A Harvard Data Commons

HUIT Tech Talk, December 8, 2020

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

IQSS
The Institute for Quantitative Social Science

HARVARD UNIVERSITY
Information Technology

Data access and data sharing are critical to today’s research. Although there has been clear progress in the last years, challenges remain to access, manage, and collaborate on research data at Harvard and beyond.

Global progress in data sharing and access

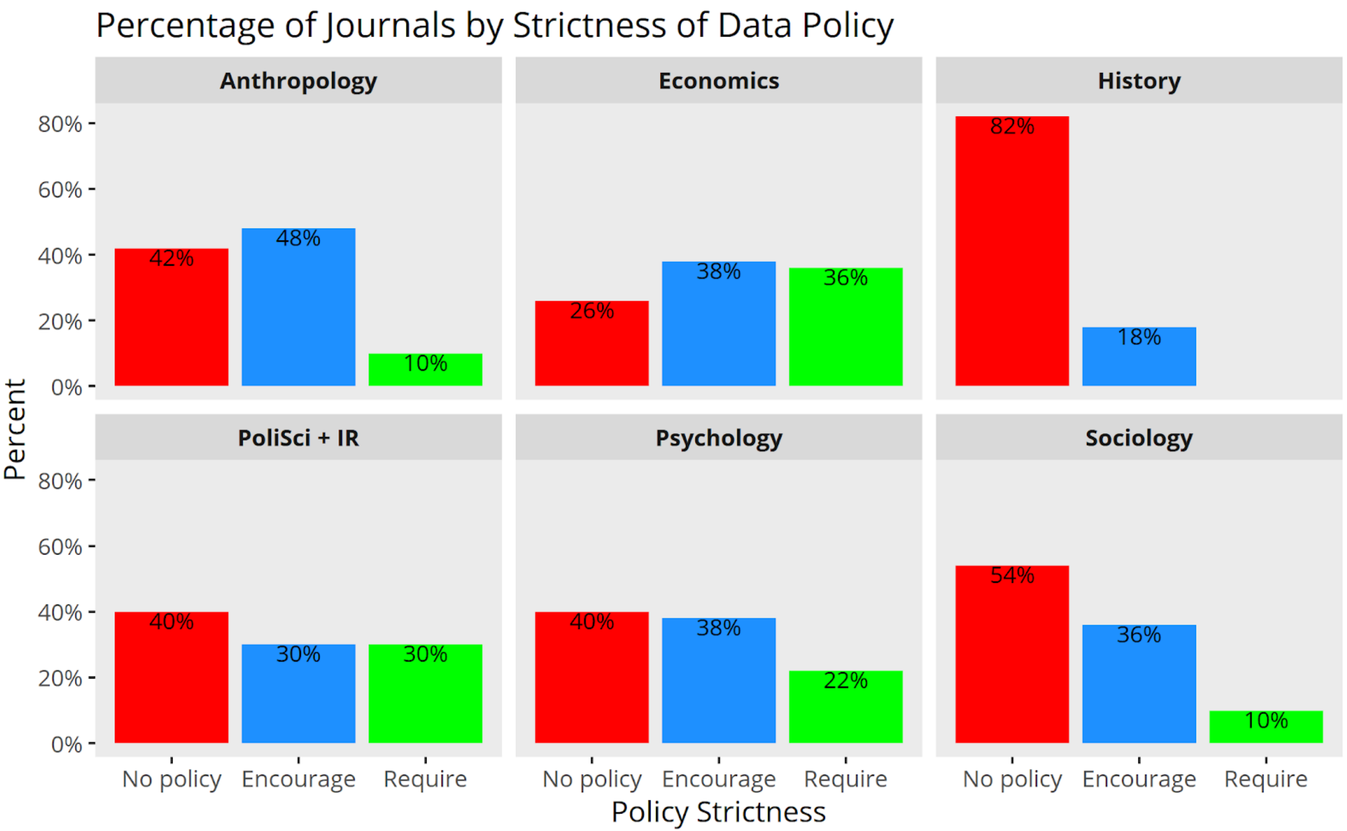
- **New data policies in journals**
  - *Example:* > 50% of top social science journals recommend or require sharing the data associated with the article
- **New data sharing mandates by funding entities**
  - *Example:* National Institutes of Health (NIH) recent release of Policy for Data Management 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


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

Global 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.”

Global 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

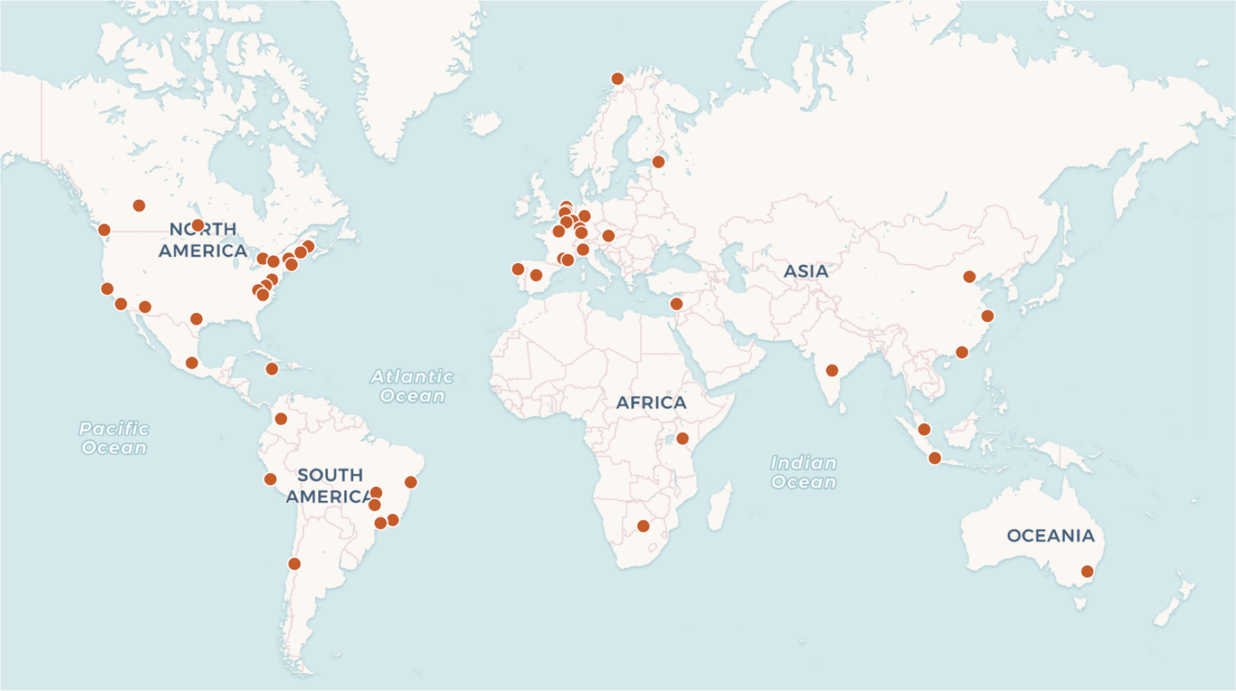
POSITION STATEMENT ON DATA

“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**.”

