

Enhancing collaboration and access to data throughout the research lifecycle

Center for Open and REproducible Science (CORES) Launch Event
Stanford | Data Science
February 18, 2021

Mercè Crosas, Ph.D., Harvard University
University Research Data Management Officer, HUIT
Chief Data Science and Technology Officer, IQSS
scholar.harvard.edu/mercecrosas @mercecrosas



The Institute for Quantitative Social Science



HARVARD
UNIVERSITY

Photo: Dwayne Liburd
<https://www.dwayneliburd.com/>

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](#)

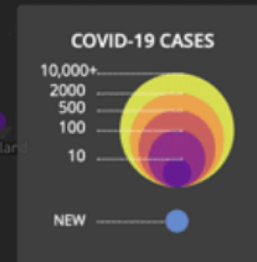
Open access epidemiological data from the COVID-19 outbreak

[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](https://doi.org/10.1016/S1473-3099(20)30119-5)



WEEK OF 2020-03-16



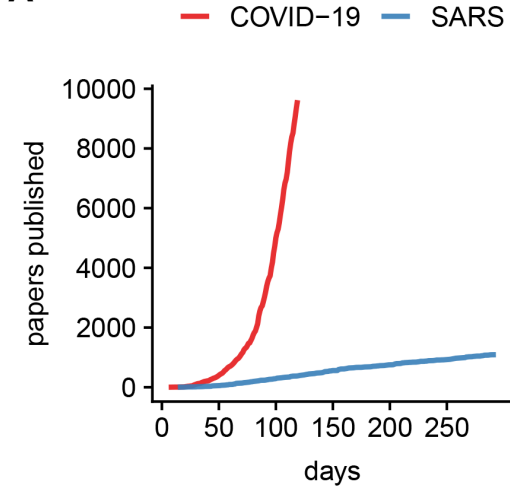
Science in the face of Covid-19: faster, better, stronger?

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

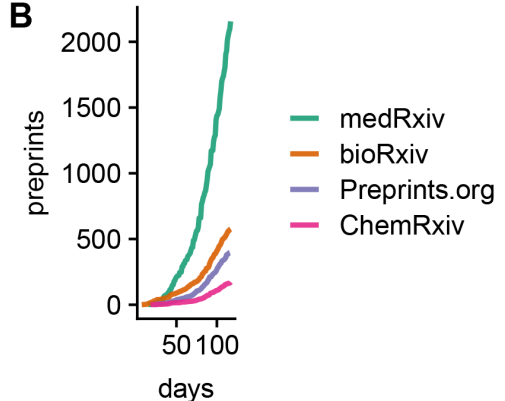


“During the pandemic, conflicting information can undermine trust in science. Openness and the **sharing of data** enable researchers to **collaborate, review and reproduce findings**, and such activities strengthen trust in science.”

A



B



“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)



- **Increase of sharing and access to data globally**
- **Increase of research data and computing service offerings in Universities**

Progress

Recent Advances in Data Sharing

- **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**

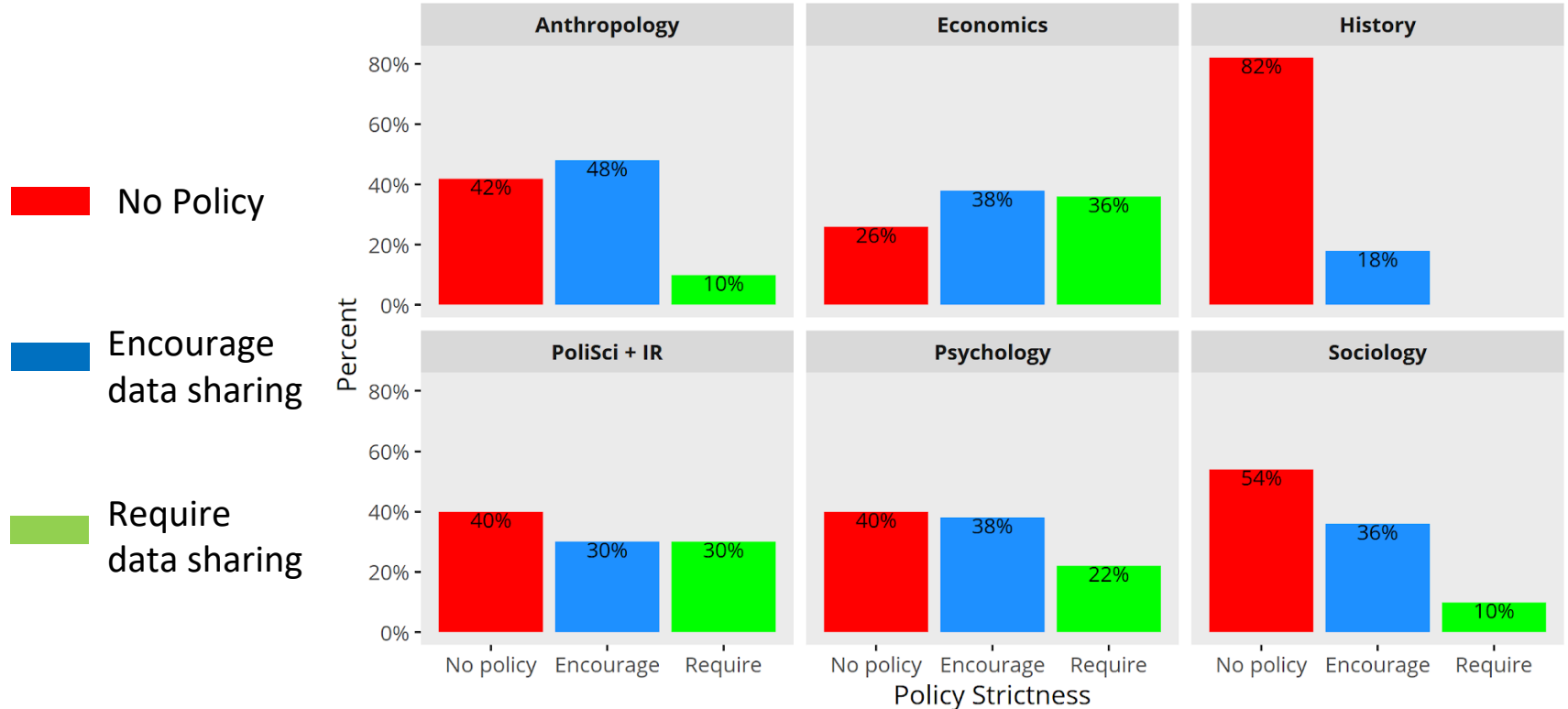
- *Example:* American Geophysical Union (AGU) Position Statement on Data

