



Saving Data, Sustaining Discovery

The Power of Research Data Curation

Sonia Barbosa

Associate Director, Dataverse Support, Data Curation and The Murray Archive

Visionary Perspectives on Data Rescue



"The work that goes into Data Rescue builds trust"



"Repositories are Infrastructure for future innovation"

”

"We want people to recognize that this (data rescue) is a public good, just like roads and bridges and other kinds of infrastructure."

— Lynda Kellam, PhD

Founding member of the Data Rescue Project



We Rescue Data Because We Want New Discoveries!

better data → better questions → better
results

A Data Nightmare



The Reality of Irretrievable Knowledge

Imagine finding the exact dataset you need, only to discover:

- Half of the files are **missing**.
- Insufficient documentation makes the data **unusable**.
- The data creator has retired and suffers from **memory loss**.

"I have sought depositors in numerous situations, but some are simply not resolvable."

The Cost and Real-World Impact of Data Loss



Invisible but Expensive

The true impact of lost research is often hidden beneath the surface, manifesting as systemic inefficiencies and sunk costs.



Wasted Resources

Time and money wasted re-collecting data that was previously available but has been lost due to **poor curation**.



Stalled Innovation

Missed chances to validate or extend earlier work, slowing down the overall pace of discovery and progress.



Scientific Irreproducibility

A long-term study that can't be replicated because raw data are gone, stalling scientific progress and wasting years of research effort.

Historical Data Loss: Lost Before Curation



Pre-Digital Collections: The Missing Ledger of the 20th Century

Substantial portions of mid-20th century research (paper, punch cards, magnetic tapes) were generated before **preservation standards existed**. This vital knowledge base is largely lost or fragmented.

- **Physical Media Decay** (brittleness, magnetic demagnetization)
- **Obsolete Formats** (forgotten coding schemes)
- **Fragmented/Lost Context**



Documented Data Disappearances & Historical 'Gaps'

We often see only the 'saves.' Some of our most fundamental data was only partially recovered, or entirely vanished, creating historical gaps.

- **Early US Census (1960)** (Major magnetic tape degradation; partial)
- **Historical Longitudinal Studies** (Inaccessible raw formats)
- **Scientific Observational Data**

***We're sitting on data we can't find, can't trust,
or can't use—and that's throttling discovery!***



Do you Rescue Data?

Rescuing Data: Chronology of Repositories

Mapping the evolution from isolated data to global interoperable networks

ERA 1



Pre-Digital / 1990s

Isolated & at-risk collections. Legacy formats, inaccessible storage, and limited sharing.

Repositories

<50

Datasets

<10k

ERA 2



Digital Dawn (1990-2005)

First curation efforts. Archive creation, metadata standards, and internet-based sharing.

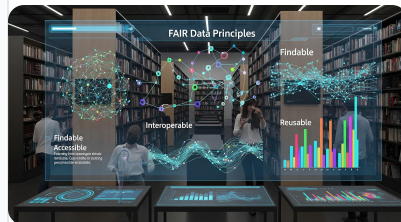
Repositories

~250

Datasets

100k

ERA 3



FAIR Modernization (2005-2018)

Open data mandates. Accessible, findable data with persistent identifiers.

Repositories

1,500+

Datasets

10m+

ERA 4



Global Integration (2018-Present)

Interoperability and AI. Cross-search, automated curation, and federated networks.

Status

Ongoing

Focus

Discovery & Reuse

Era 4: Global Integration (2018–Present)

Building a seamless, automated, and federated data ecosystem

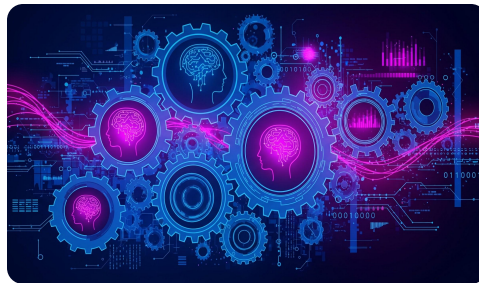
NETWORK



System Evolution

Shift toward machine-actionable DMPs, PID graph integration, and automated FAIR assessment tools to ensure data is discovery-ready.

STRATEGY



Ecosystem Strategy

Focus on cross-domain data sharing, automated provenance, and interoperable infrastructure standards across global research communities.

