A new approach to data-intensive research support:

Computational Methods and Data at Yale University Library

Rebecca B. Dikow, Ph.D.

Director of Computational Methods and Data

Yale UNIVERSITY LIBRARY

Talk outline

1

Evolving needs require a new approach

Researchers are coming to data and computation with diverse perspectives and needs and an increasingly ready access to advanced technology

2

A key collaboration

Highlighting our work with the Data Intensive Social Science Center to host and manage the Yale Dataverse

3

Next steps

Discussion of ongoing work to define services based on values of openness, collaboration, and accessibility

My path to libraries

- My background is in Evolutionary Biology
- I recently came to Yale from the Smithsonian Institution where I was a Data Scientist and led the Data Science Lab, a research team within the Office of the CIO. Our work focused on:
  - Using genomics and machine learning tools to analyze museum collections and archives data
  - Supporting researchers in their use of High-Performance Computing
  - Defining best practices for using AI on museum collections data

Researcher perspectives

Data stories

- Vicki Funk was a Senior Research Botanist at the National Museum of Natural History, Smithsonian Institution
- Vicki's impact on collections:
  - 269 families collected
  - 11,782 specimens collected from 36 countries
  - 4,504 specimens identified from 70 countries
  - 5,708 specimens used in 212 publications

  Data from bionomia.net

Researcher perspectives

Data stories

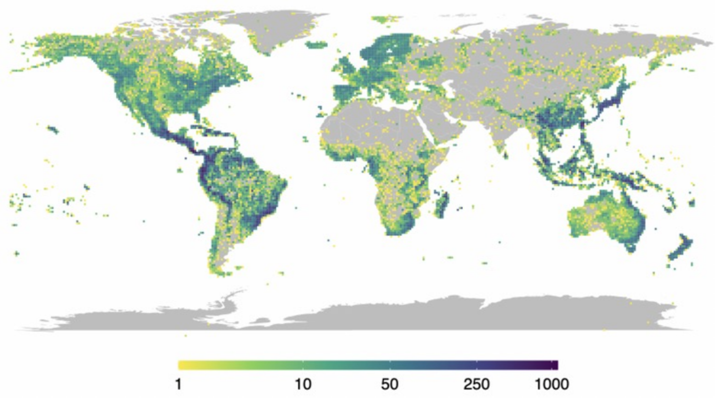
- A single plant specimen can produce diverse data

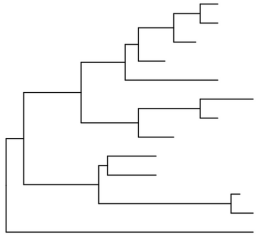
UNITED STATES
3316421
NATIONAL HERBARIUM
Image No.
04166627
UNITED STATES NATIONAL HERBARIUM
PLANTS OF BOLIVIA
COMPOSITAE
Achyrocline
La Paz: Murillo. Zonga Valley. 15 km N of cumbre, around Lga Viscachani. 16°11'30"S, 68°07'20"W. Elev. 3750 m.
V. A. Funk 11303
12 Apr 1995
N. Bernal H.
US

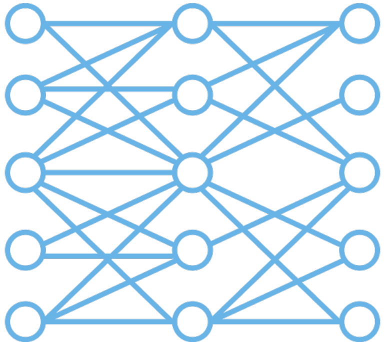
Researcher perspectives

Data stories

- These data may be stored in different places and analyzed with different tools







Researcher perspectives

Data stories

- Common sentiment: I can’t keep up with everything I have to do with my data.

