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.

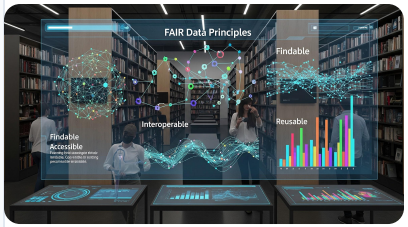
Repositories

~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)

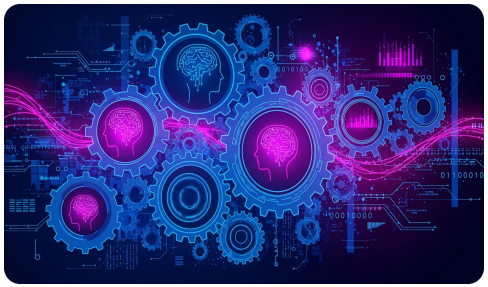
Building a seamless, automated, and federated data ecosystem

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

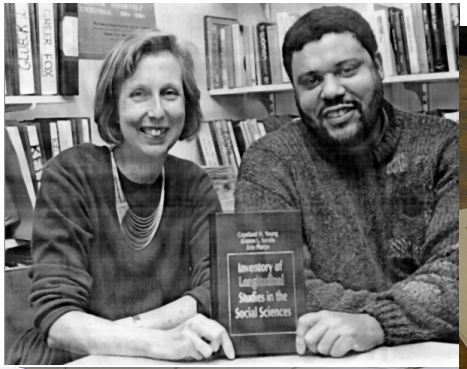
Move and standardize formats while adding basic metadata to ensure the data remains intelligible.

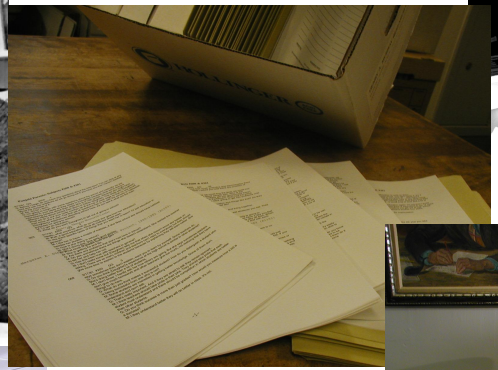
3. Prioritize

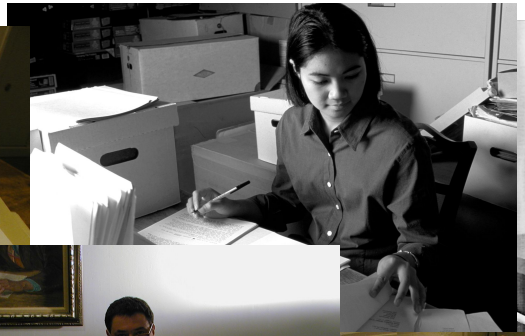
Focus on unique, reusable assets that align with the mission to maximize the value of rescue efforts.

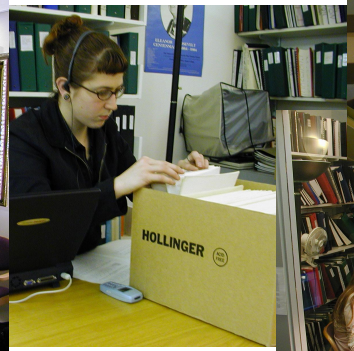
Presentation Time: 5 minutes • Goal: Clarify the scope of data curation

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.

Securing the foundation for future innovation through proactive data stewardship.

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.

Era of Isolated Data & At-Risk Collections

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.

Era of Isolated Data & At-Risk Collections

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

Findable. Accessible. Interoperable. Reusable.