- **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

Percentage of Journals by Strictness of Data Policy



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

Recent Advances in Data Sharing

- **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](#))



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

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

“[...] 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.”

Recent Advances in Data Sharing

- **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



POSITION STATEMENT ON DATA

“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.**”

Recent Advances in Data Sharing

- **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



Federated **FAIR** data repositories worldwide

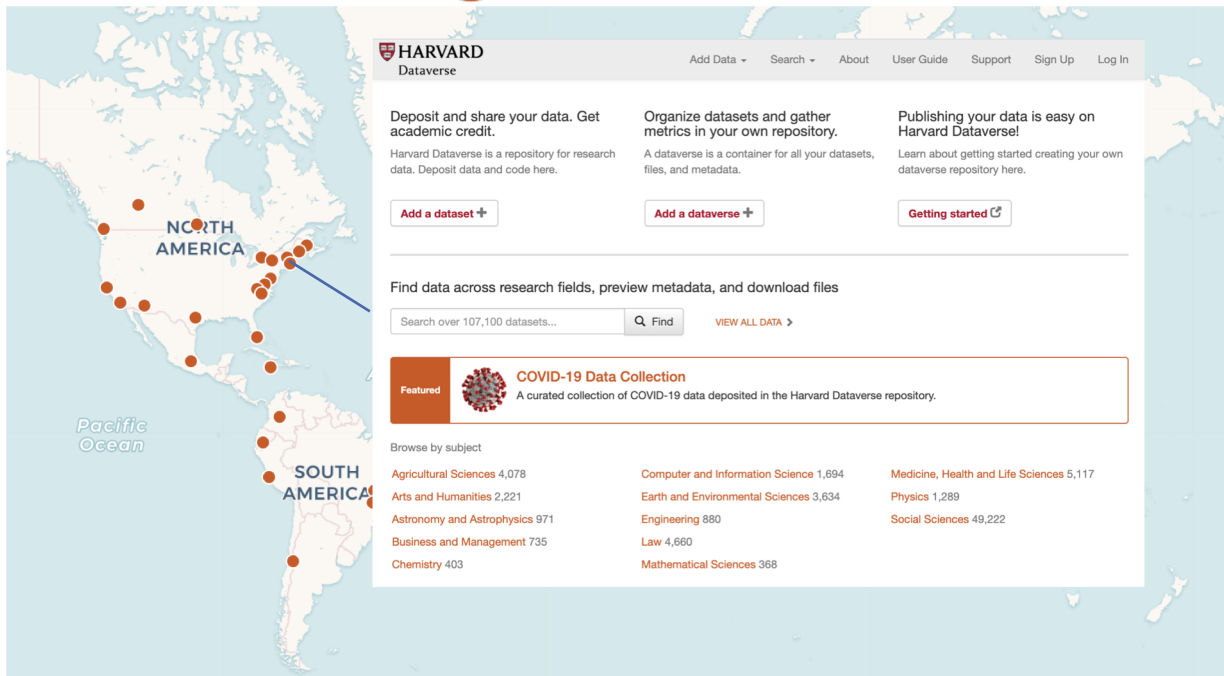


Developed at Harvard's Institute for Quantitative Social Science (IQSS) with contributions from the Dataverse community and the Global Dataverse Community Consortium (<https://dataverse.org>)

- **Open-source**
- **67** installations
- **6** continents
- **8K** Dataverse collections
- **140K** datasets
- **900K** files
- **30M** file downloads
- **Metadata** shared across repositories



Harvard Dataverse repository

A screenshot of the Harvard Dataverse website. The background is a light blue map of North and South America with orange dots indicating data locations. The website interface is white with a grey navigation bar at the top. The navigation bar includes the Harvard Dataverse logo, a search bar, and links for "Add Data", "Search", "About", "User Guide", "Support", "Sign Up", and "Log In". Below the navigation bar, there are three columns of text. The first column is titled "Deposit and share your data. Get academic credit." and includes a link to "Add a dataset". The second column is titled "Organize datasets and gather metrics in your own repository." and includes a link to "Add a dataverse". The third column is titled "Publishing your data is easy on Harvard Dataverse!" and includes a link to "Getting started". Below these columns, there is a section titled "Find data across research fields, preview metadata, and download files" with a search bar and a "VIEW ALL DATA" link. At the bottom, there is a "Featured" section titled "COVID-19 Data Collection" with a description. Below this, there is a "Browse by subject" section with a table of subjects and their counts.