IMPACT



2025 Forecast

Repositories
>2,000 Global
Datasets
~5M Rescued

What “Rescuing Data” Really Means

The Definition (Goal: More than just a backup)

“Rescuing data means taking vulnerable, valuable data and making sure we can actually find it, understand it, and use it in the future.”

1. Find

Identify at-risk data residing in old systems, personal drives, or isolated locations “Sonia’s basement.”

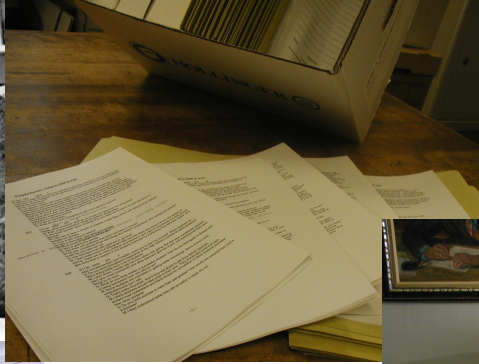
2. Stabilize

Move and standardize formats while adding basic metadata to ensure the data remains intelligible.

3. Prioritize

Focus on unique, reusable assets that align with the mission to maximize the value of rescue efforts.

The Murray Archive: Social Science Data at Risk



From Rescue to Discovery

1. Power the Future

We need our past to fuel what's next.

AI and analytics need **good, long-running data** to learn.

Breakthroughs build on **well-kept records** of the past.

2. New Possibilities

Saving data opens doors to discovery.

Spot **patterns no one saw before** using new tools.

Forecast, manage risk, and **make better decisions**.

3. Active Discovery

It's about more than just storing files.

Clear rules on ownership, editing, and usage rights.

A **culture of trust** where people build on existing work.

The Human & Organizational Side

The most important, time and cost consuming, and most overlooked aspect of data rescue.

The Problem: Silos

Data dies in silos and handoffs. Projects end, people move, and systems change—leading to orphaned data.

Misaligned Incentives

Incentives often work against rescue. Rewards are for new output, not for "cleaning up" or documenting old data ← **That's us (data curators, data Stewards)**

The Solution: Normalization

Make rescue part of normal work.

- Default workflows for depositing data at project end.
- Clear roles (owners, stewards, IT, legal).

Data Rescue is a Team Sport

Data rescue isn't just about saving files—it's a massive, collaborative effort across disciplines and roles.

It requires the combined expertise of:

- Data Stewards & Curators
- Researchers & Scientists
- IT Professionals & Librarians
- Legal & Policy Experts

Collaborating to secure the foundation of our research heritage.

The Pioneers: 1970s – 1990s

Establishing the Foundations of Research Data Archiving



ICPSR

Inter-university Consortium for Political and Social Research

- Central hub for social science data since the early era.
- Pioneered standardized data documentation.
- Focused on large-scale survey data and census records.



UK Data Archive

United Kingdom's Primary Resource

- Leading European center for social and economic data.
- Strong emphasis on longitudinal study preservation.
- Established rigorous acquisition and curation protocols.



Murray Archive

Henry A. Murray Research Archive

- Founded in 1976 at Radcliffe College.
- Specialized in qualitative and longitudinal data.
- Focus on women's lives and social change studies.

Repository Pioneers: 1990 – 2005

Establishing Digital Infrastructure and Interoperability.



The Internet Archive (1996)
Foundational digital preservation.

- Pioneered large-scale web archiving.
- Home of the "Wayback Machine".
- Collects cultural and historical web data.



arXiv.org (1991)
Pioneered open access preprint archiving.

- Pioneered open access preprint archiving.
- Validated digital self-archiving for physics/math.
- Championed early online scientific communication.
- Established standards for metadata exchange (OAI-PMH).



DSpace (2002)
Developed by MIT and HP.

- First widely adopted open-source institutional repository platform.
- Validated digital curation workflows.
- Standardized digital asset management.


Repository Pioneers: 2005 – Present

Establishing the Era of FAIR Data and Global Interoperability.



Dataverse Project (2006)

- Multi-tenant, institutional repository software.
- Supports complex metadata and data citation.
- Open-source, global adoption.




Dryad (2009)

- Curation focused repository for scientific literature data.
- Strong integration with journal publishing workflows.
- Mandates data sharing.



Zenodo (2013)

- Multidisciplinary and open access repository (CERN).
- Assigns digital object identifiers (DOIs).
- Accommodates large-scale data and software.