Global 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

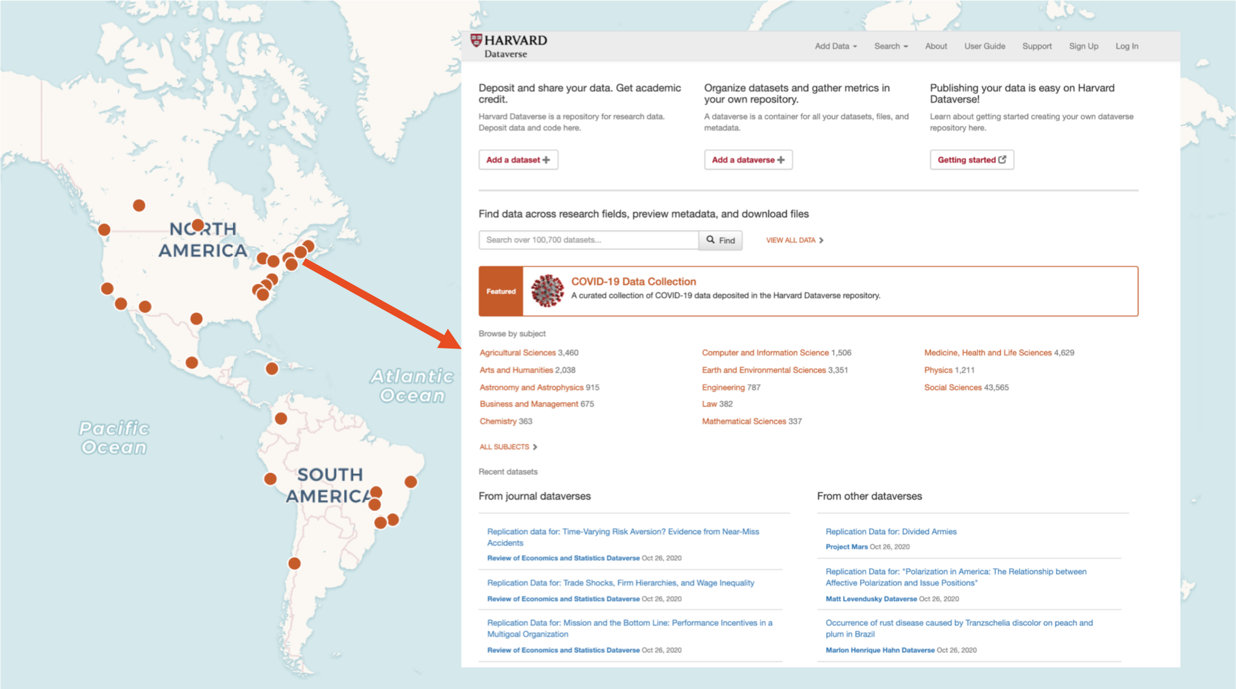
Dataverse® Federated FAIR data repositories worldwide



- **Open-source**
- **63** installations
- **6** continents
- **7K** Dataverse collections
- **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



Deposit and share your data. Get academic credit.

Harvard Dataverse is a repository for research data. Deposit data and code here.

Organize datasets and gather metrics in your own repository.

A dataverse is a container for all your datasets, files, and metadata.

Publishing your data is easy on Harvard Dataverse!

Learn about getting started creating your own dataverse repository here.

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

COVID-19 Data Collection

A curated collection of COVID-19 data deposited in the Harvard Dataverse repository.

Browse by subject

Agricultural Sciences 3,460
Arts and Humanities 2,038
Astronomy and Astrophysics 915
Business and Management 675
Chemistry 363
Computer and Information Science 1,506
Earth and Environmental Sciences 3,351
Engineering 787
Law 382
Mathematical Sciences 337
Medicine, Health and Life Sciences 4,629
Physics 1,211
Social Sciences 43,565

Recent datasets

From journal dataverses

Replication data for: Time-Varying Risk Aversion? Evidence from Near-Miss Accidents
Review of Economics and Statistics Dataverse Oct 26, 2020

Replication Data for: Trade Shocks, Firm Hierarchies, and Wage Inequality
Review of Economics and Statistics Dataverse Oct 26, 2020

Replication Data for: Mission and the Bottom Line: Performance Incentives in a Multigoal Organization
Review of Economics and Statistics Dataverse Oct 26, 2020

From other dataverses

Replication Data for: Divided Armies
Project Mars Oct 26, 2020

Replication Data for: "Polarization in America: The Relationship between Affective Polarization and Issue Positions"
Matt Levendusky Dataverse Oct 26, 2020

Occurrence of rust disease caused by Tranzschelia discolor on peach and plum in Brazil
Marlon Henrique Hahn Dataverse Oct 26, 2020

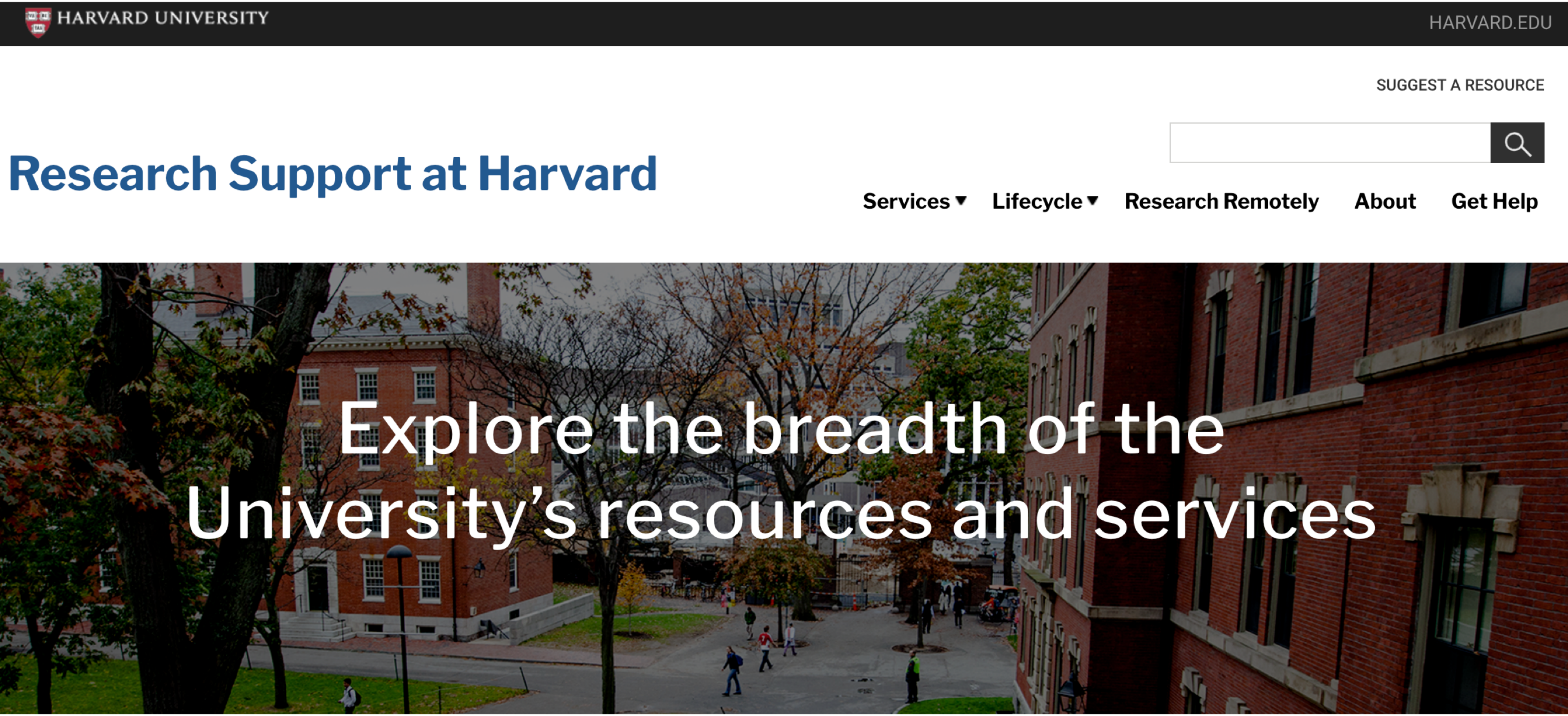
Harvard Dataverse: a collaboration between IQSS, the Library, and HUIT (https://dataverse.harvard.edu)