R Studio

jupyter

GBIF Global Biodiversity Information Facility

GitHub

figshare

NCBI

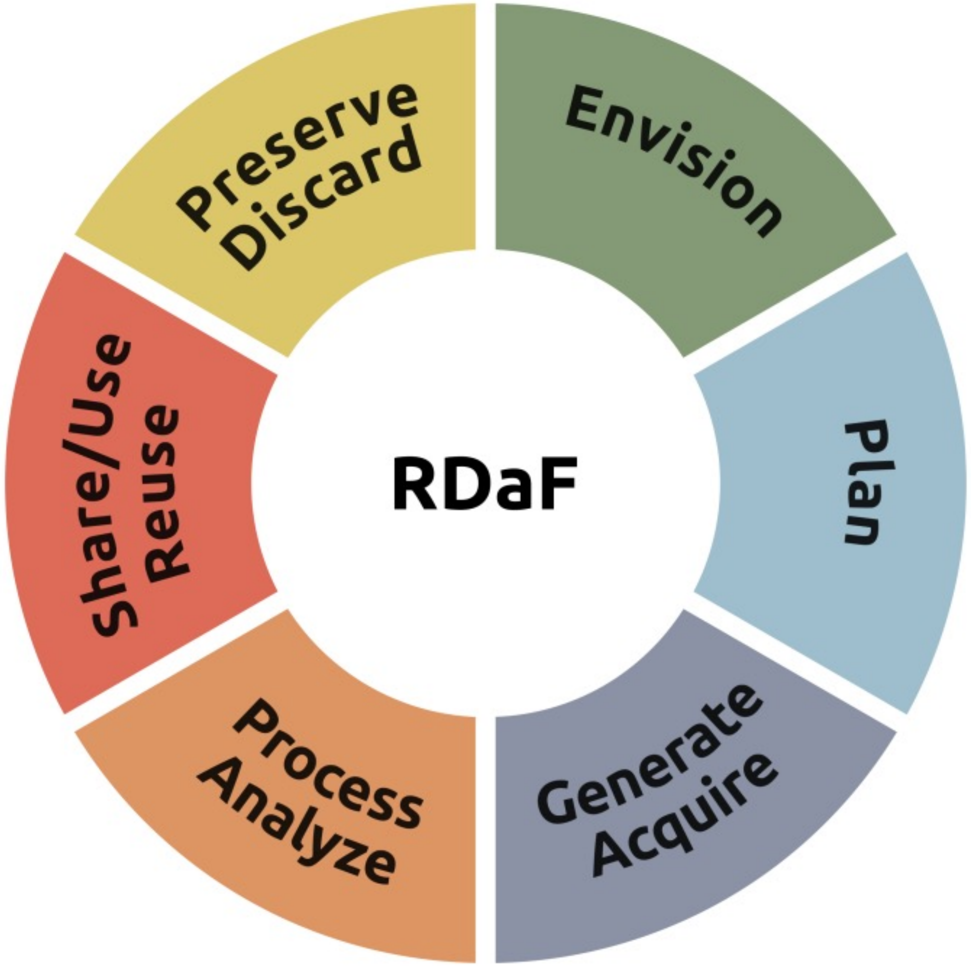
Researcher perspectives

Data stories

- While some researchers readily engage, many do not see themselves in the terms data, data science, research data, or computation
- Many, students in the humanities in particular, will say, “I don’t think I have data,” but will say they are not sure how to ask for help in organizing their materials, sources, or notes about objects studied
- What they describe is often just as complex a problem as any biomedical researcher or physicist has to face in terms of data management