FAIR Data Principles (2016)


- *The definitive standard* for data management.
- Integrated into funding and institutional policies.
- Shifted focus from simple storage to structured reuse.



figshare (2011)
OSF (2011)
elixir biology
humanitarian archive

The Evolving Ecosystem

- Rise of generalist and discipline-specific platforms.
- Increased interoperability and networked systems.
- Emphasis on data reuse, transparency, and collaboration.



Data Preservation Alliance for the Social Sciences (Data-PASS)

A voluntary partnership to acquire, archive, and preserve at-risk digital social science data. Originally funded by the Library of Congress, ensuring accessibility for future researchers.

Key Preservation Efforts

- **Data Rescue:** Identifying and acquiring legacy data like opinion polls and voting records.
- **Shared Cataloging:** Federated access across multiple institutional repositories.
- **Ethics:** Strict subject confidentiality and de-identification protocols.
- **Future Planning:** Establishing management guidelines for ongoing research.

Institutional Partners

Major data repositories and academic institutions:

- ICPSR, University of Michigan
- Roper Center for Public Opinion Research
- The Odum Institute, UNC
- The Murray Research Archive, Harvard
- National Archives (NARA)

Data-Pass: Digital Preservation

Building a Resilient Network

The NDIIPP strategy, hosted by the **Library of Congress**, invested **\$30 million** in grants and partnerships across 320+ institutions to seed a thriving digital preservation community.

Key Transitioned Programs:

- **NDSA:** Now at the Digital Library Federation (since 2016).
- **DPOE:** Now hosted by the Pratt Institute.
- **NDSR:** Follow at nds-r-program.org.

Note: While NDIIPP is no longer active at the Library of Congress, its legacy thrives in these mature communities.

The screenshot shows the Data-PASS project page on the Library of Congress website. The page header includes the Library of Congress logo and navigation buttons for 'ASK A LIBRARIAN', 'DIGITAL COLLECTIONS', and 'LIBRARY CATALOGS'. A search bar is located in the top right corner. The breadcrumb trail reads: 'The Library of Congress > Digital Preservation > Partners > Data-PASS'. The main content area features a 'DIGITAL PRESERVATION' banner with a 'JEDI advisory board group - Closed' status. Below this is a search box for the site and a list of navigation links: Home, About, Meetings & Events, and Education & Training. A 'Resources' section lists links to 'Digital Formats Sustainability', 'Federal Agencies Digitization Guidelines Initiative', 'Library of Congress Recommended Format Specifications', and 'Project Web site'. The 'Data-PASS' title is followed by 'Lead Partner: ICPSR (Inter-university Consortium for Political and Social Research), University of Michigan' and 'Project Dates: 2004-2010'. An 'Additional Partners' section states that a complete listing can be found at the 'partners viewshare' link. A paragraph describes the project's goal: 'Data-PASS has acquired and preserved social science data at risk of being lost to the research community, including opinion polls, voting records, large-scale surveys and other social science studies. While this information provides the full story of the social and cultural experience of America, a huge quantity of this data is missing or at-risk. Significant data collections that had not been deposited in a permanent archive have been identified and acquired by the partner institutions. Through NDIIPP funding, these data have been preserved and made available through a shared catalog.' Below this, it says 'More detailed project information can be found at the Project Web site'. A 'Highlights' section lists several documents: 'Data-PASS Final Report to NDIIPP (2011) (PDF)', 'Data-PASS Dataverse Shared Catalog', 'Data-PASS Metadata Requirements (2007) (PDF, 73KB)', 'Paper: Building Relationships: "A Foundation for Digital Archives" (2008) (PDF, 31KB)', 'Data-PASS Data Deposit Agreement (2006) (PDF, 74KB)', and 'Resource: Archived project web site'. A 'Back to top' link is at the bottom right. The footer contains 'Connect with the Library' with social media icons, 'Subscribe & Comment' with RSS and E-Mail options, 'Download & Play' with Podcasts, Webcasts, and iTunesU, and 'Questions' with 'Ask a Librarian' and 'Contact Us'. The bottom of the page includes links for 'About | Press | Jobs | Donate', 'Inspector General | Legal | Accessibility | External Link Disclaimer | USA.gov', and 'Speech Enabled'.