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

Progress at Harvard

A new inventory of research support services and a single resource to find them

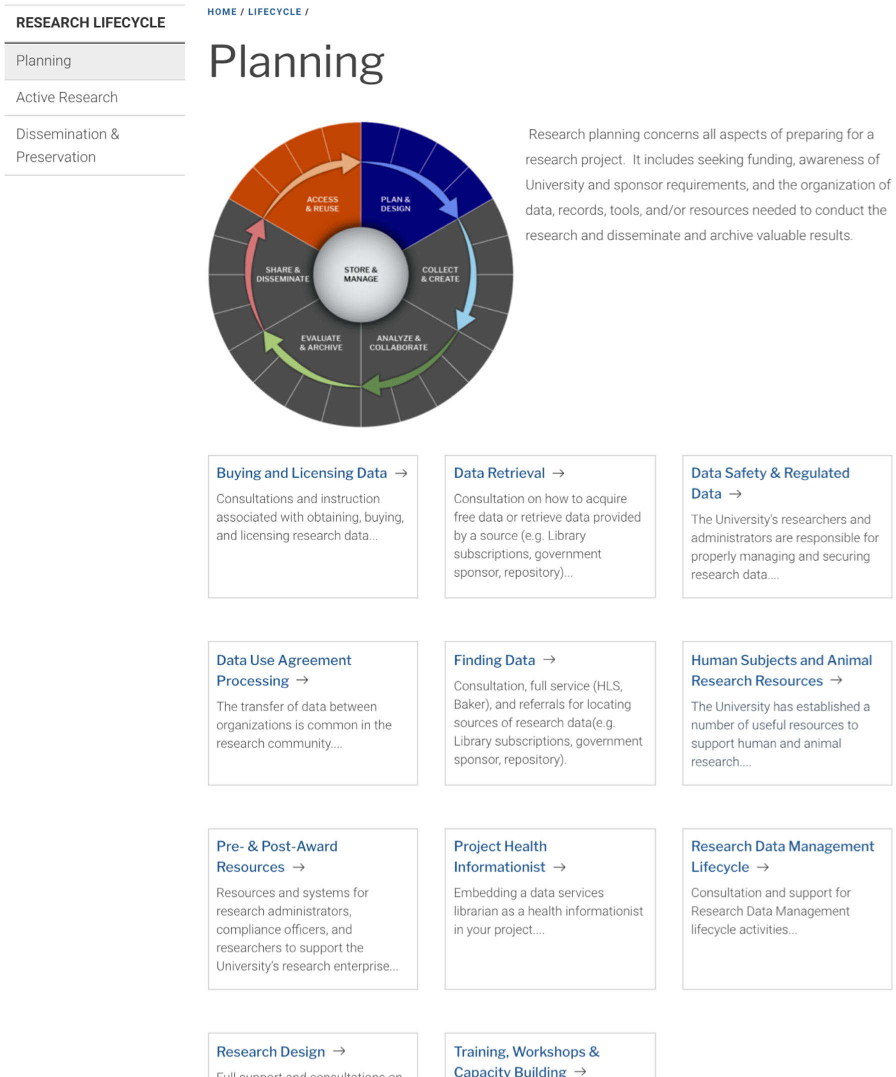
- A collaboration on building a research support services catalog and a website with a **common vision**:
  - *“To help faculty, researchers, and those who work with them to advance their research by easily finding and browsing the University’s breadth of resources and services”*
- Sponsored by the Library, HUIT, and Office of Vice Provost for Research
- Initial launch planned for: Early 2021



Research Support at Harvard

Explore the breadth of the University’s resources and services

https://researchsupport.harvard.edu/



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.

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, government sponsor, repository)...

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 and Animal Research Resources →
The University has established a number of useful resources to support human and animal research....

Pre- & Post-Award Resources →
Resources and systems for research administrators, compliance officers, and researchers to support the University's research enterprise...

Project Health Informationist →
Embedding a data services librarian as a health informationist in your project....