HARVARD
Dataverse

Add Data Search About User Guide Support Sign Up Log In


Deposit and share your data. Get academic credit.
Harvard Dataverse is a repository for research data. Deposit data and code here.
[Add a dataset](#)

Organize datasets and gather metrics in your own repository.
A dataverse is a container for all your datasets, files, and metadata.
[Add a dataverse](#)

Publishing your data is easy on Harvard Dataverse!
Learn about getting started creating your own dataverse repository here.
[Getting started](#)

Find data across research fields, preview metadata, and download files

Search over 107,100 datasets... [Find](#) [VIEW ALL DATA](#)

Featured  **COVID-19 Data Collection**
A curated collection of COVID-19 data deposited in the Harvard Dataverse repository.

Browse by subject

Agricultural Sciences 4,078	Computer and Information Science 1,694	Medicine, Health and Life Sciences 5,117
Arts and Humanities 2,221	Earth and Environmental Sciences 3,634	Physics 1,289
Astronomy and Astrophysics 971	Engineering 880	Social Sciences 49,222
Business and Management 735	Law 4,660	
Chemistry 403	Mathematical Sciences 368	

Open to all researchers
across all disciplines



- Increase of sharing and access to data globally
- Increase of research data and computing service offerings in Universities

Progress

Growth of Research Data Services Offerings in Universities

- Ithaka S+R report (Radecki & Springer, 2020, <https://doi.org/10.18665/sr.314397>):
 - Reviewed research data services from 120 U.S. Universities
- A growing number of research data services distributed across various university units:

Within Libraries and IT (main providers)

- Consulting
- Training events
- Backend work (data architecture, metadata design)
- Front end work (web development, data visualizations)

Outside Libraries and IT

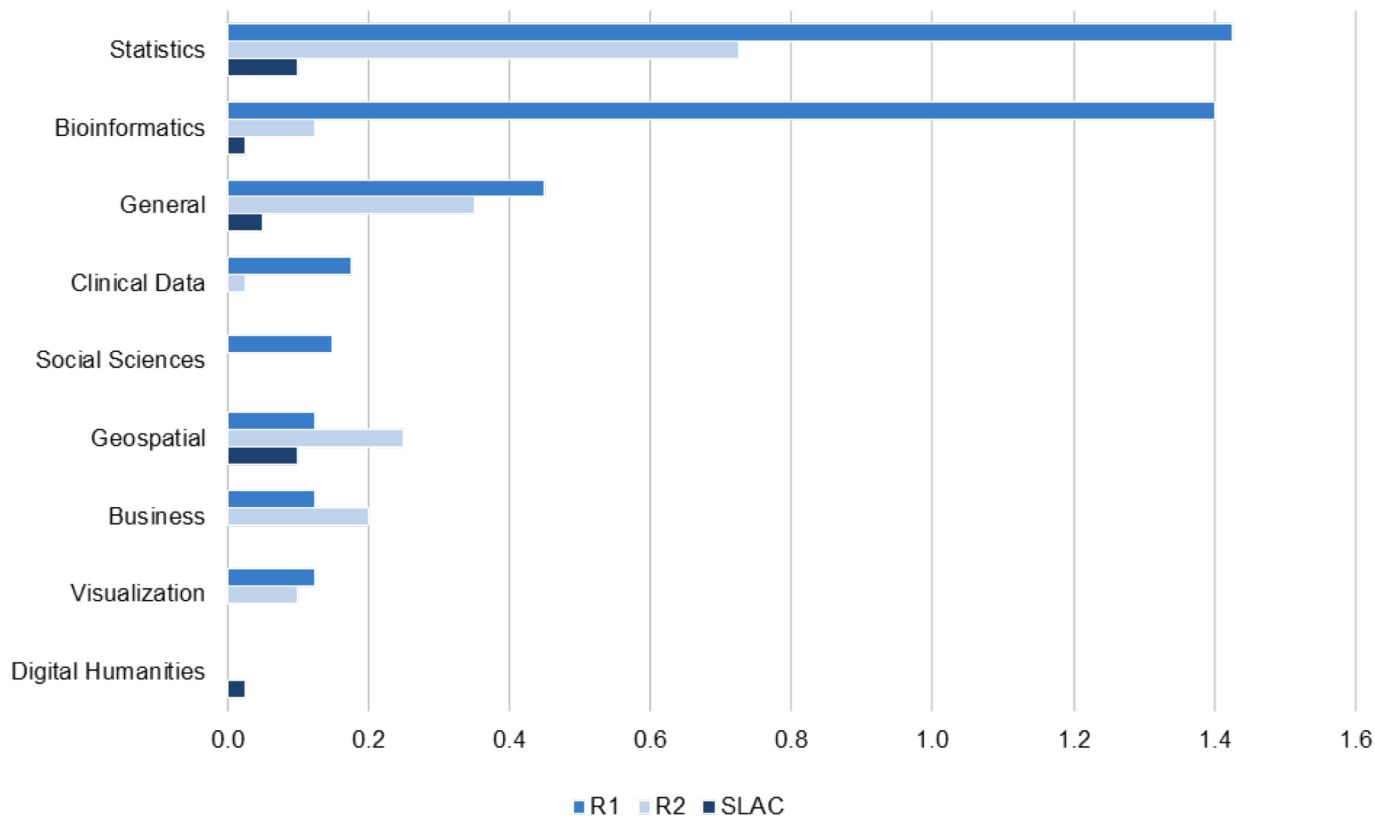
- Statistics
- Bioinformatics
- Geospatial
- Clinical data
- Business
- Social Science
- Visualizations