Data Refuge 2016

Origins & Growth

Launched November 2016 in Philadelphia to protect federal environmental data from climate denial.

Spread to 50+ cities with thousands of civic partners.

Community Action

- 50+ Data Rescue events held by libraries.
- Attracted major media attention and open gov initiatives.
- Cross-sector collaboration (universities, journalists).

The Data Ecosystem Vision

Data lives in a complex ecosystem shaped by human and nonhuman forces. Like living organisms, data must be cultivated, documented, and protected from neglect or deletion.

Building a storybank to document how data connects people, places, and species.

Collaborating to secure the foundation of our research heritage.

The Data Rescue Project

Preserving At-Risk Research Data for Future Discovery

Current Efforts & Tracker

Monitoring and rescuing critical federal data from CDC, EPA, and HRSA to ensure public access remains uninterrupted.

- **Data Rescue Tracker:** Active monitoring of data availability.
- **Community Support:** In collaboration with IASSIST, RDAP, and the Data Curation Network.



Visit www.datarescueproject.org for more information.

“Things were going dark left and right”: the race to save US government datasets before they’re deleted

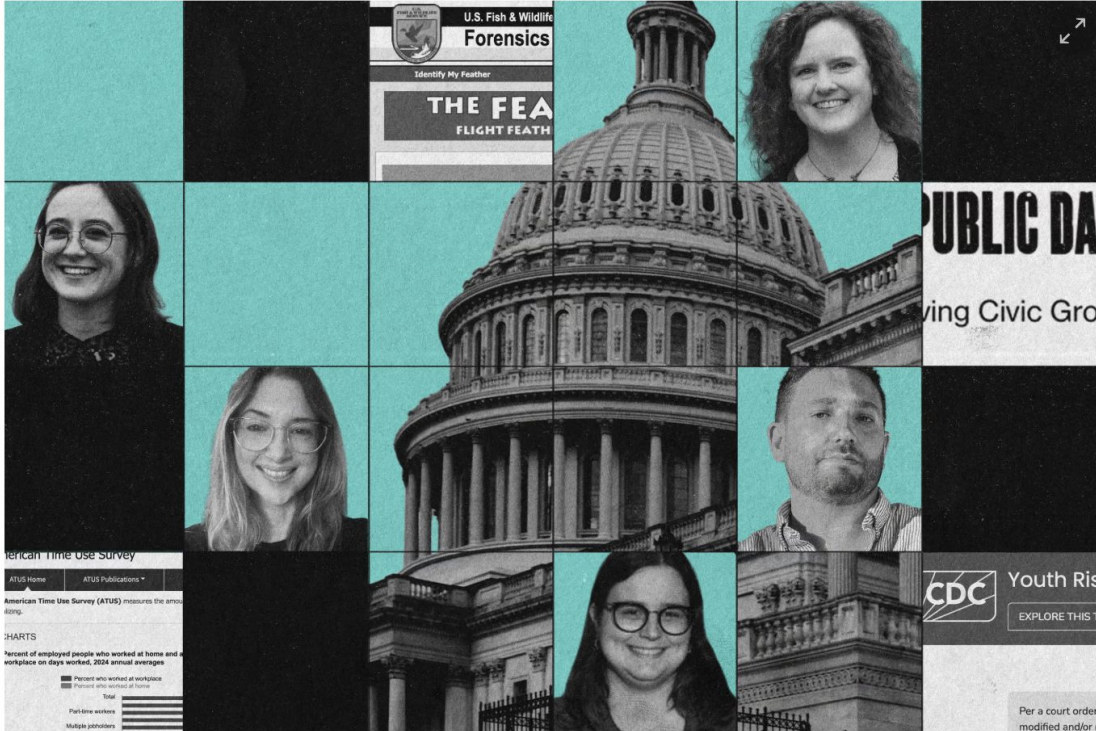


Illustration: Guardian Design/Getty Images; Data Rescue Project

Group has banded together to rescue data as Trump administration has removed or altered data on climate change, reproductive health, LGBTQ+ people and more

Most viewed

<https://www.theguardian.com/us-news/2026/may/07/save-government-datasets-deleted>

Data Rescue at Harvard Dataverse

- **2017 Rescue Efforts - EPA data:**

28 datasets and over **2k files** were rescued into the Harvard Dataverse Repository in collaboration with the Harvard Dataverse repository curation team and Harvard Library staff.