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

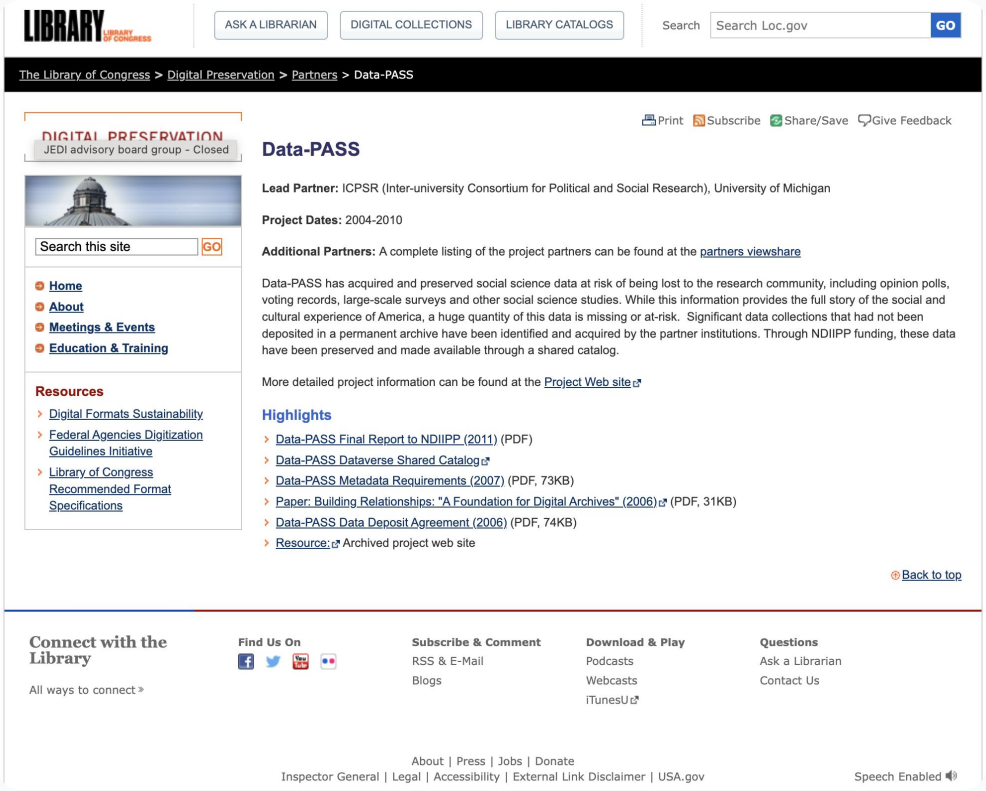
Building a Resilient Network

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 ndsr-program.org.

