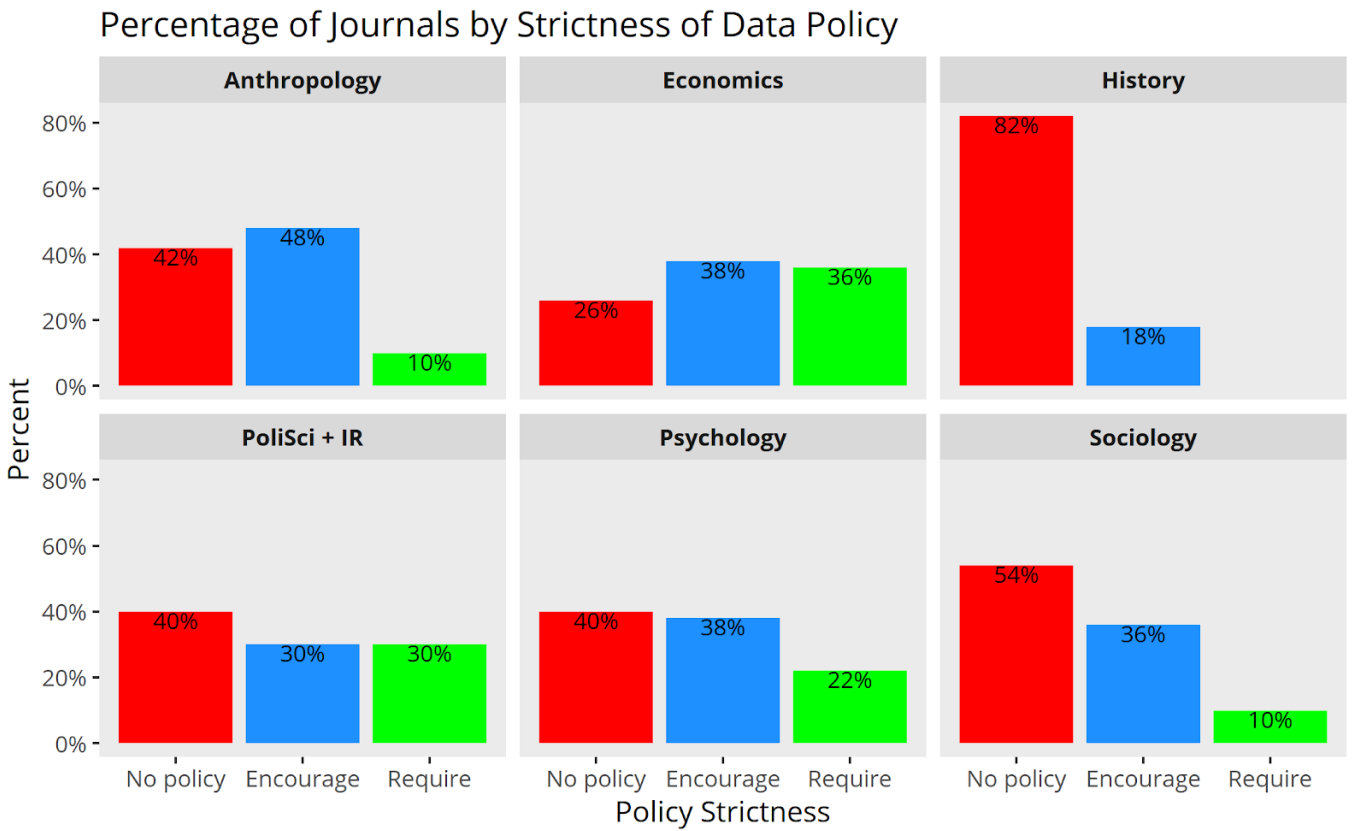
- **New data policies in journals**
  - *Example:* > 50% of top social science journals recommend or require sharing the data associated with the article
- **New data sharing mandates by funding entities**
  - *Example:* National Institutes of Health (NIH) recent release of Policy for Data Management and Sharing
- **Joint statements from scientific communities**
  - *Example:* American Geophysical Union (AGU) Position Statement on Data
- **Ubiquity of domain-specific and generalist data repositories**
  - *Example:* Dataverse software powers > 60 repositories world-wide