HARVARD
Dataverse

Add Data ▾ Search ▾ User Guide Support Sign Up Log In

DataRefuge
DataRefuge Dataverse
(Harvard University)

Harvard Dataverse > CAFE Research Coordinating Center for Health and Extreme Weather Collection > Extracted Data Contributions >

Contact Share

Data Refuge is a public, collaborative project designed to address the following concerns about federal climate and environmental data:

- What are the best ways to safeguard data?
- How do federal agencies play crucial roles in data collection, management, and distribution?
- How do government priorities impact data's accessibility?
- Which projects and research fields depend on federal data?
- Which data sets are of value to research and local communities, and why?

DataRefuge is also an initiative committed to identifying, assessing, prioritizing, securing, and distributing reliable copies of federal climate and environmental data so that it remains available to researchers. Data collected as part of the DataRefuge initiative will be shared in multiple formats to help ensure continued accessibility.

Read full Description [+]

United States Bureau of Economic Analysis
Bureau of Justice Statistics (BJS)
Bureau of Labor Statistics
Centers for Disease Control and Prevention (CDC)

Search this dataverse... Advanced Search Add Data

Dataverses (28)
 Datasets (48)
 Files (2,307)

Dataverse Category
Organization or Institution (27)
Department (1)

1 to 10 of 76 Results

Alternative Fuels Data Center
Apr 22, 2025 - United States Department of Energy
Alternative Fuels Data Center, 2025, "Alternative Fuels Data Center", <https://doi.org/10.7910/DVN/NZGZZI>, Harvard Dataverse, V1
Data related to alternative fuels and advanced vehicles.

HARVARD
Dataverse

Add Data Search User Guide Support Sign Up Log In

CAFE

CAFE Research Coordinating Center for Health and Extreme Weather Collection

(Harvard University, Boston University)

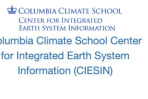
Harvard Dataverse >

Contact Share


Welcome to the CAFE Dataverse collection! This open collection is designed to support and enhance global research initiatives focused on understanding and mitigating the health impacts of environmental exposures. More information about CAFE's data management can be found on our [CAFE documentation website](#). To learn more about CAFE, please visit our [homepage](#).

Instructions


To make contributions to the CAFE collection, you'll find easy-to-follow steps in our [CAFE Dataverse upload instructions](#). Before publishing your dataset, please ensure the metadata adheres to the [CAFE Recommended README Template](#), and familiarize yourself with the [Dataverse Terms of Use](#).




Columbia Climate School Center for Integrated Earth System Information (CIESIN)



DesignSafe Data Depot Repository Harvested Subcollection



Connecting Health Outcomes Research and Data Systems (CHORDS)



Climate Health AIR pollution (CHAIR) in India

Search this dataverse... Advanced Search


Datasets (50) Datasets (1,241) Files (44,296)

1 to 10 of 1,291 Results

Extracted Data From: EPA Interactive Map of Air Quality Monitors
May 20, 2026 - Extracted Data Contributions

HARVARD
Dataverse

Add Data Search User Guide Support Sign Up Log In



Socioeconomic Data and Applications Center (SEDAC)
(Columbia University)

Datasets from the former NASA Socioeconomic Data and Applications Center (SEDAC)

Harvard Dataverse > CAFE Research Coordinating Center for Health and Extreme Weather Collection > Columbia Climate School Center for Integrated Earth System Information (CIESIN) >

Contact Share

The SEDAC collection comprises 300 datasets that were curated or developed from 1998-2025 under CIESIN's contract to run the NASA Socioeconomic Data and Applications Center (SEDAC). The data in this collection are of high relevance to researchers focused on human-environment interactions as well as for health researchers and practitioners.

Some of the data for this project are stored on the [Northeast Storage Exchange \(NESE\)](#). Follow the instructions for large data download found on our website: [Downloading data from NESE via Globus: Quick Start](#)

Search this dataverse... Advanced Search

Datasets (0) Datasets (244) Files (7,948)

1 to 10 of 244 Results Sort

Global Rural-Urban Mapping Project, Version 1 (GRUMPv1): Population Count Grid
Apr 14, 2026

The CAFE Research Coordinating Center for Health and Extreme Weather (RCC) is a joint initiative by the Boston University School of Public Health and Harvard T.H. Chan School of Public Health (**click on the images to navigate to the collection**).