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

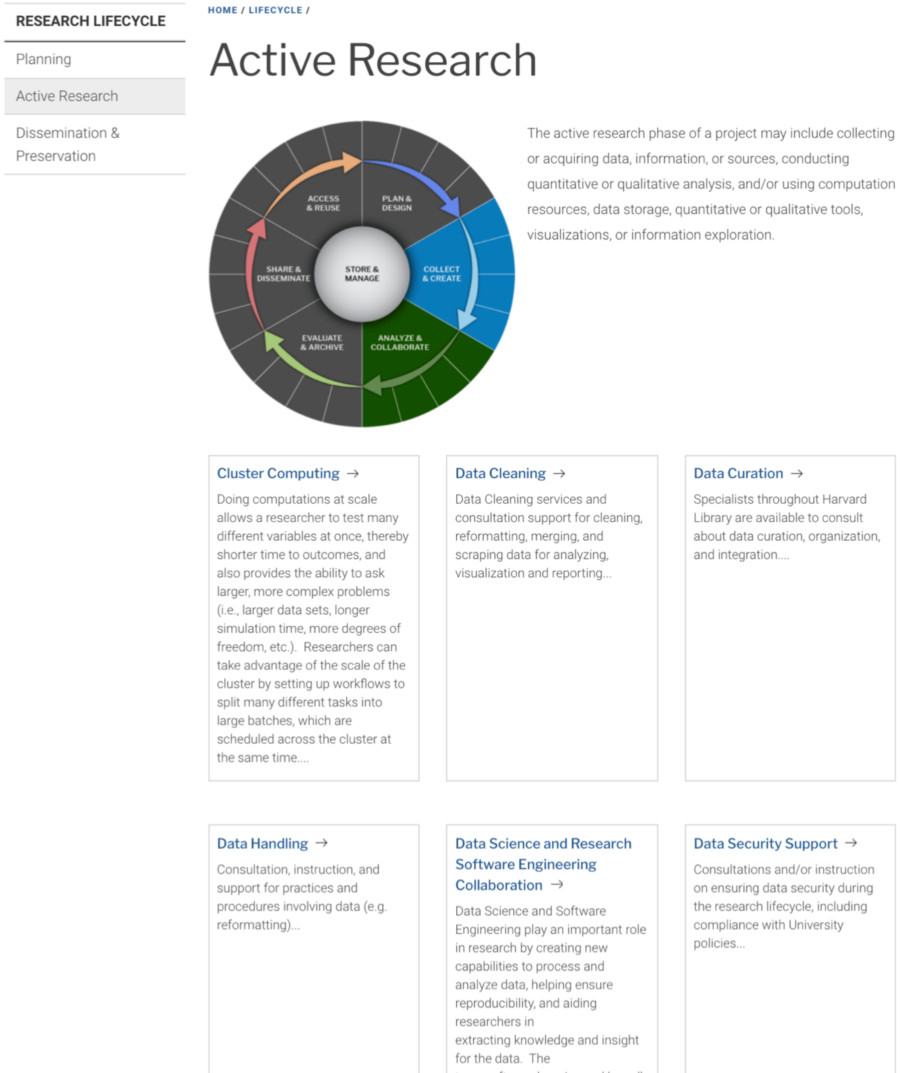
Research Design →
Full support and consultations on

Training, Workshops & Capacity Building →

Planning:
Access & Reuse
Plan & Design

11 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



RESEARCH LIFECYCLE

Planning

Active Research

Dissemination & Preservation

HOME / LIFECYCLE /

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 larger, more complex problems (i.e., larger data sets, longer simulation time, more degrees of freedom, etc.). Researchers can take advantage of the scale of the cluster by setting up workflows to split many different tasks into large batches, which are scheduled across the cluster at the same time....

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....

Data Handling →

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

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

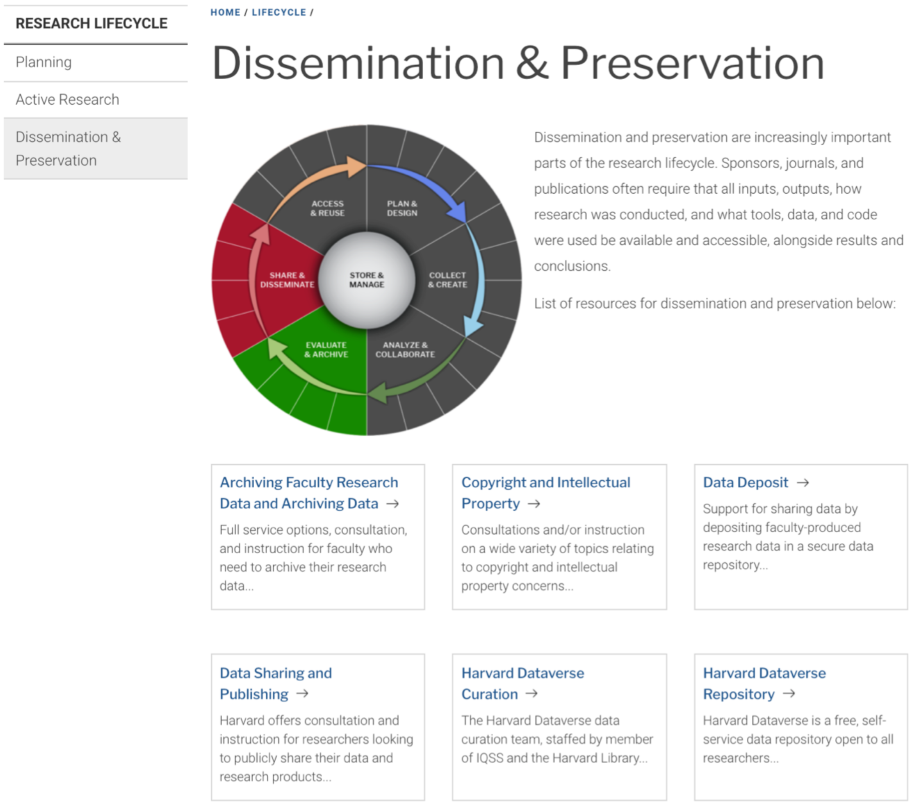
Data Security Support →

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

Active Research:
Collect & Create
Analyze & Collaborate

19 service offerings

- Cluster Computing, Virtual Instances
- Research Data Storage, Database, Data Security
- Software and Platforms
- 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
- Microbiological Safety



RESEARCH LIFECYCLE

- Planning
- Active Research
- Dissemination & Preservation

HOME / LIFECYCLE /

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...

Data Deposit →
Support for sharing data by depositing faculty-produced research data in a secure data repository...

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

Harvard Dataverse Curation →
The Harvard Dataverse data curation team, staffed by member of IQSS and the Harvard Library...

Harvard Dataverse Repository →
Harvard Dataverse is a free, self-service data repository open to all researchers...

Dissemination & Preservation:
Evaluate & Archive
Share & Disseminate

6 service offerings:

- Copyright and Intellectual Property
- Archiving data
- Data Deposit
- Data Sharing and Publishing
- Harvard Dataverse Curation
- Harvard Dataverse Repository

Services	Planning	Active Research	Dissemination & Preservation
Research Administration & Compliance	• Data Safety and Regulated Data • Data Use Agreement Processing • Human Subjects & Animal Research Resources • Pre- & Post-Award Resources	• Microbiological Safety	
Research Computing		• Cluster Computing, Virtual Instances • Research Data Storage, Database, Data Security • Research Computing Consulting & Facilitation • Data Science and Research Software Engineering	
Research Data & Scholarship	• Buying and Licensing Data • Data Retrieval, Finding Data • Research Data Management Lifecycle • Research Design • Training, Workshop, Capacity Building • Project Health Informationist	• Statistical Analysis, Text Analysis • Dataset Creation, Data Cleaning, Data Curation, Data Handling, Metadata creation • Data Visualization • Geospatial data • Qualitative Data Support • Software and Platforms	• Copyright and Intellectual Property • Archiving data • Data Deposit • Data Sharing and Publishing • Harvard Dataverse Curation • Harvard Dataverse Repository

Challenges Remain

The new Research Support Services inventory shows us that Harvard provides a fantastic wide set of services to support managing, working on, and sharing of research data. But gaps still exist to support an increasing number of external mandates and cross-disciplinary collaborations, ensure compliance, facilitate the use of sensitive data and private data from industry, and apply open science and reproducibility best practices to our research.