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



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

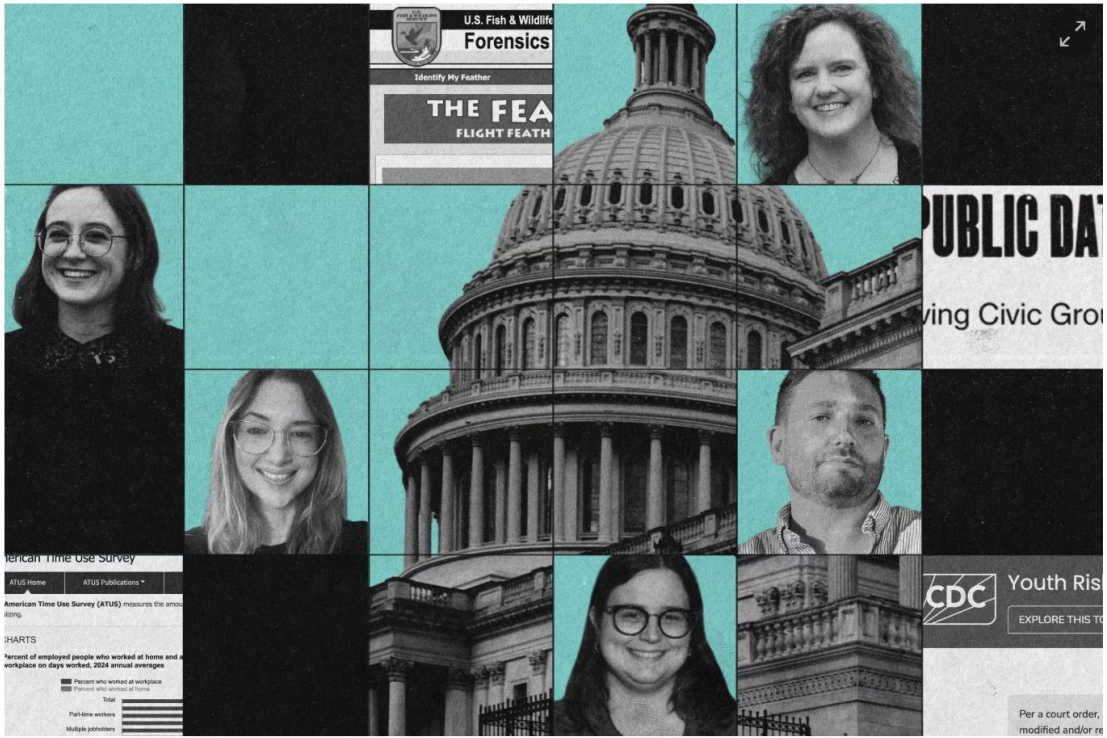


Illustration: Guardian Design/Getty Images; Data Rescue Project