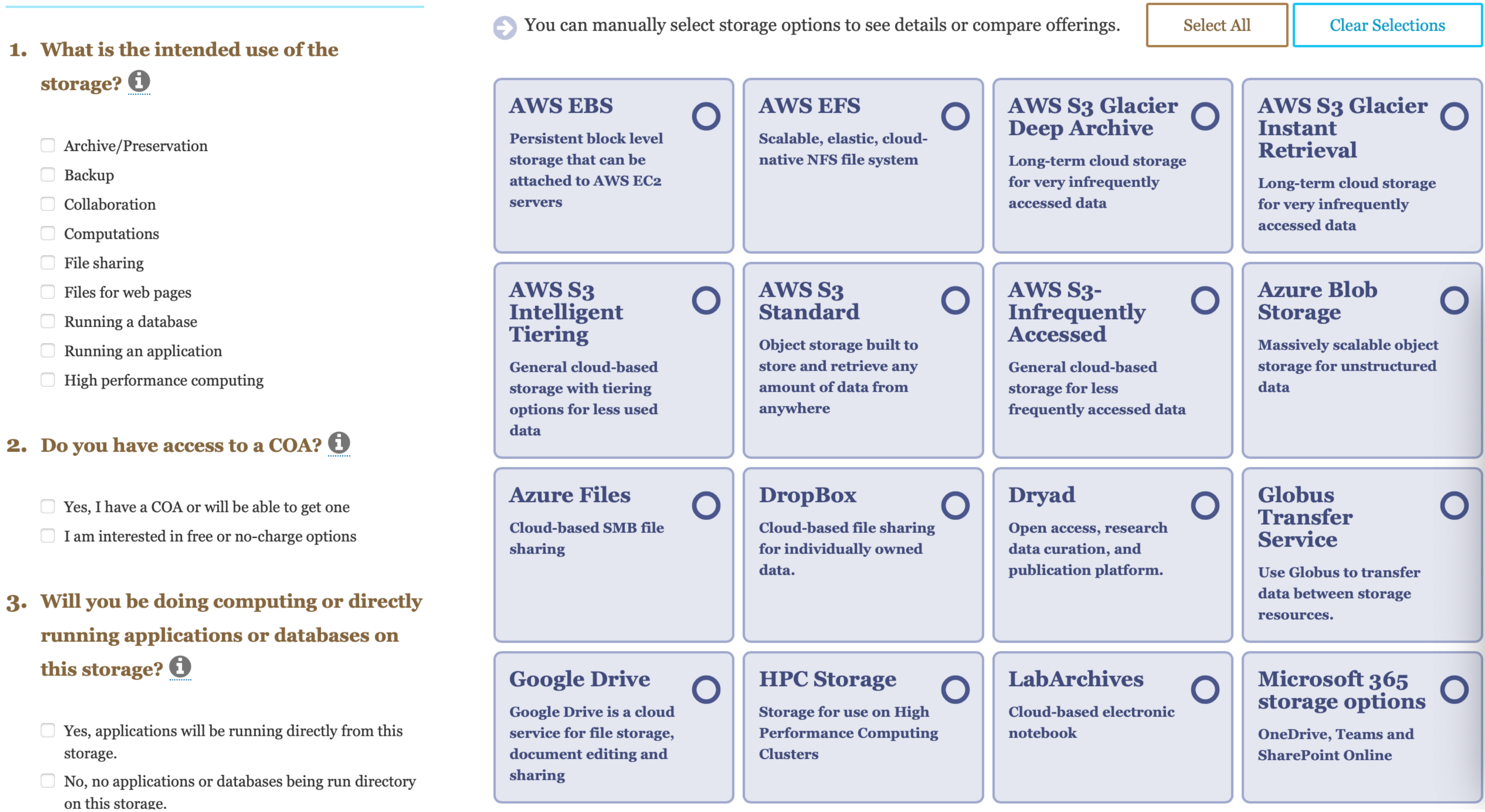
NIST Research Data Framework

- The Library sees researchers at every part of the research data lifecycle, across disciplines



Storage@Yale

While data storage is only a part of research data management, it can be overwhelming to navigate the available options.



1. **What is the intended use of the storage?** ℹ

□ Archive/Preservation
□ Backup
□ Collaboration
□ Computations
□ File sharing
□ Files for web pages
□ Running a database
□ Running an application
□ High performance computing

2. **Do you have access to a COA?** ℹ

□ Yes, I have a COA or will be able to get one
□ I am interested in free or no-charge options

3. **Will you be doing computing or directly running applications or databases on this storage?** ℹ

□ Yes, applications will be running directly from this storage.
□ No, no applications or databases being run directory on this storage.

You can manually select storage options to see details or compare offerings.

○ **AWS EBS**
Persistent block level storage that can be attached to AWS EC2 servers

○ **AWS EFS**
Scalable, elastic, cloud-native NFS file system

○ **AWS S3 Glacier Deep Archive**
Long-term cloud storage for very infrequently accessed data

○ **AWS S3 Glacier Instant Retrieval**
Long-term cloud storage for very infrequently accessed data

○ **AWS S3 Intelligent Tiering**
General cloud-based storage with tiering options for less used data

○ **AWS S3 Standard**
Object storage built to store and retrieve any amount of data from anywhere

○ **AWS S3-Infrequently Accessed**
General cloud-based storage for less frequently accessed data

○ **Azure Blob Storage**
Massively scalable object storage for unstructured data

○ **Azure Files**
Cloud-based SMB file sharing

○ **DropBox**
Cloud-based file sharing for individually owned data.

○ **Dryad**
Open access, research data curation, and publication platform.

○ **Globus Transfer Service**
Use Globus to transfer data between storage resources.

○ **Google Drive**
Google Drive is a cloud service for file storage, document editing and sharing

○ **HPC Storage**
Storage for use on High Performance Computing Clusters

○ **LabArchives**
Cloud-based electronic notebook

○ **Microsoft 365 storage options**
OneDrive, Teams and SharePoint Online

Current Computational Methods and Data team

Barbara Esty

Data and Statistical Support Services

Miriam Olivares

Geospatial Support Services

Gavi Levy Haskell

Digital Humanities

Kayla Shipp

Digital Humanities

Current patron-facing services

1

Research Consultations

Anyone can book time with our staff and more than a dozen graduate student consultants, with expertise in Geospatial methods, Statistics, and Digital Humanities