Challenges remain

Context, documentation, provenance

Collaborations

Large and complex datasets

Sensitive and private data

- Insufficient information to use/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 private data

- Difficult to find research datasets before publication
- Difficult to access data from across groups, schools, and organizations
- Often duplicative, costly efforts

Challenges remain

Context, documentation, provenance

Collaborations

Large and complex datasets

- Need for storage keeps increasing across disciplines
- Data cannot be downloaded to local computer
- Often special software is required to explore and make sense of the data

Sensitive and private data

Challenges remain

Context, documentation, provenance

Collaborations

Large and complex datasets

Sensitive and private data

- Not all data can be open
- Need to ensure compliance of security/access requirements based on data sensitivity
- Difficult to track Data Use Agreements with the actual datasets
- Access to private data from industry for research still difficult and limited

A Harvard Data Commons – Background and Vision

From Data-Centric Systems …

“NSF should support foundational cyberinfrastructure research for data science with a focus on frameworks and tools for data science, data centric systems architectures, and data repositories:

- **Data Science:** Methods and standard tools for analysis, synthesis, simulation, visualization, sharing, integration, and management.
- **Data Systems Architecture:** A fast, scalable fault-tolerant infrastructure for real-time analyses and data transfer; controlled data access.
- **Data Repositories:** Well curated, validated, open access repositories required to ensure that the results of scientific research are available.”
- [NSF CI2030: Future Advanced Cyberinfrastructure document, 2018]

… to 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."

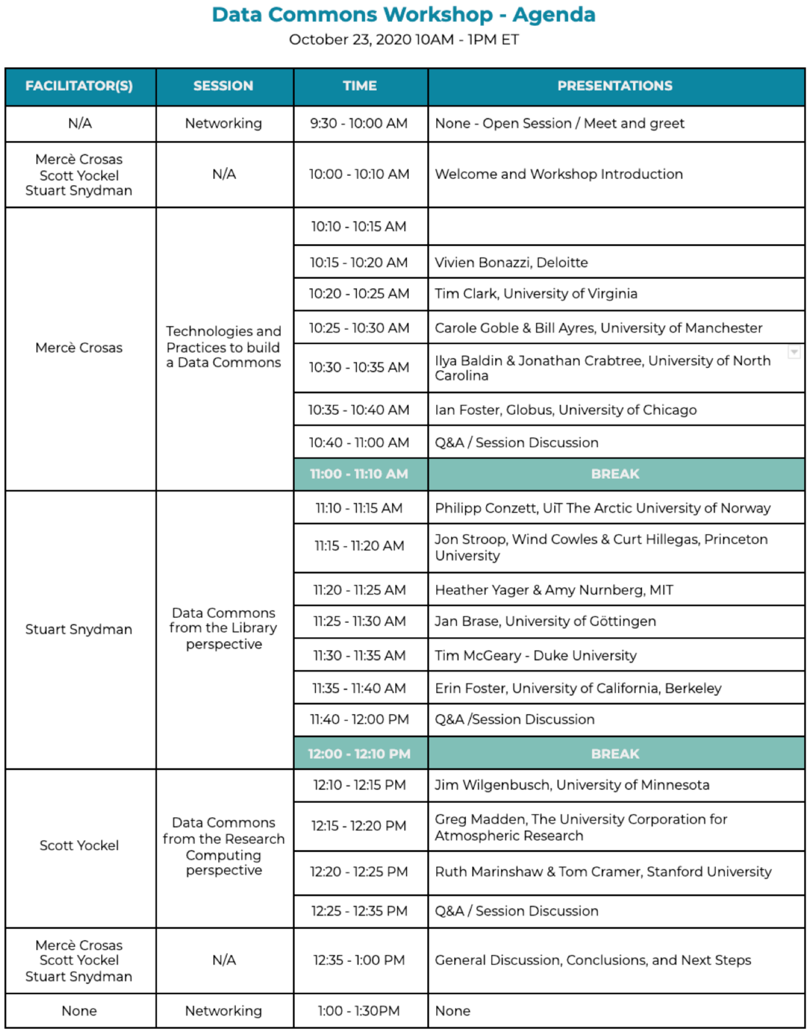
[Robert Grossman, on the NIH Data Commons Consortium initiative]

Harvard Data Commons ITCRB Proposal – October 2019

- **Project Vision:** To substantially improve and support research collaboration, data sharing, and data tracking throughout Harvard by creating a data-centric hub accessible to all Harvard researchers.
- **Who we are**:
  - **Mercè Crosas**, HUIT, IQSS
  - **Scott Yockel**, HUIT, FAS RC
  - **Piotr Sliz**, HMS, Children's Hospital
  - **Bill Barnett**, HMS RC
  - **Paul DiBello,** HBS RC
  - **Stu Snydman**, Library Technology Services, HUIT/Library
  - *We also engaged with the university's Research Data Coordination working group and the Research Computing and Data Council for input and use cases*
- Initial effort included working towards submitting a proposal for the 2019 ITCRB funding cycle, which ended up being cancelled.

Data Commons for Institutions Workshop – October 2020

We brought together more than 30 experts from 15 universities for a single day workshop to discuss ongoing efforts to develop institutional data commons.



Data Commons Workshop - Agenda

October 23, 2020 10AM - 1PM ET

FACILITATOR(S)	SESSION	TIME	PRESENTATIONS
N/A	Networking	9:30 - 10:00 AM	None - Open Session / Meet and greet
Mercè Crosas Scott Yockel Stuart Snydman	N/A	10:00 - 10:10 AM	Welcome and Workshop Introduction
Mercè Crosas	Technologies and Practices to build a Data Commons	10:10 - 10:15 AM	
		10:15 - 10:20 AM	Vivien Bonazzi, Deloitte
		10:20 - 10:25 AM	Tim Clark, University of Virginia
		10:25 - 10:30 AM	Carole Goble & Bill Ayres, University of Manchester
		10:30 - 10:35 AM	Ilya Baldin & Jonathan Crabtree, University of North Carolina
		10:35 - 10:40 AM	Ian Foster, Globus, University of Chicago
		10:40 - 11:00 AM	Q&A / Session Discussion
		11:00 - 11:10 AM	BREAK
Stuart Snydman	Data Commons from the Library perspective	11:10 - 11:15 AM	Philipp Conzett, UiT The Arctic University of Norway
		11:15 - 11:20 AM	Jon Stroop, Wind Cowles & Curt Hillegas, Princeton University
		11:20 - 11:25 AM	Heather Yager & Amy Nurnberg, MIT
		11:25 - 11:30 AM	Jan Brase, University of Göttingen
		11:30 - 11:35 AM	Tim McGeary - Duke University
		11:35 - 11:40 AM	Erin Foster, University of California, Berkeley
		11:40 - 12:00 PM	Q&A /Session Discussion
		12:00 - 12:10 PM	BREAK
Scott Yockel	Data Commons from the Research Computing perspective	12:10 - 12:15 PM	Jim Wilgenbusch, University of Minnesota
		12:15 - 12:20 PM	Greg Madden, The University Corporation for Atmospheric Research
		12:20 - 12:25 PM	Ruth Marinshaw & Tom Cramer, Stanford University
		12:25 - 12:35 PM	Q&A / Session Discussion
Mercè Crosas Scott Yockel Stuart Snydman	N/A	12:35 - 1:00 PM	General Discussion, Conclusions, and Next Steps
None	Networking	1:00 - 1:30PM	None