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

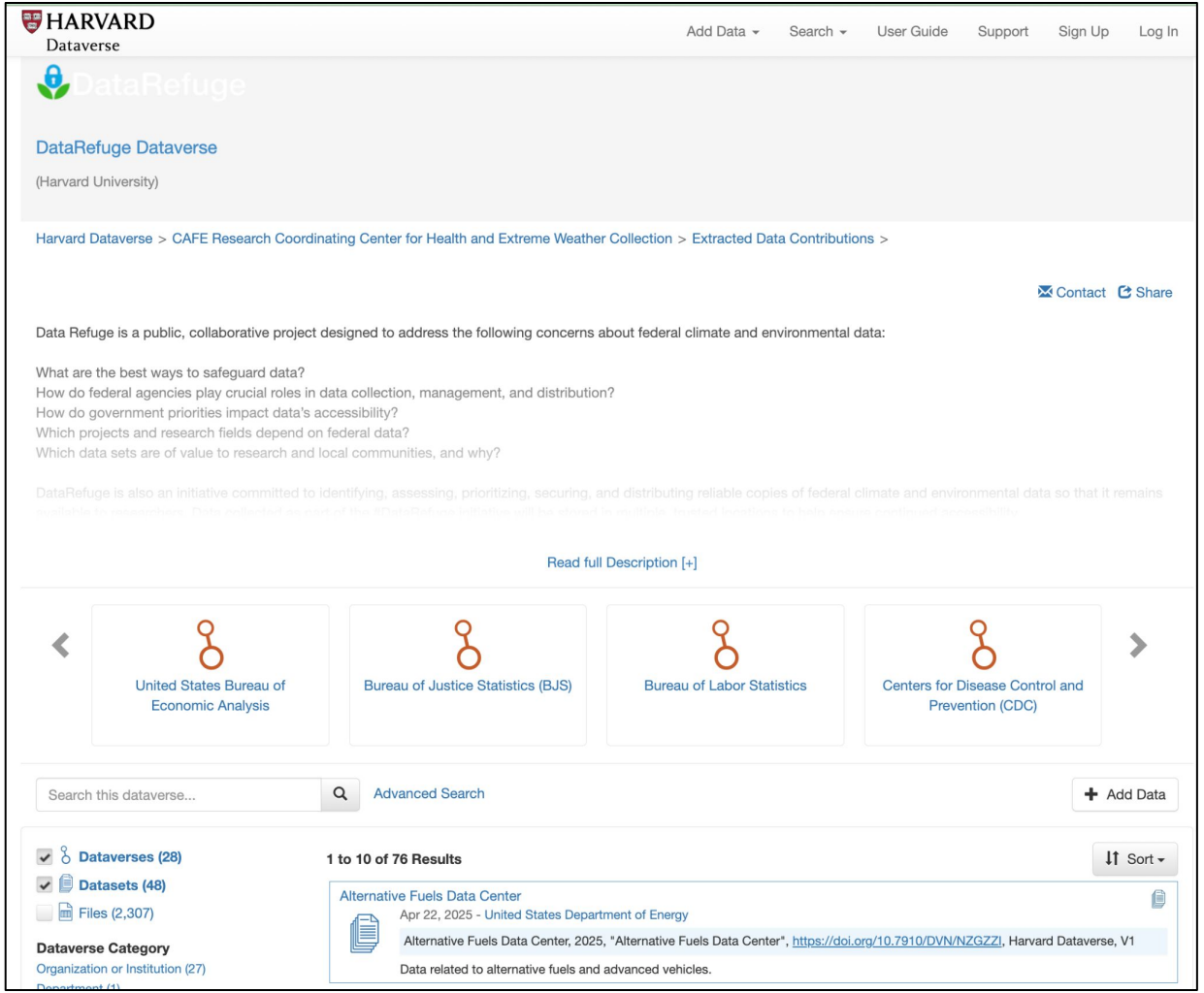
Most viewed

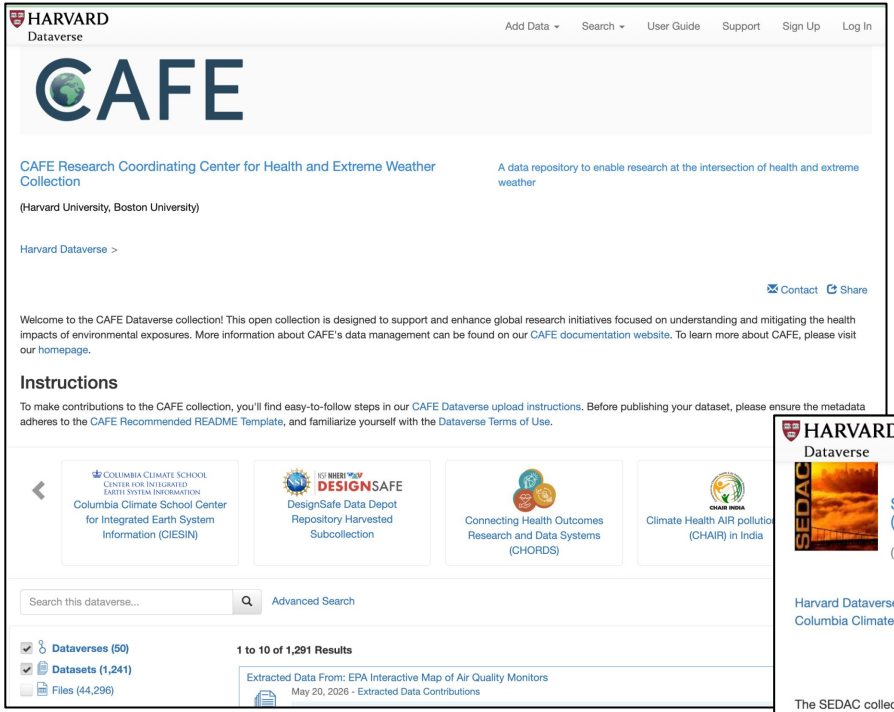
https://www.theguardian.com/us-news/2026/may/07/save-government-datasets-deleted