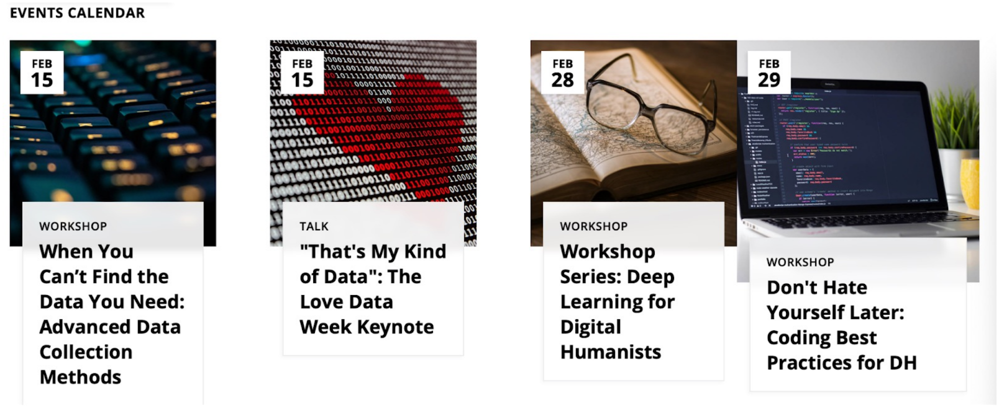
2

Access to Technology

Our staff provide researchers access to hardware, software, and digital wayfinding to resources across campus

3

Community building



EVENTS CALENDAR

FEB 15
WORKSHOP
When You Can't Find the Data You Need: Advanced Data Collection Methods

FEB 15
TALK
"That's My Kind of Data": The Love Data Week Keynote

FEB 28
WORKSHOP
Workshop Series: Deep Learning for Digital Humanists

FEB 29
WORKSHOP
Don't Hate Yourself Later: Coding Best Practices for DH

Current patron-facing services

4

Instruction

Our staff design and deliver workshops and course modules. Offerings reflect needs as communicated by faculty, staff, and students

5

Collaboration

We are available to collaborate on projects driven by researchers across campus

6

Digital Wayfinding

Our staff has a bird's eye view of campus resources and can help make connections and navigate inefficiencies

Current back-end services

1

Data acquisition

Our team plays a key role in acquiring datasets, giving access to researchers and keeping track of user agreements and licenses

2

Maintaining software

We work with Library IT to ensure availability of dozens of software packages from R to Stata, etc.

3

Advocating for access to resources

We bring researcher needs to the attention of Library IT, central Yale ITS, and the Yale Center for Research Computing

Current challenges

1

Maintaining digital projects

Collaborative projects are exciting and energizing, but can lead to our staff needing to maintain systems

2

Navigating AI

There is increasing access to AI tools, but not necessarily the AI literacy needed to navigate how to use them responsibly in research

3

Staffing!

As demand for research data support increases, we need additional staff in both patron-facing and Library IT roles to keep up