Data Commons Workshop – Outcome

- **Key themes** for developing a Data Commons for an institution included:
  - Be service oriented and user-focused
  - Host tools and offer training, and aim to integrate them to facilitate seamless support throughout the research lifecycle
  - Align with common institutional data sharing and management policies
  - Establish leading practices for creating, accessing, using, and collaborating on data within and across institutions
  - Enable ease of data flowing in and out of as well as across data commons by defining a minimum set of standards

A Harvard Data Commons should provide the integrated technology behind the services that support the research data lifecycle.

A Harvard Data Commons vision

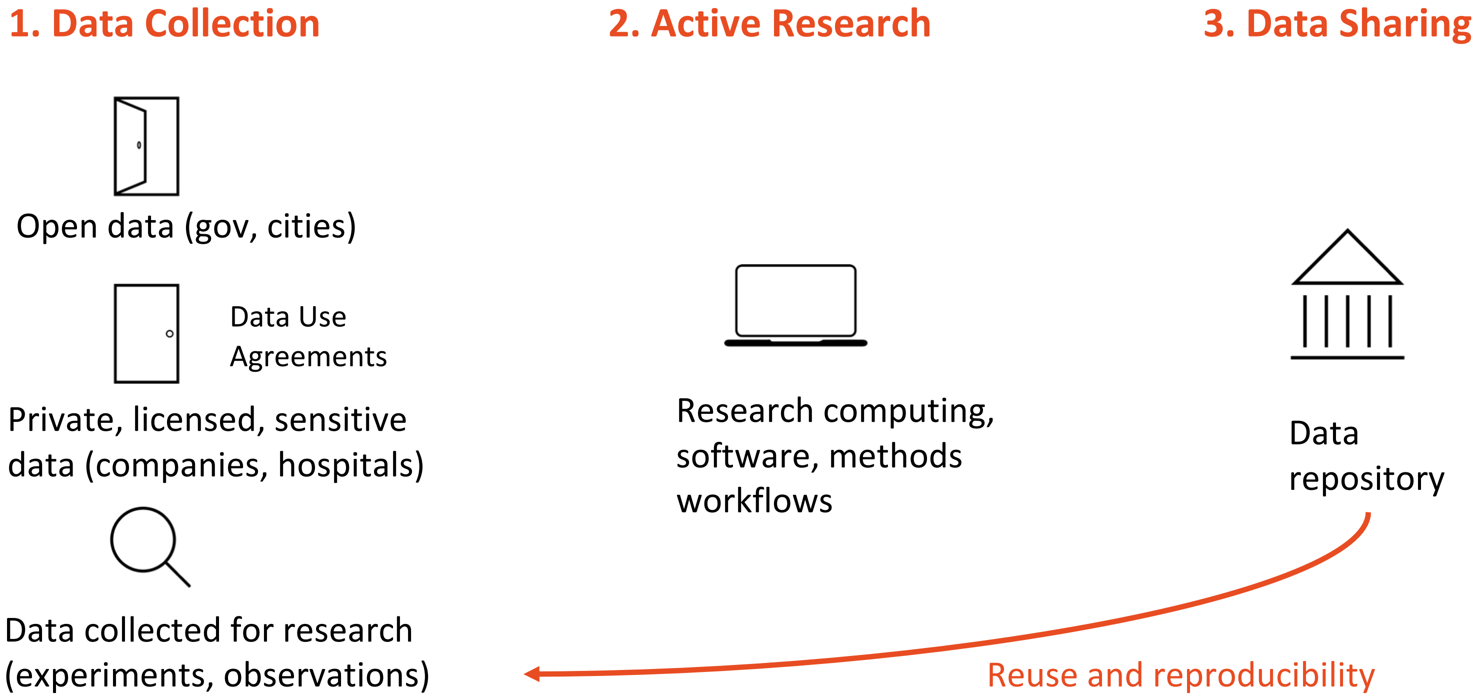
Context, documentation, provenance

Collaborations

Large and complex datasets

Sensitive and private data

Data Lifecycle



1. Data Collection

Open data (gov, cities)

Data Use Agreements

Private, licensed, sensitive data (companies, hospitals)

Data collected for research (experiments, observations)

2. Active Research

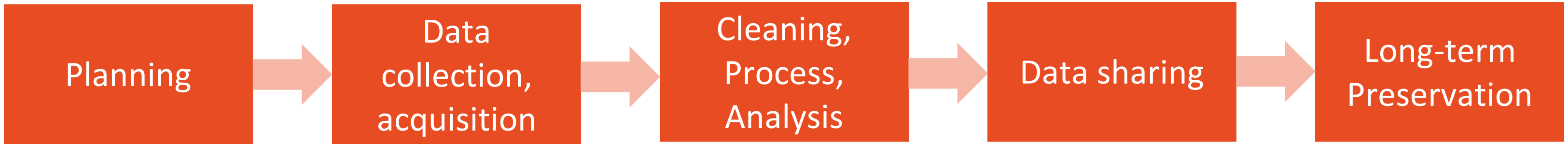
Research computing, software, methods workflows

3. Data Sharing

Data repository

Reuse and reproducibility

Seamless integration of tools and data packages across the data lifecycle

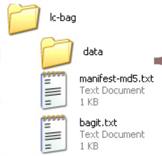


Machine-Actionable Data Management Plan (DMP)

Machine-Actionable Data Use Agreement (DUA) and IRB review

Data,
Code,
Workflows,
Computational Notebooks,
Electronic Lab Notebooks

Package w/ Metadata & Provenance



lc-bag
data
manifest-md5.txt
Text Document
1 KB
bagit.txt
Text Document
1 KB

Harvard managed research data collections

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 Harvard Data Commons vision