Data Policies of top 50 journals in 6 disciplines

No Policy

Encourage data sharing

Require data sharing



Crosas, Gautier, Karcher, Kirilova, Otalora, Schwartz. Data Policies of Highly-Ranked Social Science Journals, *preprint,* https://osf.io/preprints/socarxiv/9h7ay

Clear progress in data sharing and access

- **New data policies in journals**
  - *Example:* > 50% of top social science journals recommend or require sharing the data associated with the article
- **New data sharing mandates by funding entities**
  - *Example:* National Institutes of Health (NIH) recent release of Policy for Data Management and Sharing
- **Joint statements from scientific communities**
  - *Example:* American Geophysical Union (AGU) Position Statement on Data
- **Ubiquity of domain-specific and generalist data repositories**
  - *Example:* Dataverse software powers > 60 repositories world-wide

Final NIH Policy for Data Management and Sharing

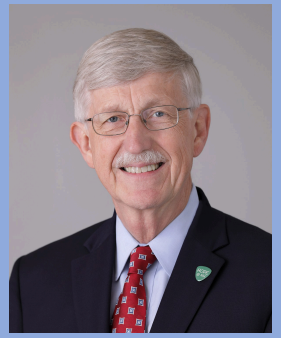
Notice Number:
NOT-OD-21-013

Key Dates

Release Date:
Effective Date:
October 29, 2020
January 25, 2023

Issued by

Office of The Director, National Institutes of Health (OD)



"This policy establishes the baseline expectation that data sharing is a fundamental component of the research process"

Francis S. Collins, M.D., Ph.D.
Director, National Institutes of Health

"[…] NIH encourages data management and sharing practices to be consistent with the **FAIR (Findable, Accessible, Interoperable, and Reusable)** data principles and reflective of practices within specific research communities."

Clear progress in data sharing and access

- **New data policies in journals**
  - *Example:* > 50% of top social science journals recommend or require sharing the data associated with the article
- **New data sharing mandates by funding entities**
  - *Example:* National Institutes of Health (NIH) recent release of Policy for Data Management and Sharing
- **Joint statements from scientific communities**
  - *Example:* American Geophysical Union (AGU) Position Statement on Data
- **Ubiquity of domain-specific and generalist data repositories**
  - *Example:* Dataverse software powers > 60 repositories world-wide

AGU ADVANCING EARTH AND SPACE SCIENCE

JOIN RENEW GIVE LOGIN

“Robust, verifiable, and reproducible science requires that evidence behind an assertion be accessible for evaluation. Researchers have a responsibility to collect, develop, and **share this evidence** in an ethical manner, that is **as open and transparent as possible**.”

Clear progress in data sharing and access

- **New data policies in journals**
  - *Example:* > 50% of top social science journals recommend or require sharing the data associated with the article
- **New data sharing mandates by funding entities**
  - *Example:* National Institutes of Health (NIH) recent release of Policy for Data Management and Sharing
- **Joint statements from scientific communities**
  - *Example:* American Geophysical Union (AGU) Position Statement on Data
- **Ubiquity of domain-specific and generalist data repositories**
  - *Example:* Dataverse software platform powers > 60 repositories worldwide