DISSC: Data Intensive Social Science Center

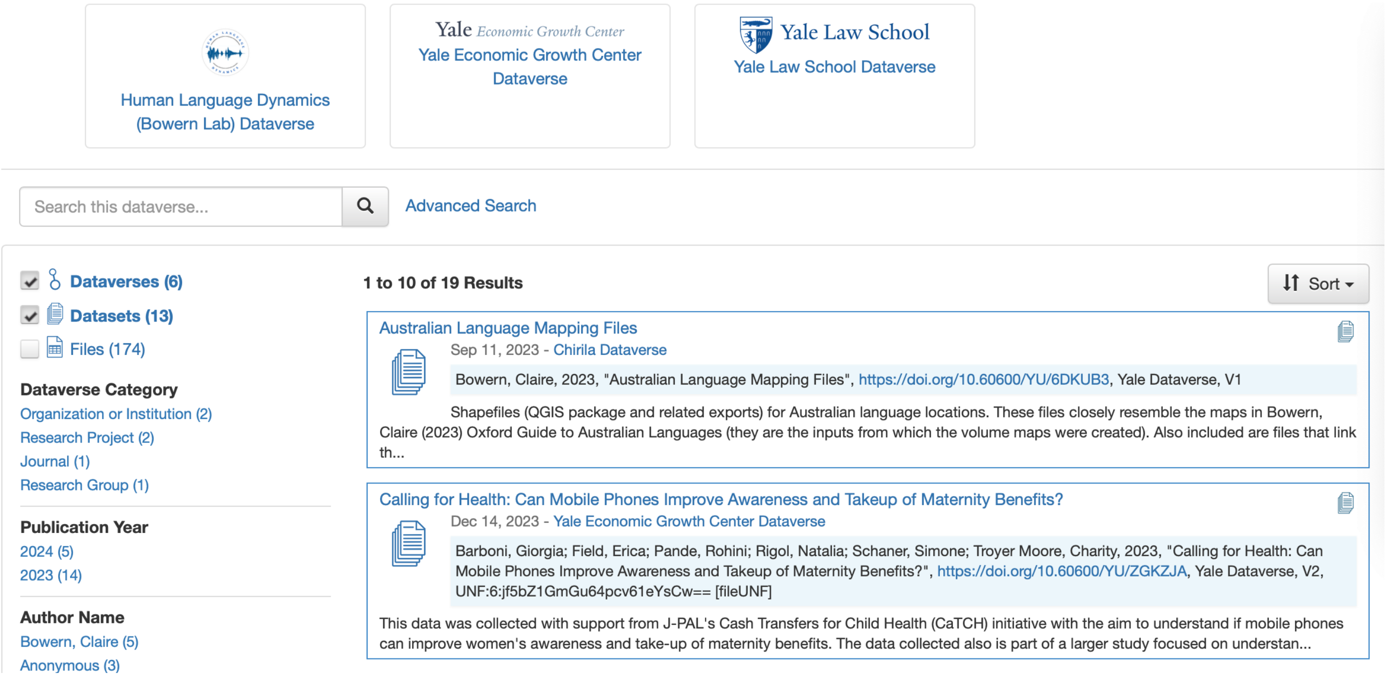
Yale Dataverse

dataverse.yale.edu

Limor Peer, Institute for Social and Policy Studies

The Dataverse® Project

- Yale Library is working with DISSC to host and manage the Yale Dataverse, a data repository to share, preserve and cite research data



Human Language Dynamics (Bowern Lab) Dataverse

Yale Economic Growth Center
Yale Economic Growth Center Dataverse

Yale Law School
Yale Law School Dataverse

Search this dataverse... | Advanced Search

☑ Dataverses (6)
☑ Datasets (13)
□ Files (174)

Dataverse Category
Organization or Institution (2)
Research Project (2)
Journal (1)
Research Group (1)

Publication Year
2024 (5)
2023 (14)

Author Name
Bowern, Claire (5)
Anonymous (3)

1 to 10 of 19 Results

Sort

Australian Language Mapping Files
Sep 11, 2023 - Chirila Dataverse
Bowern, Claire, 2023, "Australian Language Mapping Files", https://doi.org/10.60600/YU/6DKUB3, Yale Dataverse, V1
Shapefiles (QGIS package and related exports) for Australian language locations. These files closely resemble the maps in Bowern, Claire (2023) Oxford Guide to Australian Languages (they are the inputs from which the volume maps were created). Also included are files that link th...

Calling for Health: Can Mobile Phones Improve Awareness and Takeup of Maternity Benefits?
Dec 14, 2023 - Yale Economic Growth Center Dataverse
Barboni, Giorgia; Field, Erica; Pande, Rohini; Rigol, Natalia; Schaner, Simone; Troyer Moore, Charity, 2023, "Calling for Health: Can Mobile Phones Improve Awareness and Takeup of Maternity Benefits?", https://doi.org/10.60600/YU/ZGKZJA, Yale Dataverse, V2, UNF:6:jf5bZ1GmGu64pcv61eYsCw== [fileUNF]
This data was collected with support from J-PAL's Cash Transfers for Child Health (CaTCH) initiative with the aim to understand if mobile phones can improve women's awareness and take-up of maternity benefits. The data collected also is part of a larger study focused on understan...