Extracted Data Use Case



This use case highlights ways researchers can use Dataverse repositories to share publicly available extracted datasets for reuse.

DV Features used:
Deposit Templates

Metadata:
Subsets of a "source" dataset

Licensing

Chain of Custody

File Types

Provenance File, if available

Background:

In the context of **data rescue**, extracted data sharing involves identifying, recovering, and reformatting valuable information from at-risk or inaccessible sources—such as outdated systems, legacy formats, or deteriorating media—and making curated subsets of that data available for reuse. These extracted datasets are often cleaned, restructured, and documented to preserve their utility and historical value. Sharing them ensures that important data—especially from government, scientific, or community sources—is not lost and can continue to support research, policy, and public knowledge. Careful attention is given to licensing, provenance, and ethical considerations, particularly when the rescued data involve human subjects or sensitive content.

Additional Considerations:

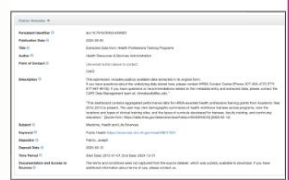
Incomplete or unknown licensing: You may not know the original terms. When in doubt, use restrictive or custom licensing, and consult legal/ethics experts if possible.

Unclear consent or ethics coverage: If human subjects data are involved, and original consent is missing or outdated, consider restricted access or consult IRBs.

Fragile or undocumented formats: Rescued data may need extensive reformatting and contextual documentation to ensure usability.

Chain of custody and integrity: Document how the data were obtained and preserved to establish authenticity.

Legacy data literacy: Explain historical terminology, units, or data collection methods that might be unfamiliar to modern users.



The **Harvard Dataverse Repository**, the endowed permanent data sharing repository of Harvard University, is a managed data repository open to all researchers from any discipline, both inside and outside of the Harvard community, where you can share, archive, cite, access, and explore research data. Powered by the **Data Rescue Project**, the repository was released in 2006 by the **Institute for Quantitative Social Science (IQSS)** at Harvard University and is one of the largest repositories of open research data in the world, hosting thousands of datasets across a wide range of disciplines. The repository is free and open to the research community for data sharing and is supported by **Harvard Library** and **Harvard University Information Technology**.



Extracted Data Contributions

Harvard Dataverse > CAFE Climate and Health Research Coordinating Center Collection >

Contact Share

It is important to read all of the information below before depositing data into this collection.

Metadata should be entered using both the **Deposit Template** guidance and the information below. Before beginning, **ensure all deposits are in the public domain** to avoid copyright infringement.

A **Deposit Template** has been added by default to guide the process of completing the metadata information for datasets in this collection. Select the **ADD DATA** and **New Dataset** option on the right-hand side of this page (not the option at the top of the page) when ready to start depositing data.

The purpose of this sub-collection is to store critical climate and health datasets accessible at various locations in one place. Because these datasets are extracted with minimal modification, complete metadata that notes where appropriate citation data can be found is especially important to note.

Additional deposit guidance:

- 1) Subsets of a source dataset:** If the deposit is a subset of the source data (ie: selecting subset of variables, specific geographic area, etc.), this should be noted in the **Title and Description** metadata fields. Include any available metadata, and other information available from the source data website.
- 2) Metadata About Data Sources field:** include all information that is available. Specific fields to include have been noted in the **Deposit Template**
- 3) Licensing:** Licensing determines the possibility of extraction and posting in this sub-collection. Please review any relevant information to ensure data extraction and posting is permitted. By default, this collection includes "CC0" as the license. This can be modified per data deposit and before publishing by clicking on the "TERMS" tab of a DRAFT dataset.
- 4) Adding additional metadata:** After uploading the initial requested information and saving your deposit, click the **Edit Dataset** option on the dataset page to add any additional metadata.
- 5) Submit for Review:** Continue to "add" the DRAFT dataset until you have entered all the available metadata. Once your draft is ready for review, select the "Submit for Review" option on your dataset page's top right-hand corner.

Happy Dataho!