Dataverse®

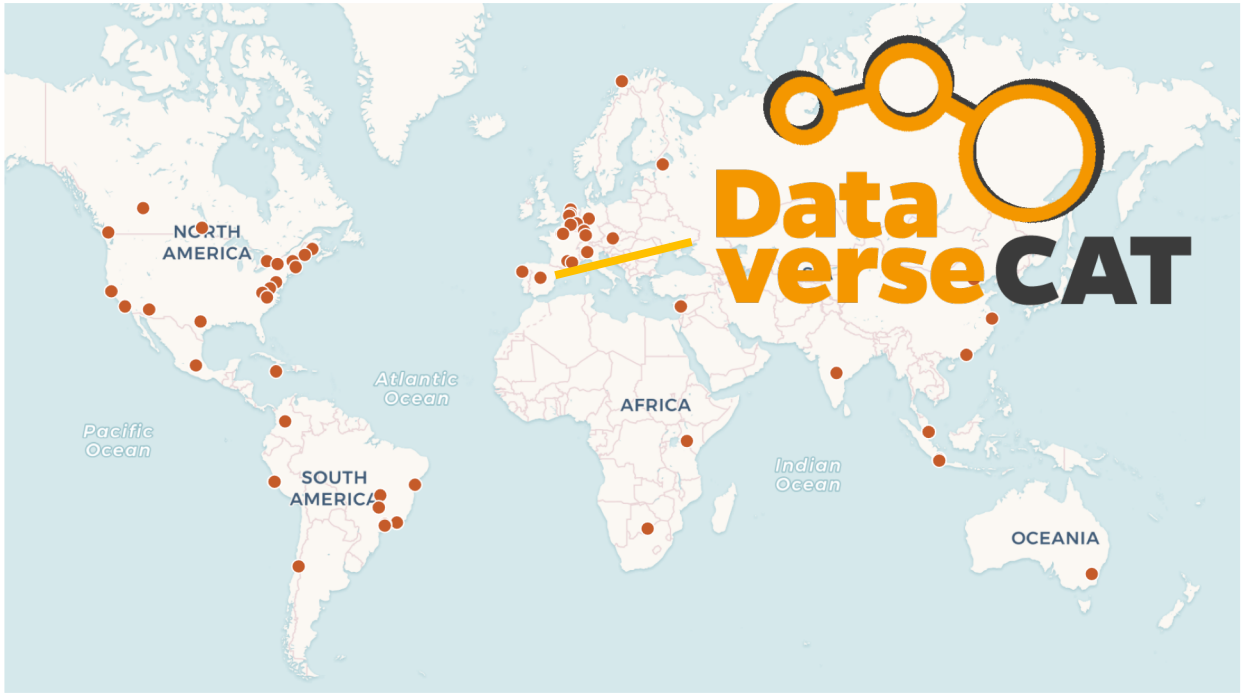
Federated FAIR data repositories worldwide

- **Open-source**
- **63** installations
- **6** continents
- **7K** dataverses
- **135K** datasets
- **800K** files
- **28M** file downloads
- **Metadata** shared across repositories

Developed at Harvard's Institute for Quantitative Social Science (IQSS) with contributions from the Dataverse community (https://dataverse.org)

Dataverse®

Federated FAIR data repositories worldwide



- **Open-source**
- **63** installations
- **6** continents
- **7K** dataverses
- **135K** datasets
- **800K** files
- **28M** file downloads
- **Metadata** shared across repositories

Developed at Harvard's Institute for Quantitative Social Science (IQSS) with contributions from the Dataverse community (https://dataverse.org)

Challenges

Challenges remain

Context, documentation, provenance

Collaborations

Large and complex datasets

Sensitive and proprietary data

- Insufficient information to reuse the data
- Incomplete code to reproduce results
- Lack of data source and transformations to understand validity

Challenges remain

Context, documentation, provenance

Collaborations

Large and complex datasets

Sensitive and proprietary data

- Difficult to find research datasets before publication
- Difficult to access data from other groups or organizations
- Often duplicative, costly efforts

Challenges remain

Context, documentation, provenance

Collaborations

Large and complex datasets

Sensitive and proprietary data

- Data cannot be downloaded to local computer
- Often special software is required to explore and make sense of the data

Challenges remain

Context, documentation, provenance

Collaborations

Large and complex datasets

Sensitive and proprietary data

- Not all data can be open
- Security and access requirements depend on data sensitivity
- Difficult to negotiate Data Use Agreements
- Access to proprietary data for research very limited

Vision

A Data Commons

"... brings together (or co-locates) **data with cloud computing** infrastructure and commonly used **software services, tools & applications** for **managing, analyzing, and sharing data** to create an **interoperable** resource for a research community."