Yale Dataverse

dataverse.yale.edu

The Dataverse® Project

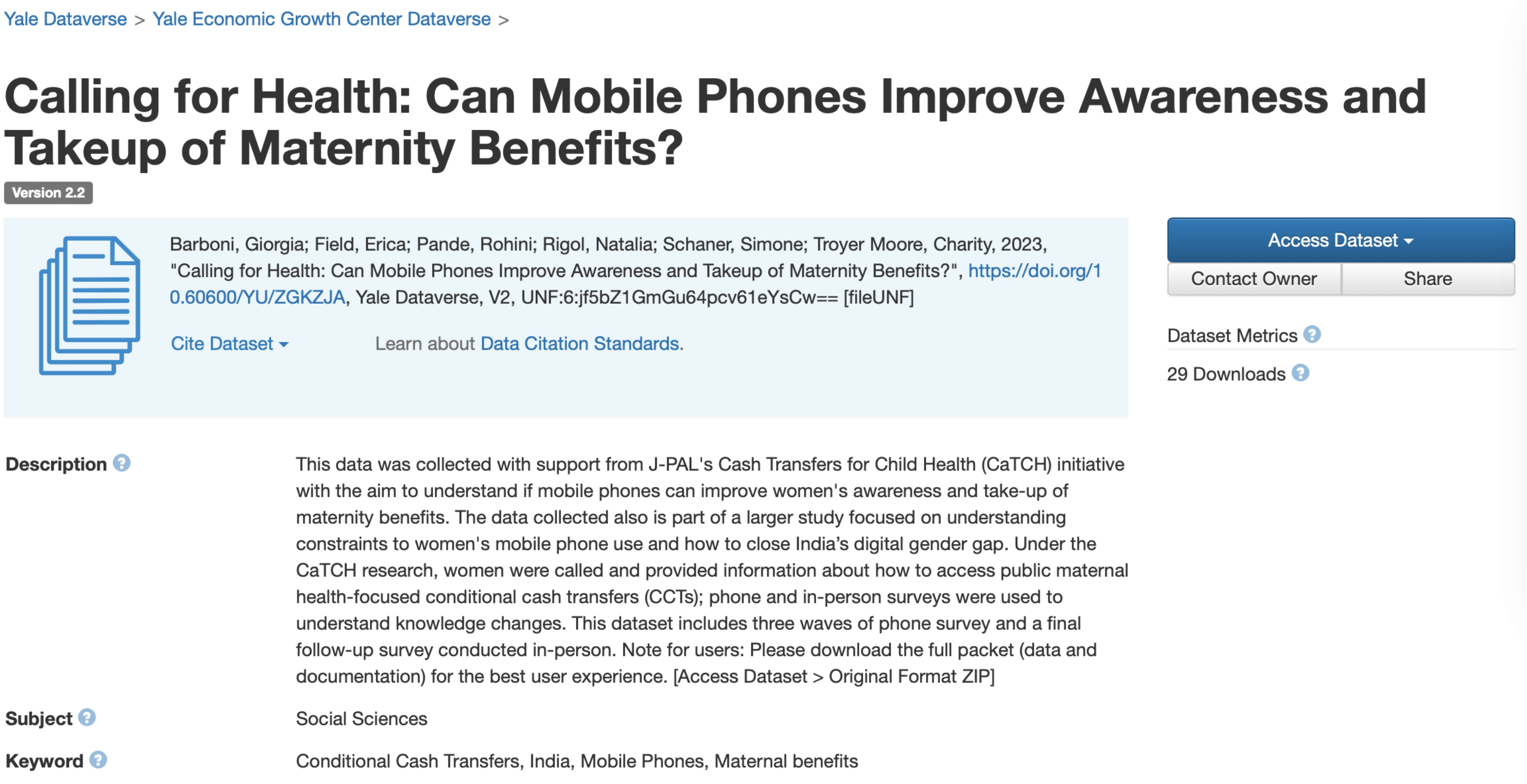
- Why a research data repository is necessary:
  - The Library spends significant funds on datasets, but researchers cannot always easily find them
  - The Library catalog is not sufficiently flexible to capture needed metadata for datasets
  - Discipline-specific repositories are not available to all researchers, and there are often gaps in what types of data they accept
  - Data generated by researchers are at risk of being lost or useless without metadata
  - Datasets are increasingly large, meaning it is difficult to move them around
  - Researchers need to connect compute to data without having to download data to their local devices

Yale Dataverse

dataverse.yale.edu

The Dataverse® Project

- Proof of concept phase (current):
  - A handful of faculty “curators” in social science departments are depositing datasets and making recommendations for changes and additional features
  - Limited to data that are meant to be public with a CC0 license



Yale Dataverse > Yale Economic Growth Center Dataverse >

Calling for Health: Can Mobile Phones Improve Awareness and Takeup of Maternity Benefits?

Version 2.2

Barboni, Giorgia; Field, Erica; Pande, Rohini; Rigol, Natalia; Schaner, Simone; Troyer Moore, Charity, 2023, "Calling for Health: Can Mobile Phones Improve Awareness and Takeup of Maternity Benefits?", https://doi.org/10.60600/YU/ZGKZJA, Yale Dataverse, V2, UNF:6:jf5bZ1GmGu64pcv61eYsCw== [fileUNF]

Learn about Data Citation Standards.

Dataset Metrics

29 Downloads

Description This data was collected with support from J-PAL's Cash Transfers for Child Health (CaTCH) initiative with the aim to understand if mobile phones can improve women's awareness and take-up of maternity benefits. The data collected also is part of a larger study focused on understanding constraints to women's mobile phone use and how to close India's digital gender gap. Under the CaTCH research, women were called and provided information about how to access public maternal health-focused conditional cash transfers (CCTs); phone and in-person surveys were used to understand knowledge changes. This dataset includes three waves of phone survey and a final follow-up survey conducted in-person. Note for users: Please download the full packet (data and documentation) for the best user experience. [Access Dataset > Original Format ZIP]