Profile of types of Library Data Services

Generalist consultation is the most common service offered by the libraries

	Consulting	Training Events	Front End Work	Back End Work	Total
General	35.9%	16.0%	3.2%	2.6%	57.7%
Geospatial	16.7%	9.0%	0.0%	0.0%	25.6%
Statistics	7.1%	1.3%	0.0%	0.0%	8.3%
Digital Humanities	2.6%	1.3%	0.0%	0.0%	3.8%
Social Sciences	0.6%	0.6%	0.0%	0.0%	1.3%
Health Sciences	0.6%	0.0%	0.0%	0.0%	0.6%
Other	1.3%	1.3%	0.0%	0.0%	2.6%
Total	64.7%	29.5%	3.2%	2.6%	100%

Average number of research data services per institution offered by centers and facilities, departments, and schools



Research Data and Computing Services Offerings at Harvard

- **A need to increase** research data and computing services
 - Along with increase in data-centric and data science research
 - To support funders and journals requirements
 - Services distributes across units and schools
- **Collaboration** between Research, Library, and IT/Research Computing is key
- **Build an inventory** to learn what services are provided across units
 - Understand what is available
 - Standardize the information
- **One research support site** to find all service offerings in a common way *(to be launched in 2021)*
- **Find gaps and connect** services, tools, and teams
- **Foster a community** of research computing and data teams at the University (working groups, events)

Services offerings throughout the **research lifecycle**

Research Lifecycle



The research lifecycle refers to the (often iterative) process of conducting research, from the initial planning, funding, and research project design to publishing and disseminating the conclusions or work of scholarship. Although the research process varies across disciplines and research domains, it often includes validating a model or hypothesis by using information and data. In turn, the results from the data help improve the model and thus, gather additional data to validate the new model. On this site, we refer to data in the broadest sense of the word, including experimental, observational, acquired, and simulated data, as well as any relevant information, artifacts, and original sources. In recent years, the research lifecycle has also included publishing

the data, code, and workflows to facilitate the reproducibility of the published results.

Browse by Research Lifecycle

Planning →

[Buying and Licensing Data](#)

[Data Retrieval](#)

Active Research →

[Cluster Computing](#)

[Data Cleaning](#)

Dissemination & Preservation →

[Archiving Faculty Research Data and Archiving Data](#)

Planning:

Access & Reuse
Plan & Design

Active Research:

Collect & Create
Analyze & Collaborate

Dissemination & Preservation:

Evaluate & Archive
Share & Disseminate

Planning



Research planning concerns all aspects of preparing for a research project. It includes seeking funding, awareness of University and sponsor requirements, and the organization of data, records, tools, and/or resources needed to conduct the research and disseminate and archive valuable results.

Animal Research Resources →

The University has established a number of useful resources to support animal research...

Buying and Licensing Data →

Consultations and instruction associated with obtaining, buying, and licensing research data...

Data Retrieval →

Consultation on how to acquire free data or retrieve data provided by a source (e.g. Library subscriptions...

Data Safety & Regulated Data →

The University's researchers and administrators are responsible for properly managing and securing research data...

Data Use Agreement Processing →

The transfer of data between organizations is common in the research community...

Finding Data →

Consultation, full service (HLS, Baker), and referrals for locating sources of research data (e.g. Library subscriptions, government sponsor, repository).

Human Subjects Research Resources →

The University has established a number of useful resources to support human research...

Longwood Health Informationist →

Some researchers may wish to embed a data services librarian as a health informationist in their projects...

Pre- & Post-Award Resources →

Resources and systems for research administrators, compliance officers, and researchers to support the University's research enterprise...

Research Data Management Lifecycle →

Consultation and support for Research Data Management lifecycle activities...

Research Design →

Full support and consultations on the design of research projects to streamline the research process...

Training, Workshops & Capacity Building →

Ongoing training and workshops are available across the University online, and in person when available...

Planning: Access & Reuse Plan & Design

12 service offerings:

- Buying and Licensing Data
- Data Retrieval, Finding Data
- Data Safety and Regulated Data
- Data Use Agreement Processing
- Human Subjects
- Animal Research Resources
- Pre- & Post-Award Resources
- Research Data Management Lifecycle
- Research Design
- Training, Workshop, Capacity Building
- Project Health Informationist

SERVICES

▼ Research Administration & Compliance

Data Safety &
Regulated Data

Data Use Agreement
Processing

eSupport - Committee
on Microbiological
Safety (eCOMS)

Human Subjects and
Animal Research
Resources

Pre- & Post-Award
Resources

• Research Computing

• Research Data and Scholarship

[HOME](#) / [SERVICES](#) / [RESEARCH ADMINISTRATION & COMPLIANCE](#) /

Data Use Agreement Processing

The transfer of data between organizations is common in the research community. When the data is confidential, proprietary, or otherwise considered sensitive or protected, the organization providing the data, whether that is Harvard or a third party, may require a Data Use Agreement (DUA) to govern the exchange of data. This section includes guidance on the DUA-Agreements Application which supports the review, approval and management process for DUAs, and other related resources.

Data Use Agreement review and compliance application for researchers at Harvard interested in requesting data from a third party, or providing data to a third party. This application supports the review, approval and management process for Data Use Agreements.

A DUA is a binding contract governing access to and treatment of nonpublic data provided by one party (a "Provider") to another party (a "Recipient"). DUAs are often required by external parties, and may also be necessary for Harvard data to be disclosed to another organization. DUA terms and conditions vary depending on the laws and regulations governing the specific type of data to be shared, as well as the policies and/or requirements of the Provider and Recipient. If you are unsure whether a DUA is necessary, feel free to reach out to your sponsored research office.