[Robert Grossman, on the NIH Data Commons Consortium initiative]

A Data Commons vision with Dataverse

Context, documentation, provenance

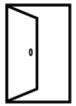
Collaborations

Large and complex datasets

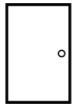
Sensitive and proprietary data

Data Lifecycle

1. Data Collection



Open data (gov, cities)



Data Use Agreements

Private, sensitive data (companies, hospitals)

Data collected for research (experiments, observations)

2. Active Research

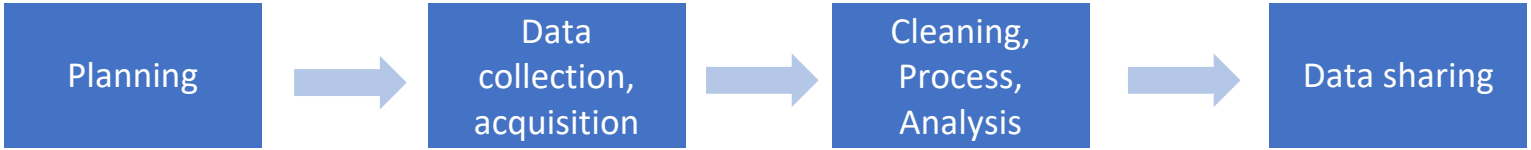
Research computing, software, methods workflows

3. Data Sharing

Data repository

Reuse and reproducibility

Integration with containers, packaging standards



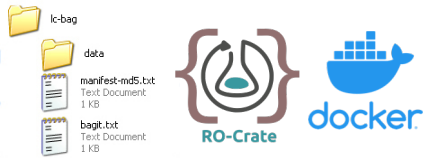
Data Management Plan (DMP)

Data Use Agreement (DUA)

Data
Code
Workflows
Notebooks
Codebooks

Dataverse®

Package with Metadata & Provenance



Containers (Docker): Encapsulate data + code + environment for **computational reproducibility** and be **ready for analysis**

Packaging Standards (RDA Bags , Research Objects-Crate): Package data with associated files, metadata, and provenance for **sharing across systems**

A Data Commons vision with Dataverse

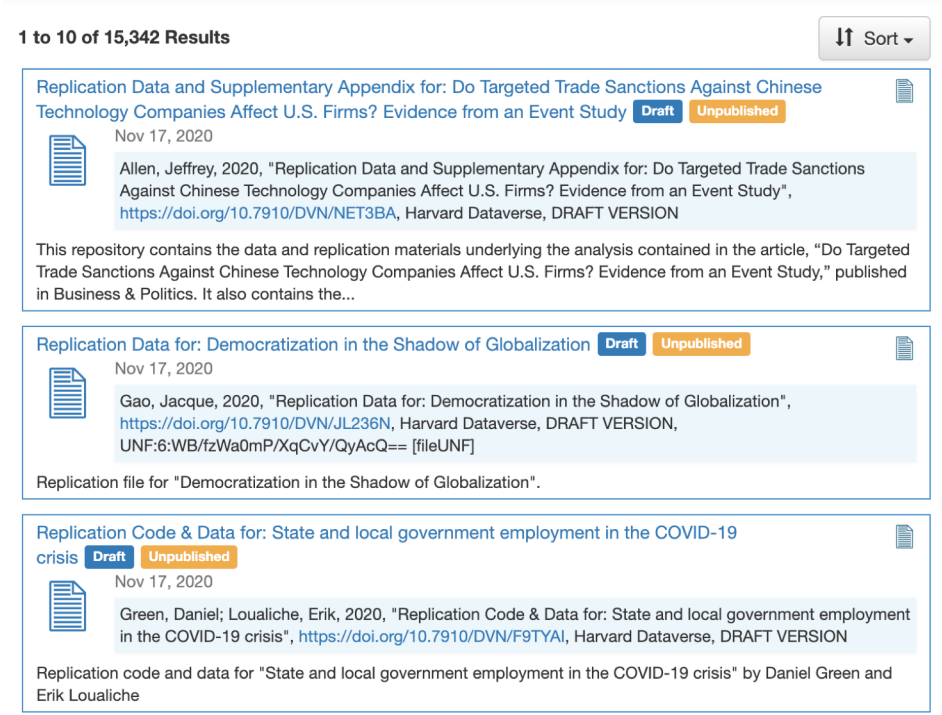
Context, documentation, provenance

Collaborations

Large and complex datasets

Sensitive and proprietary data

A common registry for active research datasets



- Adding to Dataverse a **metadata catalog** for unpublished datasets and datasets published elsewhere
- **Permissions** to access data granted to **collaborators**
- **Standard authentication** mechanism to facilitate access