Subject Social Sciences

Keyword Conditional Cash Transfers, India, Mobile Phones, Maternal benefits

Example Dataverse Dataset Record

Yale Dataverse

dataverse.yale.edu

The Dataverse® Project

- Proof of concept phase (current):
  - A handful of faculty "curators" in social science departments are depositing datasets and making recommendations for changes and additional features
  - Limited to data that are meant to be public with a CC0 license
- Phase 1 (May 2024):
  - Broader recruitment for users across social science departments – at least 10 additional faculty
  - Continue to focus on CC0 data
  - Explore additional functionality on our test instance
- Phase 2 (Fall 2024):
  - Dataverse open to the entire Yale research community

Yale Dataverse

dataverse.yale.edu

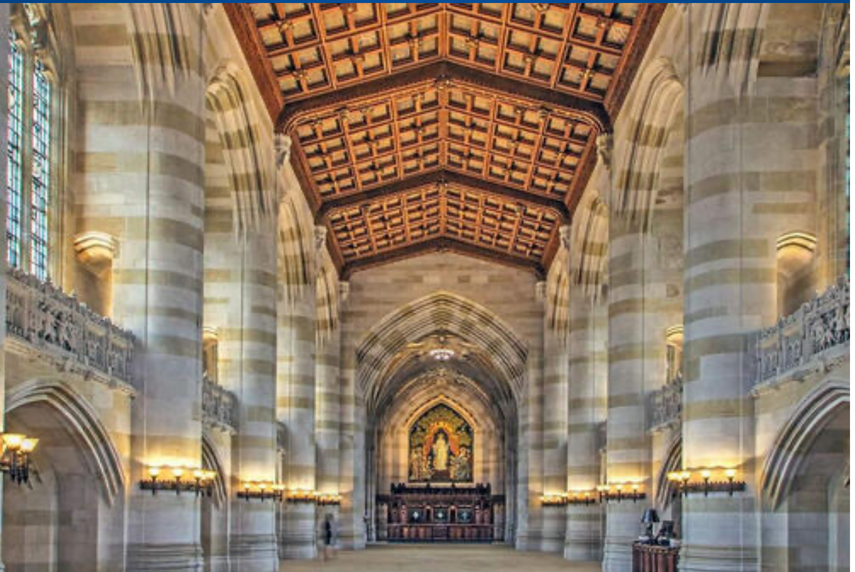
The Dataverse® Project

- Opening Dataverse up to broader use requires us to consider additional needs:
    - Some data are restricted access, temporarily or permanently
    - Some biomedical data need to comply with HIPAA regulations
    - Some data cannot be downloaded but need ready access to advanced computing
    - Some datasets are very large, or have many very small files
    - We need to be able to remove licensed data if license is not renewed

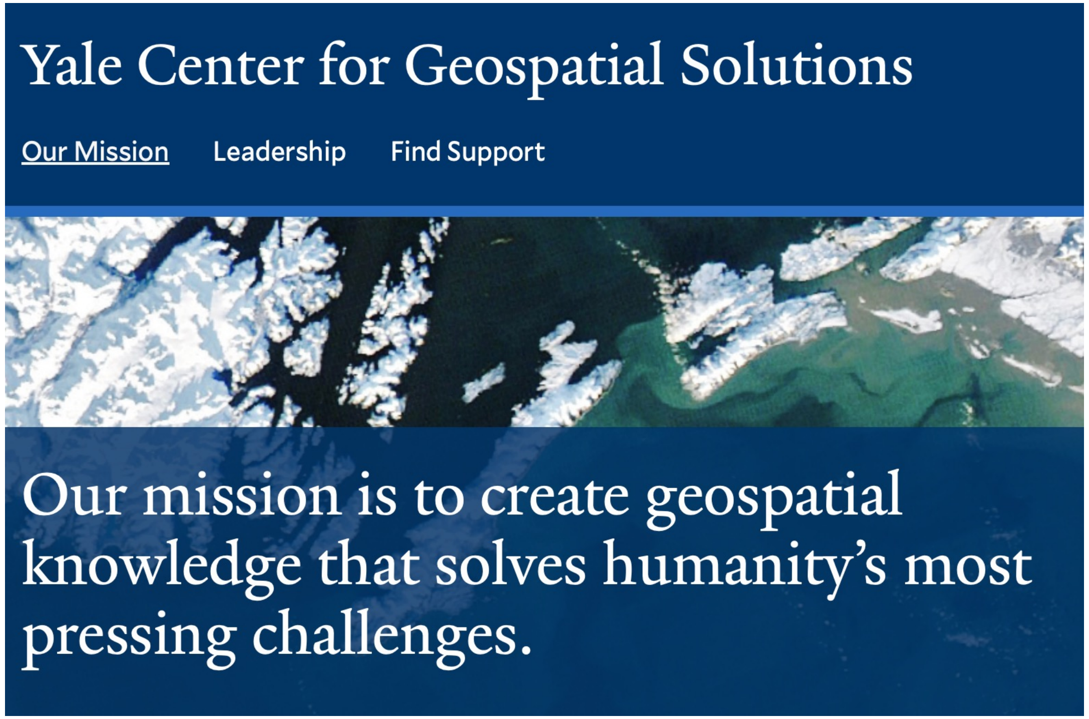
globus

Y | CRC

Next steps as we define services



- Launch additional collaborations:
  - Yale Center for Geospatial Solutions
    - Embedded Library staff
    - New Geospatial Scientist position