Details by Provider

- HUIT Administrative Technology Services, Research Administration and Compliance
- Harvard Medical and Dental Schools
- Harvard T. H. Chan School of Public Health
- Harvard University Area

Example of Service offering in Planning Phase:

- DUA and Safety System to:
 - Track all DUAs for incoming and outgoing compliant data
 - Manage DUA while data are used for research
- Assistance with DUA negotiation
- Connect process with IRB and Security officers

Active Research



The active research phase of a project may include collecting or acquiring data, information, or sources, conducting quantitative or qualitative analysis, and/or using computation resources, data storage, quantitative or qualitative tools, visualizations, or information exploration.

Cluster Computing →

Doing computations at scale allows a researcher to test many different variables at once, thereby shorter time to outcomes, and also provides the ability to ask...

Data Cleaning →

Data Cleaning services and consultation support for cleaning, reformatting, merging, and scraping data for analyzing, visualization and reporting.

Data Curation →

Specialists throughout Harvard Library are available to consult about data curation, organization, and integration. In order to maintain the availability...

Data Handling →

Consultation, instruction, and support for practices and procedures involving data (e.g., reformatting).

Data Science and Research Computing Facilitation →

A research team can often benefit from incremental help to expand their knowledge and skills, to augment their collective skill set...

Data Science and Research Software Engineering Collaboration →

Data Science and Software Engineering play an important role in research by creating new...

Data Security →

Consultations and/or instruction on ensuring data security during the research lifecycle, including compliance with University policies.

Data Visualization →

Data visualization creation and support (i.e. specialized referrals) for research projects.

Database →

In a data analysis environment, organized collections of data need to be hosted and access granted to set of researchers. A database service provides an interface to...

Dataset Creation →

Across the University, experts are available to consult on creating data and datasets using tools like mturk, qualtrics, and other surveys and field experiments...

Geospatial Data →

Experts are available to consult with researchers on finding, preparing, creating, and/or analyzing geospatial data...

Lab and Biological Safety Resources →

University-wide tracking of lab safety.

Active Research: Collect & Create Analyze & Collaborate

19 service offerings

- Cluster Computing, Virtual Instances
- Research Data Storage, Database, Security
- Software and Platforms
- Electronic Lab Notebooks, Computational Notebooks
- Research Computing Consulting & Facilitation
- Data Science and Research Software Engineering, Statistical Analysis, Text Analysis
- Dataset Creation, Data Cleaning, Data Curation, Data Handling, Metadata creation
- Data Visualization; Geospatial data
- Qualitative Data Support
- Lab and biological Safety

SERVICES
▸ Research Administration & Compliance
▸ Research Computing
Cluster Computing
Data Science and Research Software Engineering Collaboration
Database
Research Computing Consulting and Facilitation
Research Data Storage
Virtual Instances
▸ Research Data and Scholarship

[HOME](#) / [SERVICES](#) / [RESEARCH COMPUTING](#) /

Data Science and Research Software Engineering Collaboration

Data Science and Software Engineering play an important role in research by creating new capabilities to process and analyze data, helping ensure reproducibility, and aiding researchers in extracting knowledge and insight for the data. The term software here is used broadly to include all the ways in which one creates and analyses data. Researchers utilize software in their research by using scripts, tools, open-source software, and licensed software. Data science also covers a wide range of skills and techniques applied to cleaning (aka wrangling), processing, and statistics that are typically beyond what a researcher from a specific domain might have. Due to the rapidly evolving nature of research, there are not always codes for all functions needed, nor are their clean data sources; therefore, the software or data pipelines are developed specifically for a given project. Traditionally, this development was done with researchers (graduate students and postdocs) or independent contractors. This approach poses several issues in terms of maintenance, optimization, reproducibility, and cost. RSE or Data Scientist team can work closely with other Research Computing Systems teams to design, develop, deploy, optimize, and maintain software packages/tools and data pipelines that are paired with specific hardware architectures to accelerate cutting-edge research at Harvard University.

Details by Provider

- Faculty of Arts and Sciences, Research Computing
- Institute for Quantitative Social Sciences
- Harvard Business School

Example of Service offering in Active Research phase:

Data Science Services offered by the Institute for Quantitative Social Science (IQSS)

- Focuses on **social science** support, but includes other scientific domains
- **Consulting:** short term
- **Collaboration:** longer project (fee)
- **Training materials** (in collaboration with Harvard Business School):
 - **Python:** Introduction, web scraping
 - **R:** Introduction, regressions models, graphics, data wrangling
 - **Stata:** Introduction, data management, regression models, graphics
 - **Other:** Introduction to programming, SAS introduction, data science tools

Dissemination & Preservation



Dissemination and preservation are increasingly important parts of the research lifecycle. Sponsors, journals, and publications often require that all inputs, outputs, how research was conducted, and what tools, data, and code were used be available and accessible, alongside results and conclusions.

List of resources for dissemination and preservation below:

Archiving Faculty Research Data and Archiving Data →

Full service options, consultation, and instruction for faculty who need to archive their research data...

Copyright and Intellectual Property →

Consultations and/or instruction on a wide variety of topics relating to copyright and intellectual property concerns...

DASH Open-Access Repository →

DASH is Harvard's central, open-access repository for research by Harvard community members...

Data Sharing and Publishing →

Harvard offers consultation and instruction for researchers looking to publicly share their data and research products...

Harvard Dataverse Repository →

Harvard Dataverse is a free, self-service data repository open to all researchers provided by any discipline, both inside...