Collapse Description [-]

DataRefuge Dataverse

Search this dataverse... Advanced Search + Add Data

Dataverses (32)
Datasets (164)
Files (13,700)

Dataverse Category
Organization or Institution (30)
Department (1)
Research Group (1)

Publication Year
2025 (186)
2024 (1)
2017 (29)

License
CC BY-NC-SA 4.0 (8)
CC0 1.0 (0)
CC BY-SA 4.0 (18)

Author Name
Environmental Protection Agency (12)
Federal Emergency Management Agency (8)
U.S. Census Bureau (7)

1 to 10 of 218 Results

Extracted Data From: EPA Risk-Screening Environmental Indicators (RSEI), Microdata Disaggregated 2017, 1988-2017 datasets
Nov 13, 2025
US EPA, 2025, "Extracted Data From: EPA Risk-Screening Environmental Indicators (RSEI), Microdata Disaggregated 2017, 1988-2017 datasets", <https://doi.org/10.7927/H73329/00024>, Harvard Dataverse, V1
This submission includes publicly available data extracted in its original form. Please reference the Related Publication listed here for source and citation information. If you have questions about the underlying data stored here, please contact the Environmental Protection Agency using their RSEI Contact Form: <https://www.epa.gov/rseif/forms/contact-us>

Replication Data From: EPA Risk-Screening Environmental Indicators (RSEI), 2022 Water Microdata
Oct 20, 2025
US EPA, 2025, "Replication Data From: EPA Risk-Screening Environmental Indicators (RSEI), 2022 Water Microdata", <https://doi.org/10.7927/H73329/00024>, Harvard Dataverse, V1
This submission includes publicly available data extracted in its original form. Please reference the Related Publication listed here for source and citation information. If you have questions about the underlying data stored here, please contact the EPA at <https://www.epa.gov/rseif/forms/contact-us-about-rsei-model>. If you have questions or comments...

Extracted Data From: PLACES: Local Data for Better Health, Census Tract Data 2023 release
Oct 17, 2025
US Centers for Disease Control and Prevention, 2025, "Extracted Data From: PLACES: Local Data for Better Health, Census Tract Data 2023 release", <https://doi.org/10.7927/H73329/00024>, Harvard Dataverse, V1

Extracted metadata deposit template:

Clear provenance: Cite the original source, even if partially incomplete or obscure.

Document extraction and transformation: Explain how the data were selected, reformatted, or cleaned.

Ensure ethical use: Anonymize sensitive content and align with any known or presumed consent frameworks.

Provide metadata and context: Include data dictionaries, methods, and limitations.

License appropriately: Choose or assign a license based on what's known about rights and permissions.

Constitutional Workflow Metadata

Workflow Type: Custom Instructions: If you are uploading code to an external repository (e.g., GitHub) that describes how your dataset was created, please complete this section. If you are uploading code inside your dataset on Dataverse, please indicate that in the Documentation field below near the URL field.

Metadata About Data Sources: Custom Instructions: PREFILED to "YES" - do not modify

Derived From Another Dataset: Custom Instructions: Please provide the coordinate reference system (CRS) used in the spatial file.

Spatial Resolution: Custom Instructions: If you are uploading geospatial raster data, please complete this section. Otherwise, leave it blank.

Dataset Terms: License/View Agreement: [Our Community Data](#) is not an open access publication and proper credit is given to the creator. Please use the data citation shown on this dataset page. CC BY-NC-SA 4.0

In the context of **data rescue**, extracted data sharing involves identifying, recovering, and reformatting valuable information from at-risk or inaccessible sources—such as outdated systems, legacy formats, or deteriorating media—and making curated subsets of that data available for reuse. These extracted datasets are often cleaned, restructured, and documented to preserve their utility and historical value. Sharing them ensures that important data—especially from government, scientific, or community sources—is not lost and can continue to support research, policy, and public knowledge. Careful attention is given to licensing, provenance, and ethical considerations, particularly when the rescued data involve human subjects or sensitive content.

Harvard Innovation Lab and Harvard Dataverse Repository

Amicus Libris
Briefs from the Harvard Law School Library

New in the Collections / Databases & E-resources

The Data.gov Archive at the Harvard Law School Library Innovation Lab

April 20, 2015 By Jack Cushman

Data.gov Archive

HARVARD LAW SCHOOL LIBRARY | Library Innovation Lab

At the Harvard Law School Library, we have 39 early manuscript copies of Magna Carta, and now we also have over 300,000 public datasets published by the United States federal government.

In February, our Library Innovation Lab launched the Data.gov Archive, a 17-terabyte archive of every dataset published on data.gov by the U.S. federal government. The archive allows our research community — meaning anyone in the world! — to access reliable data whether or not it remains available from its original source.