Yale Center for Geospatial Solutions

Our Mission Leadership Find Support

Our mission is to create geospatial knowledge that solves humanity's most pressing challenges.

Next steps as we define services

- Develop resources for patrons for whom the term “data” does not resonate:
  - Work toward ensuring Library web presence is responsive to the terms researchers use to describe their work and support needs
  - Develop new workshops specifically for undergraduates and early graduate students on research design and organization
  - Work with the Beinecke to identify courses using special collections and their data and develop data-focused content for faculty
  - Document use cases intentionally to ensure we do not always use examples from traditionally data-intensive disciplines

Next steps as we define services

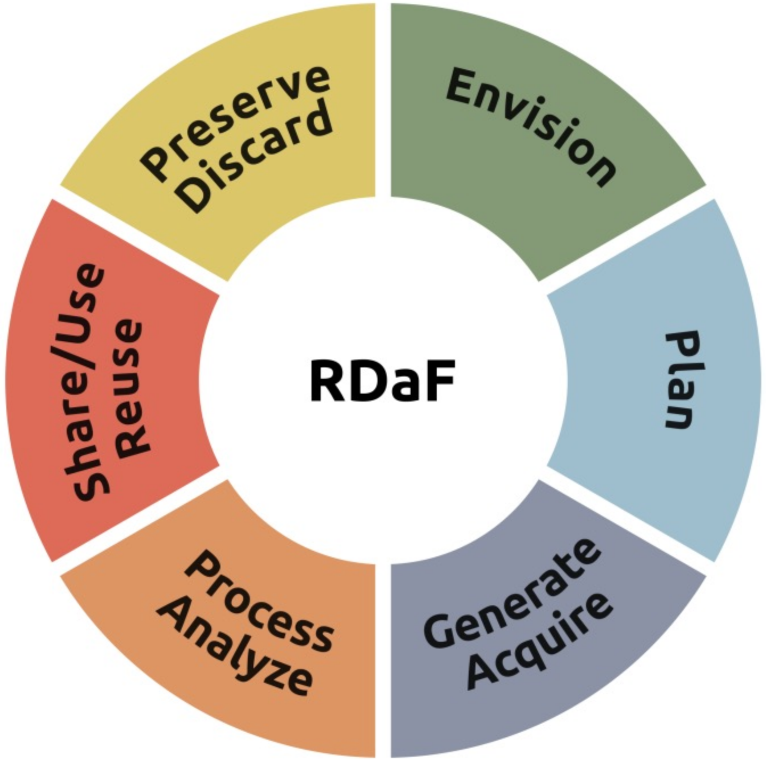
- Develop a new framework for collaboration: we are thinking about the best balance for our staff collaborating on custom software and solutions with the challenges that come with hosting digital projects long-term
  - Work more closely with Library IT to document the most sustainable ways for researchers to go about digital projects
  - Work to identify common needs, e.g. dashboards, web hosting, and pathways to finding the right resources on campus
  - Normalize and encourage planning in advance!

Next steps as we define services

- What role do we play in enforcing data use agreements that are increasingly restrictive in terms of AI?
- What role do we play in the equitable access to compute across campus?
  - Yale Center for Research Computing operates on a “PI” model that many of our patrons do not easily fit into
  - Many users may need more of a “mid-size,” interactive computing environment
  - What kind of instruction from our team is most impactful?
    - Users can learn R anywhere, but where are they learning how to design a research project that uses data and computational methods?
    - Where are the gaps in formal courses?
    - How can we provide training in professional skills that are applicable to students regardless or major or career plans

Next steps as we define services

- Research data governance and storage are hot topics – how can we partner with central IT to tackle these and ensure researcher needs are met?



Broader impacts

- How can we make sure the work we do serves the broader community:
  - We are exploring the possibility of internship programs for first-generation college students
  - We are exploring partnerships with smaller institutions
  - Instructional materials and documentation will be made available to the broader community
  - Promoting open access data and helping researchers describe and share their data means it is more available to everyone

Thank you

- New job announcements coming soon!
- Keep in touch: rebecca.dikow@yale.edu

Yale UNIVERSITY LIBRARY