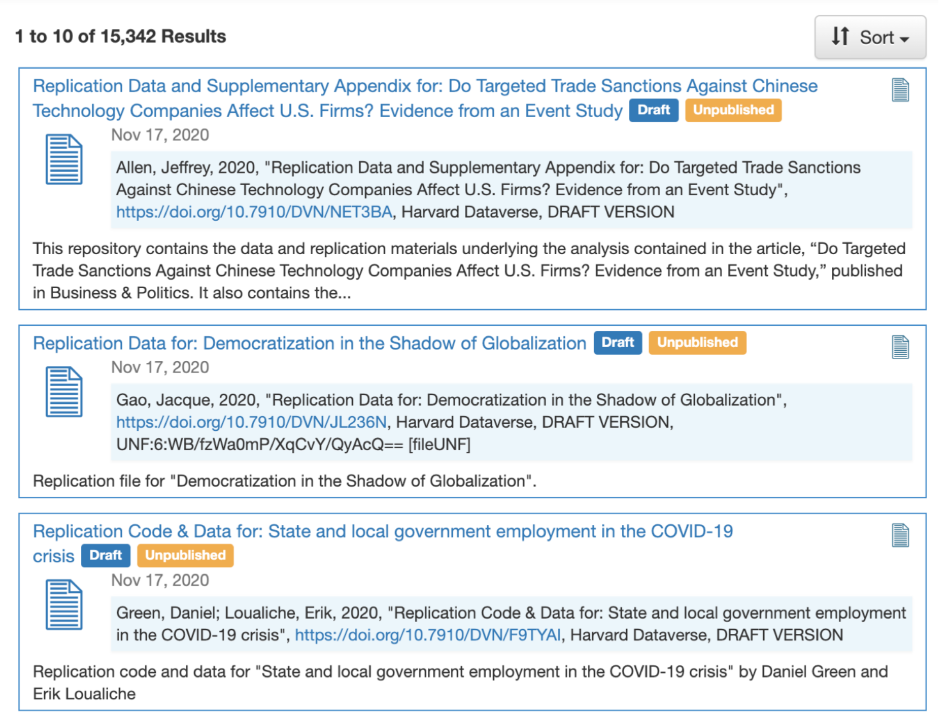
Context, documentation, provenance

Collaborations

Large and complex datasets

Sensitive and private data

A common registry for active research datasets



1 to 10 of 15,342 Results

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

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

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

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

- **Metadata catalog** for unpublished and published Harvard datasets
- **Permissions** to access data granted to **collaborators**
- **InCommon/HarvardKey/ORCID authentication** mechanism to facilitate access

A Data Commons vision

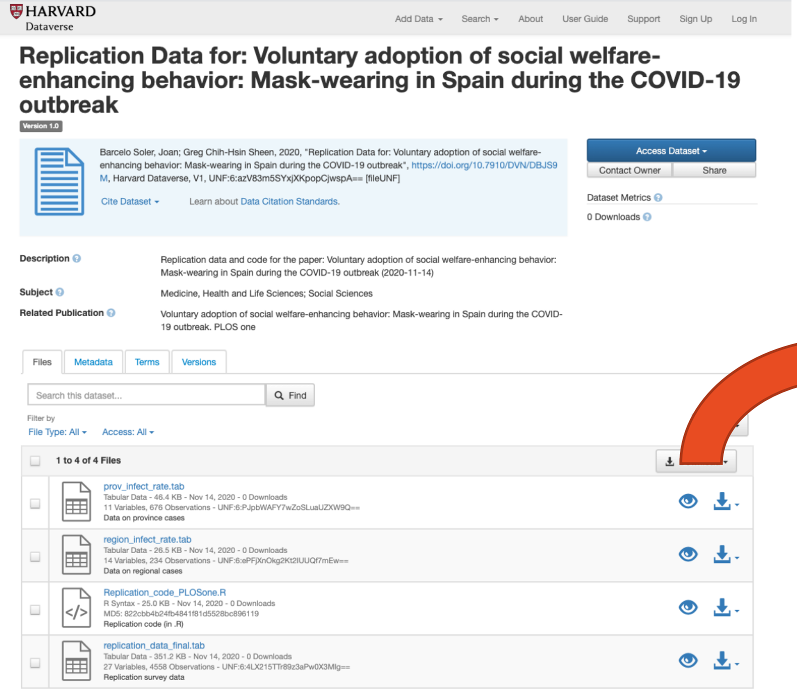
Context, documentation, provenance

Collaborations

Large and complex datasets

Sensitive and private data

Co-location of data and computing on the cloud



Replication Data for: Voluntary adoption of social welfare-enhancing behavior: Mask-wearing in Spain during the COVID-19 outbreak

Barcelo Soler, Joan; Greg Chih-Hsin Sheen, 2020, "Replication Data for: Voluntary adoption of social welfare-enhancing behavior: Mask-wearing in Spain during the COVID-19 outbreak", https://doi.org/10.7910/DVN/DBJS9M, Harvard Dataverse, V1, UNF:6:azV83m5SYxjXKpopCjwspA== [fileUNF]

Learn about Data Citation Standards.

Dataset Metrics

0 Downloads

Description 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)

Subject Medicine, Health and Life Sciences; Social Sciences

Related Publication Voluntary adoption of social welfare-enhancing behavior: Mask-wearing in Spain during the COVID-19 outbreak. PLOS one

1 to 4 of 4 Files

prov_infect_rate.tab
Tabular Data - 46.4 KB - Nov 14, 2020 - 0 Downloads
11 Variables, 676 Observations - UNF:6:PJpbWAFY7wZoSLuaUZXW9Q==
Data on province cases

region_infect_rate.tab
Tabular Data - 26.5 KB - Nov 14, 2020 - 0 Downloads
14 Variables, 234 Observations - UNF:6:ePFjXnOkg2Kt2IUUQf7mEw==
Data on regional cases

Replication_code_PLOSone.R
R Syntax - 25.0 KB - Nov 14, 2020 - 0 Downloads
MD5: 822cbb4b24fb4841f81d5528bc896119
Replication code (in .R)

replication_data_final.tab
Tabular Data - 351.2 KB - Nov 14, 2020 - 0 Downloads
27 Variables, 4558 Observations - UNF:6:4LX215TTr89z3aPw0X3Mlg==
Replication survey data

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 an OpenStack open-source cloud