Dissemination & Preservation: Evaluate & Archive Share & Disseminate

5 service offerings:

- Copyright and Intellectual Property
- Archiving data
- Data Sharing and Publishing
- DASH Open Access Repository
- Harvard Dataverse Repository

SERVICES
▸ Research Administration & Compliance
▸ Research Computing
▸ Research Data and Scholarship
Archiving Faculty Research Data and Archiving Data
Buying and Licensing Data
Copyright and Intellectual Property
Data Cleaning
Data Curation
Data Deposit
Data Handling
Data Retrieval
Data Security Support
Data Sharing and Publishing
Data Visualization
Dataset Creation
Finding Data
Geospatial Library, Data Analysis, Creation, Visualization

HOME / SERVICES / RESEARCH DATA AND SCHOLARSHIP /

Harvard Dataverse Curation

The Harvard Dataverse data curation team, staffed by member of IQSS and the Harvard Library (and separately, the Harvard Kennedy School Library), provides fee-based curation services and free consultations to researchers around the world who are depositing data into the Harvard Dataverse.

Research data replication datasets, data for related publications, and all file types and domains are welcomed in the Harvard Dataverse. Through this engagement, the curation services team will ensure that deposited datasets are discoverable, accessible, interoperable, and reusable (FAIR). (IQSS)

Details by Provider

● Institute for Quantitative Social Sciences

Audience

- All Affiliates
- All Faculty
- All Graduate Students
- All Undergraduate Students
- Public

Service Provider

Institute for Quantitative Social Sciences (IQSS)

Service Fee

Yes

Service Website

<https://support.dataverse.harvard.edu/curation-services>

Contact Information

support@dataverse.harvard.edu

Example of Service Offering in Dissemination phase:

Dataverse Curation services

- A **collaboration** between IQSS and the Harvard Library
- **Tiered service offerings:**
 - Free consultation (< 3 hours)
 - Extended consultation services
 - Dataverse collection set-up
 - administration and curation services
 - Custom services
- In 2021, new service for supporting “managed collections” interested in receiving **Core Trust Seal** certification.



Vision

what we are working on

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

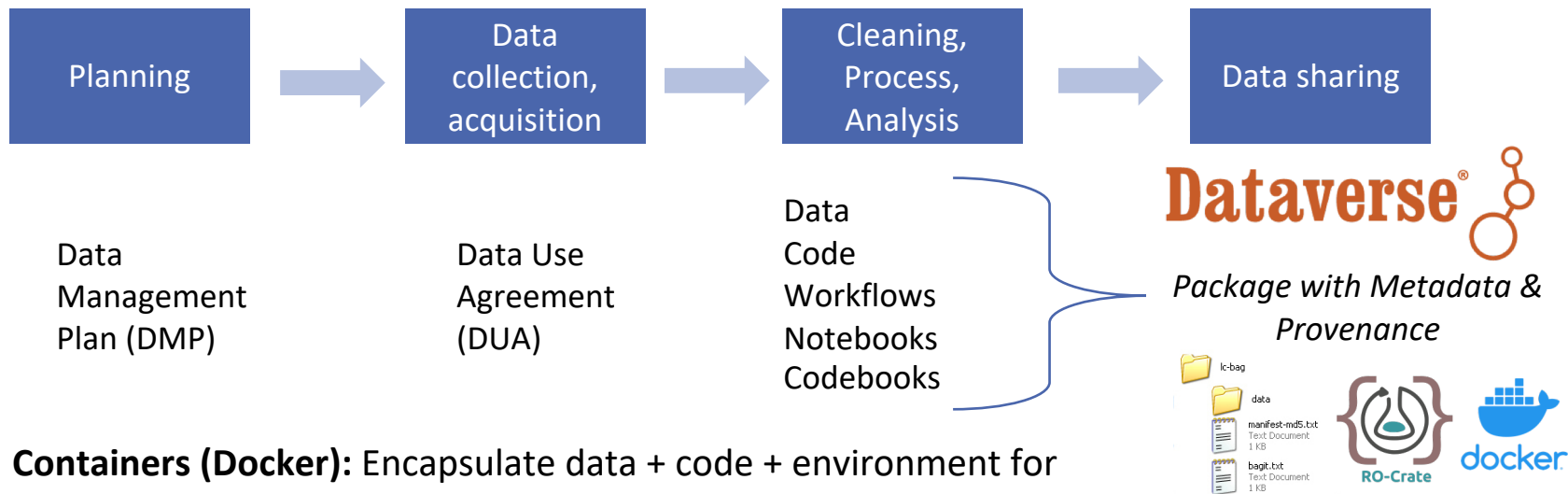
Large and complex datasets

Sensitive and proprietary data

Challenges:

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

Integration with containers, packaging standards



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

Challenges:

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

A common registry for active research datasets

Dataverse metadata catalog

1 to 10 of 15,342 Results

Sort

Replication Data and Supplementary Appendix for: Do Targeted Trade Sanctions Against Chinese Technology Companies Affect U.S. Firms? Evidence from an Event Study **Draft** **Unpublished**

Nov 17, 2020



Allen, Jeffrey, 2020, "Replication Data and Supplementary Appendix for: Do Targeted Trade Sanctions Against Chinese Technology Companies Affect U.S. Firms? Evidence from an Event Study", <https://doi.org/10.7910/DVN/NET3BA>, Harvard Dataverse, DRAFT VERSION

This repository contains the data and replication materials underlying the analysis contained in the article, "Do Targeted Trade Sanctions Against Chinese Technology Companies Affect U.S. Firms? Evidence from an Event Study," published in Business & Politics. It also contains the...