A Data Commons vision with Dataverse

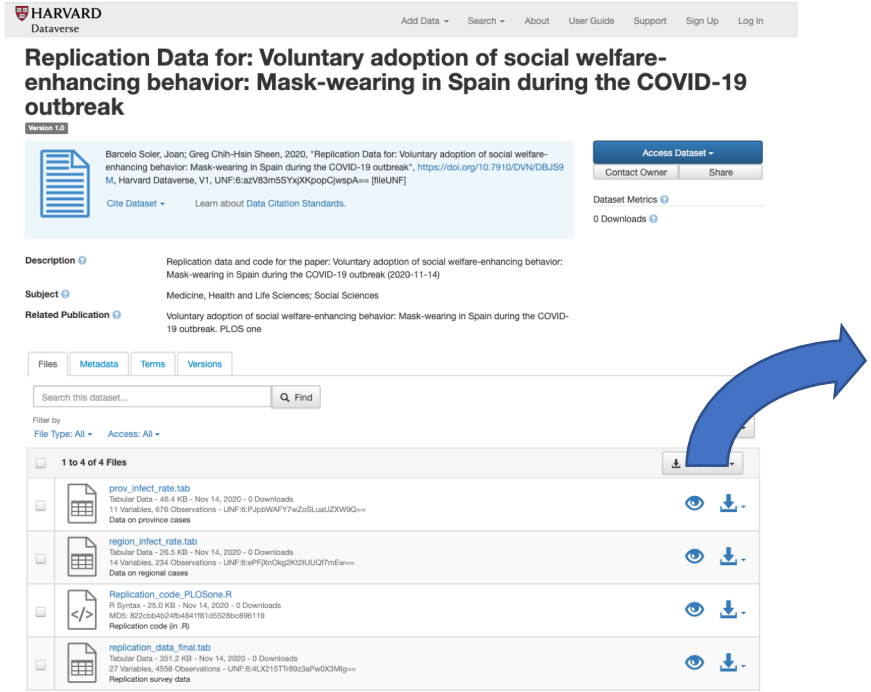
Context, documentation, provenance

Collaborations

Large and complex datasets

Sensitive and proprietary data

Co-location of data and computing on the cloud



Enable access to data on the cloud, with software needed for analysis

Massachusetts Green High Performance Computing Center + New England Research Cloud + Northeast Storage Exchange on OpenStack open-source cloud

Dataverse is also being integrated with the European Open Science Cloud.

A Data Commons vision with Dataverse

Context, documentation, provenance

Collaborations

Large and complex datasets

Sensitive and proprietary data

Data classification for security and access requirements

Non-Sensitive DataTags

Blue	Publicly open, no barriers
Green	Publicly open, but need to register to access
Yellow	Restricted, need to be granted permissions, but non-sensitive

Sensitive DataTags

Orange	Requires Data Use Agreement (DUA); requires data enclave ***(moderate sensitivity)***
Red	Requires DUA; stricter security requirements and audits ***(high sensitivity)***
Crimson	Only metadata and no link to data; data stored outside network ***(maximum sensitivity)***

Sweeney, Crosas, Bar-Sinai, 2015, Sharing Sensitive Data with Confidence: the DataTags System
https://techscience.org/a/2015101601/

Openly findable data, secure computation and storage

Data owners keep control of their data

Public Repository

Dataverse®

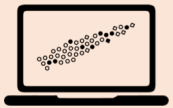
Blue | Green | Yellow

Private datasets discoverable via repository (only metadata open)

Notary Service

Digitally signed attestations confirming compliance with data access policies (DUA)

Approved Secure Compute Environment



Orange | Red

Trusted Remote Storage Agents (TRSA) or data enclaves

Orange | Red

impact

Impact: Infrastructure for Privacy-Assured Computation https://cyberimpact.us/

Differential Privacy tools to explore sensitive data

- A **differentially private** algorithm adds a minimum amount of noise to its output to mathematically guarantee the privacy of any individual in the dataset.
- **OpenDP** is a community effort to build a trustworthy and open-source suite of **differential privacy tools** to explore sensitive data
- We are currently working on the first release of **OpenDP and Dataverse integration**