“This work also shows the extraordinary value of having software engineers embedded within one of the world’s largest law libraries.”

Jack Cushman, Director of Library Innovation Lab

We did this work for the same reason we collected 39 early manuscript copies of the Magna Carta: because it is vital — in planning our laws, our government, and our future — to be able to remember what happened and from where we came.

The Data.gov Archive was created with custom software, developed at the Lab, to download datasets and attach full metadata, cryptographic signatures, and timestamps, which makes the dataset easier and cheaper to host and to share with other archives. This is part of our larger work, such as [Century Scale Storage](#) and [Perma.cc](#), in figuring out how to make digital archives as robust and durable as physical ones so they can last for as long as our Magna Carta copies have.

Explore
ES Seq
In the Community
New in the Collections
Research Type

Search

An official website of the United States government. [Here's how you know](#)

DATA.GOV | DATA | METRICS | OPEN GOVERNMENT | CONTACT

User Guide

The Home of the U.S. Government's Open Data

Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.

364,404 DATASETS AVAILABLE

Search

Most Viewed Datasets | Recently Added Datasets | Datasets by Organization | Geospatial

Mission

The United States Government's open data site is designed to unleash the power of government open data to **inform decisions by the public and policymakers, drive innovation and economic activity, achieve agency missions, and strengthen the foundation of an open and transparent government.**

About Us → | Explore the timeline →

In February, our Library Innovation Lab [launched the Data.gov Archive](#), a 17-terabyte archive of every dataset published on data.gov by the U.S. federal government. The archive allows our research community — meaning anyone in the world! — to access reliable data whether or not it remains available from its original source.

Practical Principles: "Rescue & Sustain"

An actionable, simple 4-step framework for data preservation.

1. Find

Inventory systems, drives, archives, and "shadow data."

"If X person left tomorrow, what data would leave with them?"

2. Fix

- Stabilize formats and locations.
- Add metadata: title, creator, date, method, contacts.

When possible, document discrepancies between the rescued data and any documentation, work with the data creator to ensure maximum data trustworthiness.

3. Feature

- Make data discoverable (catalog, portal, list).
- Highlight "rescued data → big insight" stories.

4. Future-proof

- Define retention and migration rules.
- Automate exports, backups, and checks.

Data Curation & Invisible Work

Source: Boyd, Ceilyn. (2025). [“Is this data?” Investigating How Curators Define, Recognize, and Repair Research Data in Data Repositories](#) [Dissertation].

Research Findings

Curators perform definitional, coordination, and technical work during dataset review and repair, primarily out of view of users.

This work is obscured in two ways:

- Repository workflows hide effort until an anomalous dataset is encountered.
- Infrastructural characteristics and conventions of practice render the work invisible.

The Invisible Curation Process

Curators only become visible when researchers are contacted about unrepairable data.

Three Key Steps (Figure 6.4):

1. **Step 1:** Dataset Review
2. **Step 1.1:** Indicator Evaluation
3. **Step 2:** Data Repair

This is intensive, hidden definitional and operationalization work.

Call to Action: Process Innovation



“Change one process so that every project ends with a **data handoff**, not a data loss.”



Working on data processing in the Murray, circa 1995 and a donated/rescue dataset in April 2026

Resources & References

Institutional Repositories & Projects

- The Harvard Dataverse Repository
<https://dataverse.harvard.edu/>
- The Murray Archive to The Harvard Dataverse: <https://dataverse.org/blog/analog-bytes-transformation-murray-research-archive>
- Data-Pass: <https://www.digitalpreservation.gov/partners/datapass.html>
- Data-Pass (Harvard): <https://gking.harvard.edu/publication/from-preserving-the-past-to-preserving-the-future-the-data-pass-project>

Technical Foundations & Software

- DSPACE: <https://dl.acm.org/doi/10.5555/827140.827151>

Advocacy & Rescue Initiatives

- Saving Government Datasets: <https://www.theguardian.com/us-news/2026/may/07/save-government-datasets-deleted>
- The Data Rescue Project: <https://www.datarescueproject.org/every-data-its-user/>

Publications:

Boyd, Ceilyn. (2025). ["Is this data?" Investigating How Curators Define, Recognize, and Repair Research Data in Data Repositories](#) [Dissertation]. Simmons University.