A Data Commons vision

Context, documentation, provenance

Collaborations

Large and complex datasets

Sensitive and private data

Non-sensitive vs. sensitive datasets using DataTags and Harvard Security Levels

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

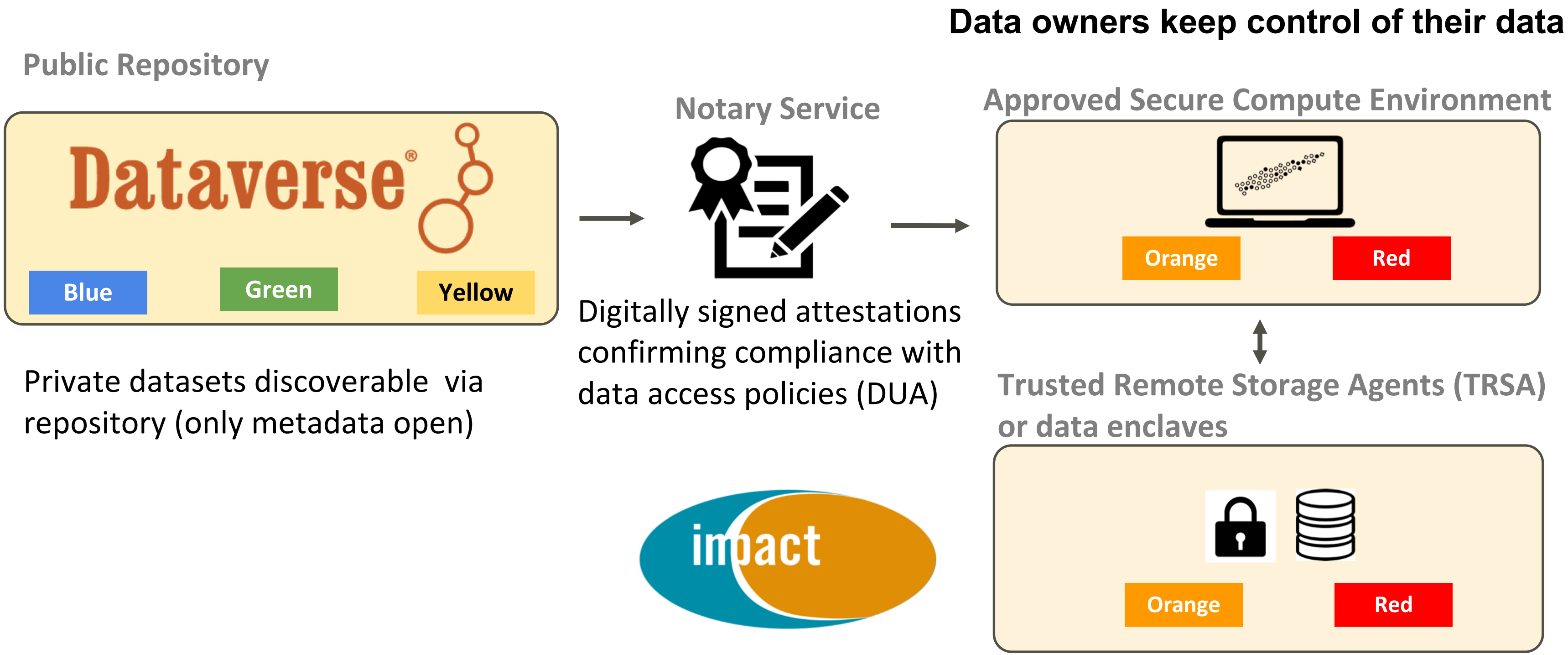
Blue	Green	Yellow	Orange	Red	Crimson

IRB determines whether the data are sensitive or non-sensitive; SSOs determine Security Level

Non-Sensitive			**Sensitive**		
Security Level 1	*Security Level 1*	*Security Level 2*	*Security Level 3*	*Security Level 4*	*Security Level 5*
Do not need to register Public data	Need to register De-identified data with low risk of re-identification	Need permission De-identified data with risk of re-identification Identifiable data but not considered sensitive	Education data (FERPA) Datasets under contractual agreement GDPR **not extra sensitive** level	Government issued identifiers HIPAA regulated - Personal Health Information GDPR **extra sensitive** level	It would put subject’s life at risk if disclosed Data locked in a physically secure room not connected

https://security.harvard.edu/files/it-security/files/rdslexamples.pdf

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

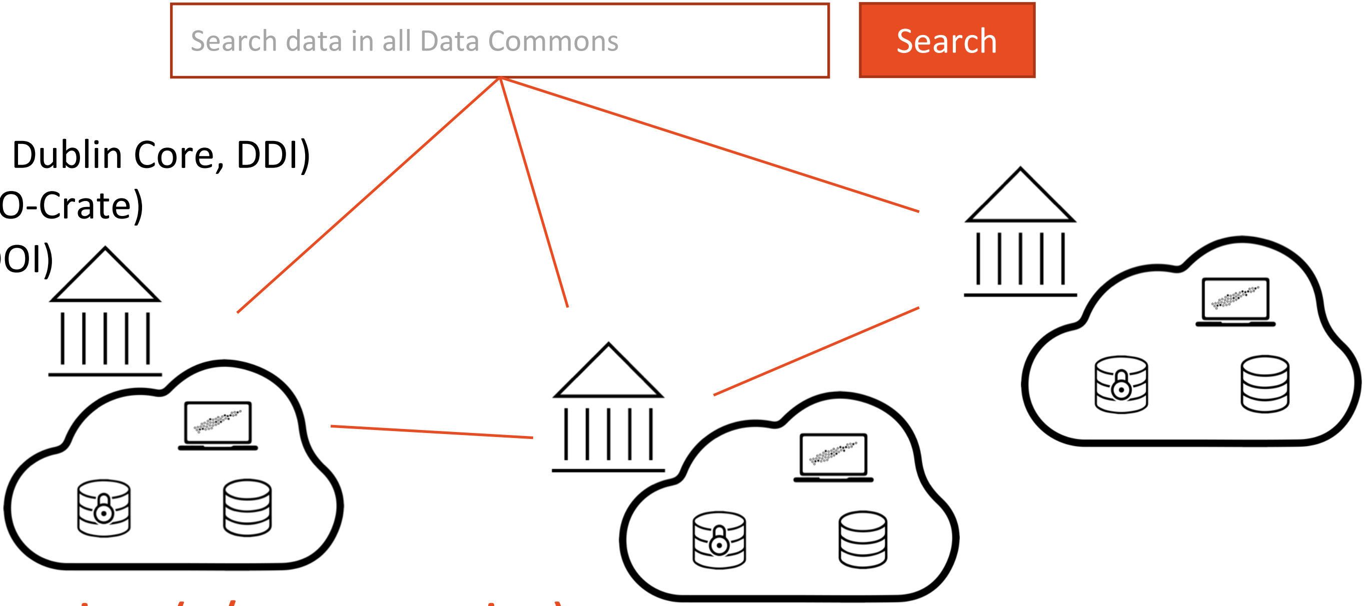
Trusted Remote Storage Agents (TRSA) or data enclaves

Orange Red

impact

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

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 Harvard Data Commons would **lower the barrier** to:

- Finding active research data, in addition to data already publicly published
- Accessing and tracking the data in one place for collaboration
- Managing data in a repository, while being connected to research cloud computing
- Ensuring compliance from agreements to actual data access and use
- Sharing private data for research securely between industry and the university
- Distributing data across systems in a standardized form
- (Re)Analyzing published results in a simple step

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.