What will this mean:

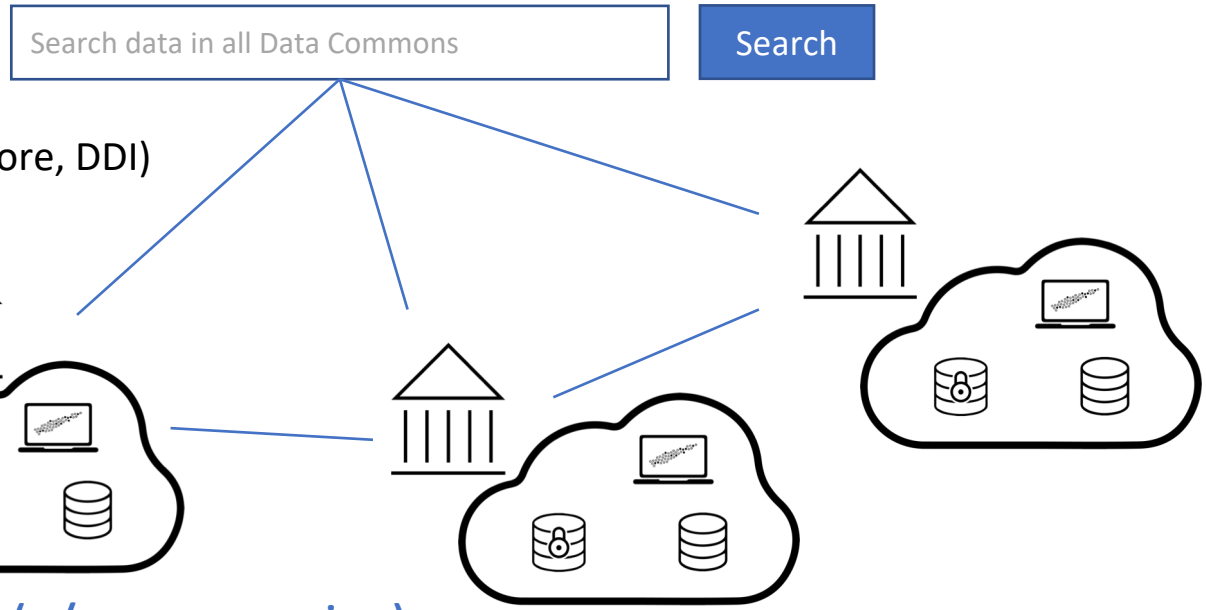
- Sensitive datasets findable in Dataverse will be explorable through **differentially private statistics**, without ever accessing the original dataset
- Opens up sensitive data to the research community

OpenDP

https://opendp.io

A global vision:
A Federated Data Commons

Agreement on community standards needed for a federated Data Commons



Common services:
Metadata (Schema.org, Dublin Core, DDI)
Packaging (RDA Bags, RO-Crate)
Persistent Identifiers (DOI)
Authentication, APIs

FAIR repository (w/ common services)
integrated with **cloud infrastructure**
and **applications and tools**

Summary

The proposed Data Commons **lowers the barrier** to:

- Finding active research data, in addition to data already publicly published
- Accessing and tracking the data in one place for collaboration
- Analyzing published results in a simple step
- Distributing data across systems in a standardized form
- Sharing private data for research securely between industry and academia

A Commons will allow researchers to explore the chain of data and computations behind a discovery as easily as they currently explore the chain of papers via citations – but with all the tools needed to immediately take the next step.

“Since the novel coronavirus struck, scientific research has been shared, and built upon, at an unprecedented pace. An **open and deeply collaborative academic enterprise has emerged**, with **scientists from around the world sharing data** and working together […]

we must not revert to our old ways.”

Janet Napolitano, President, University of California

https://www.insidehighered.com/views/2020/07/31/universities-should-commit-opening-their-research-everyone-opinion

Via Heather Joseph (SPARC)

Gràcies

Mercè Crosas, Ph.D., Harvard University
scholar.harvard.edu/mercecrosas @mercecrosas