Replication Data for: Democratization in the Shadow of Globalization **Draft** **Unpublished**

Nov 17, 2020



Gao, Jacque, 2020, "Replication Data for: Democratization in the Shadow of Globalization", <https://doi.org/10.7910/DVN/JL236N>, Harvard Dataverse, DRAFT VERSION, UNF:6:WB/fzWa0mP/XqCvY/QyAcQ== [fileUNF]

Replication file for "Democratization in the Shadow of Globalization".

Replication Code & Data for: State and local government employment in the COVID-19 crisis **Draft** **Unpublished**

Nov 17, 2020



Green, Daniel; Loualiche, Erik, 2020, "Replication Code & Data for: State and local government employment in the COVID-19 crisis", <https://doi.org/10.7910/DVN/F9TYAI>, Harvard Dataverse, DRAFT VERSION

Replication code and data for "State and local government employment in the COVID-19 crisis" by Daniel Green and Erik Loualiche

- We are working on a **metadata catalog** for unpublished datasets or datasets published elsewhere (*Dataverse Metadata Working Group*)
- Metadata findable via the repository
- But data might be elsewhere:
 - Restricted w/ access to collaborators
 - Or in another repository

A Data Commons vision with **Dataverse**

Context, documentation, provenance

Collaborations

Large and complex datasets

Sensitive and proprietary data

Challenges:

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

Integration of data repositories with computing

Dataverse Repository

The screenshot shows the Harvard Dataverse interface. At the top, the Harvard Dataverse logo is on the left, and navigation links (Add Data, Search, About, User Guide, Support, Sign Up, Log In) are on the right. The main title is 'Replication Data for: Voluntary adoption of social welfare-enhancing behavior: Mask-wearing in Spain during the COVID-19 outbreak'. Below the title, it says 'Version 1.0'. A document icon is next to the title. To the right of the title, there is a blue button 'Access Dataset' with a dropdown arrow, and below it, 'Contact Owner' and 'Share' buttons. Further right, 'Dataset Metrics' shows '0 Downloads'. Below the title, there is a 'Description' section with the text: 'Replication data and code for the paper: Voluntary adoption of social welfare-enhancing behavior: Mask-wearing in Spain during the COVID-19 outbreak (2020-11-14)'. Below that is a 'Subject' section with the text: 'Medicine, Health and Life Sciences; Social Sciences'. Below that is a 'Related Publication' section with the text: 'Voluntary adoption of social welfare-enhancing behavior: Mask-wearing in Spain during the COVID-19 outbreak. PLOS one'. At the bottom, there is a 'Files' section with a search bar and a 'Find' button. Below the search bar, there are tabs for 'Files', 'Metadata', 'Terms', and 'Versions'. Below the tabs, there is a 'Filter by' section with 'File Type: All' and 'Access: All'. Below the filter section, there is a list of files. The first file is 'prov_infect_rate.tab' with a download icon. The second file is 'region_infect_rate.tab' with a download icon. The third file is 'Replication_code_PLOSone.R' with a download icon. The fourth file is 'replication_data_final.tab' with a download icon. A large blue arrow points from the 'Files' section of the Dataverse page to the right.

On-premise and cloud computing

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

A Data Commons vision with **Dataverse**

Context, documentation, provenance

Collaborations

Large and complex datasets

Sensitive and proprietary data

Challenges:

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

Data classification for security, access requirements

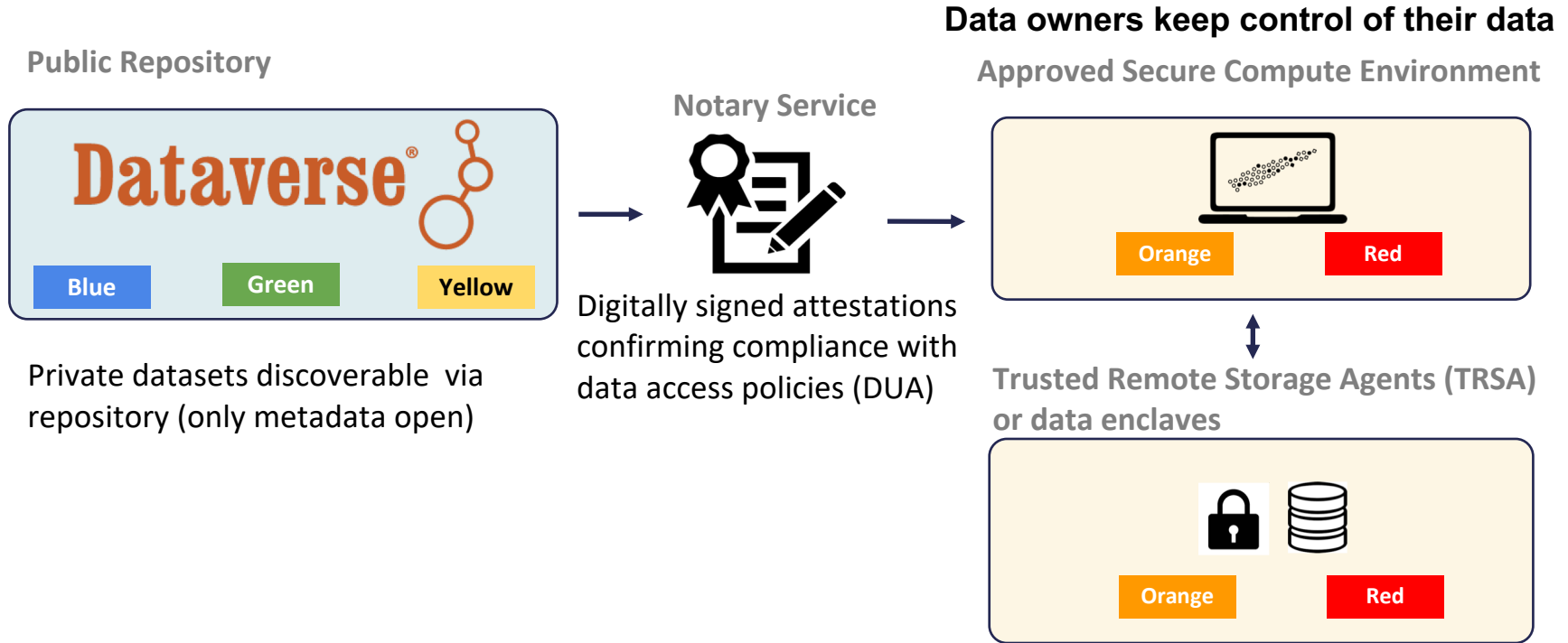
Non-Sensitive DataTags

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 <i>(moderate sensitivity)</i>
Red	Requires DUA; stricter security requirements and audits <i>(high sensitivity)</i>
Crimson	Only metadata and no link to data; data stored outside network <i>(maximum sensitivity)</i>

Openly findable data, secure computation and storage



Differential Privacy tools to explore sensitive data

- A **differentially private** algorithm introduces a minimum amount of noise to released statistics to mathematically guarantee the privacy of any individual in the dataset
- **OpenDP** (<https://opendp.org>) 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:

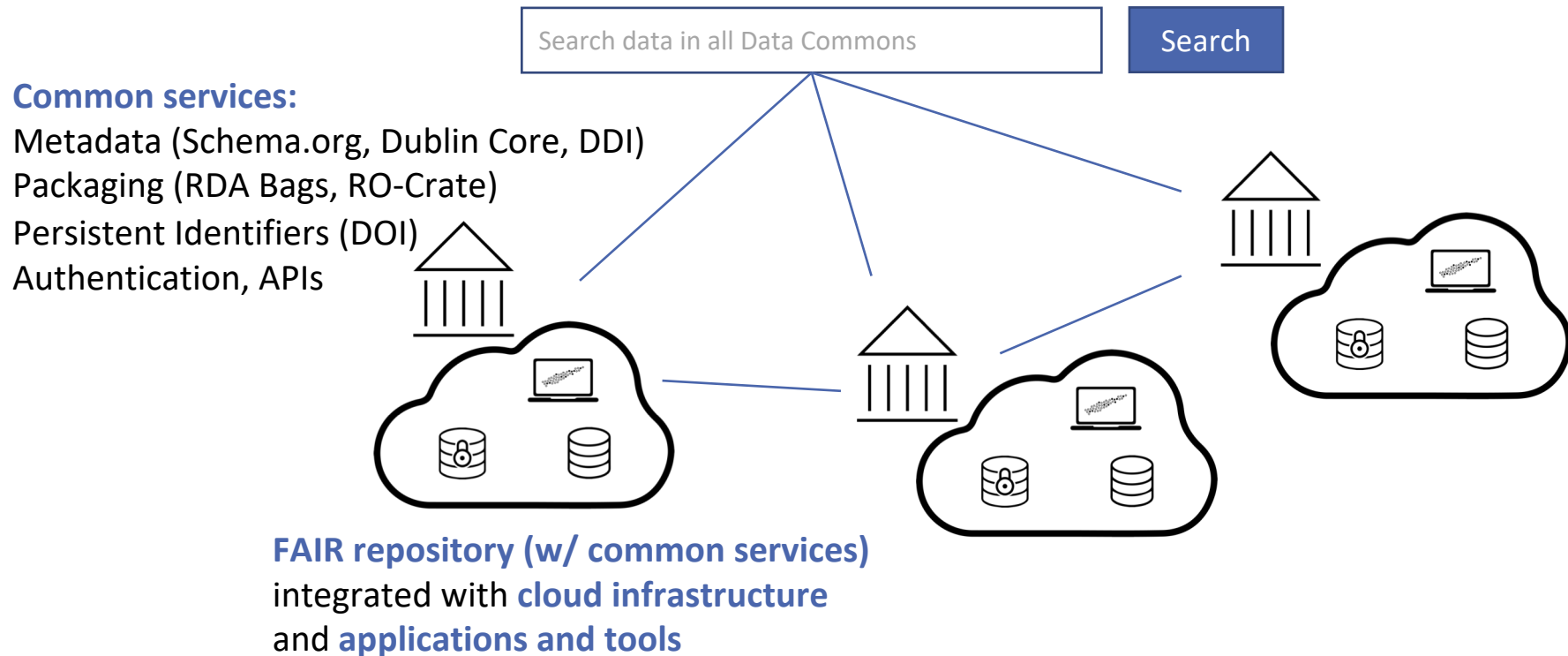
- Opens up sensitive data to the research community
- Sensitive datasets findable in a Dataverse repository will be explorable through **differentially private statistics**, without ever accessing the original dataset
- Statistics included: mean, histogram, quantile, median, variance, OLS regression, logistic regression, probit regression, difference of means, unbiased privacy



OpenDP

<https://opendp.io>

Agreement on community standards needed for a federated Data Commons



Summary

Universities need to **collaborate** between research, libraries, and IT to:

- Catalog all the services offered to support research data and computing
- Centralize costly and complex services
- Build workflows to integrate services, research tools, computing, and repositories

The proposed Data Commons with **Dataverse** repositories **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 and computing
- Distributing data and context across systems in a standardized form
- Sharing sensitive, private data for research between industry, gov, and academia

THANKS!