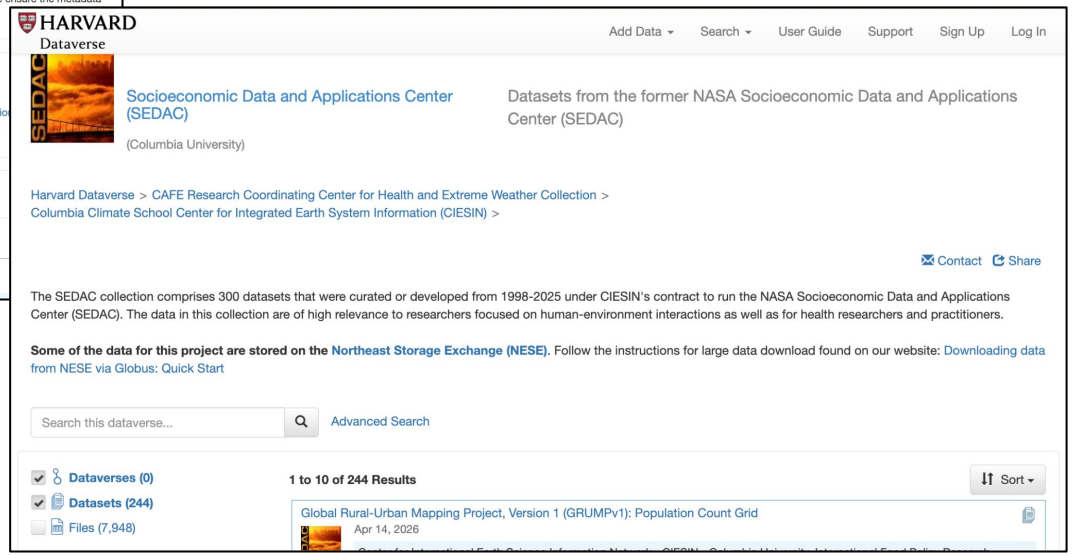
Data Rescue at Harvard Dataverse

- **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.

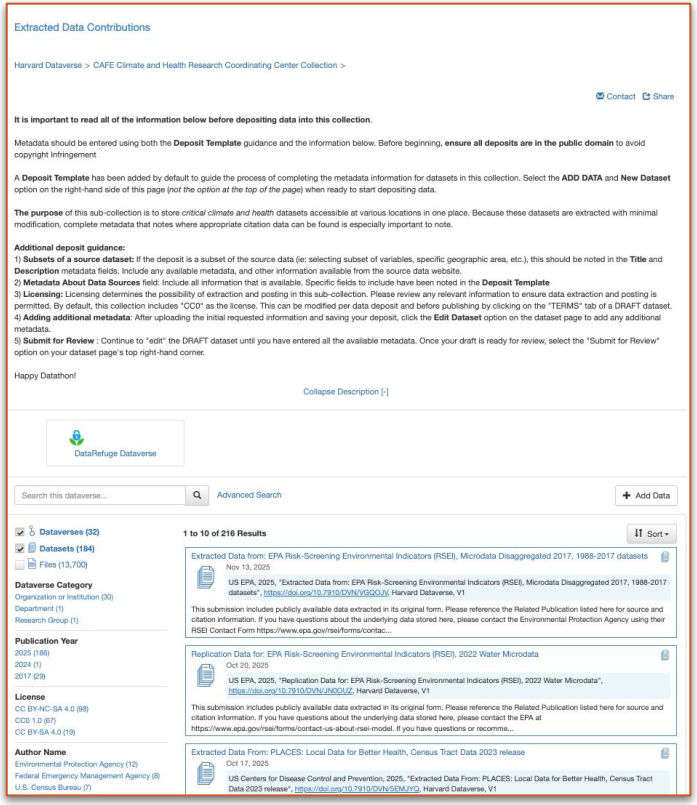


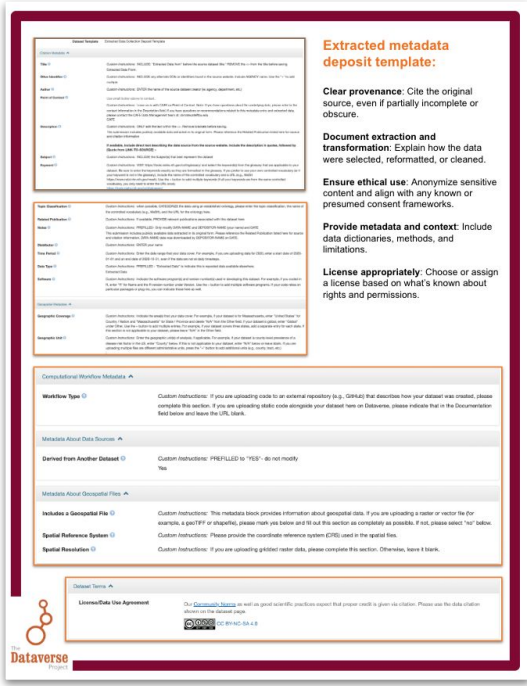




The CAFE Research Coordinating Center for Health and Extreme Weather (RCC) is a joint initiative by the Boston University School of Public Health and Harvard T.H. Chan School of Public Health (**click on the images to navigate to the collection**).

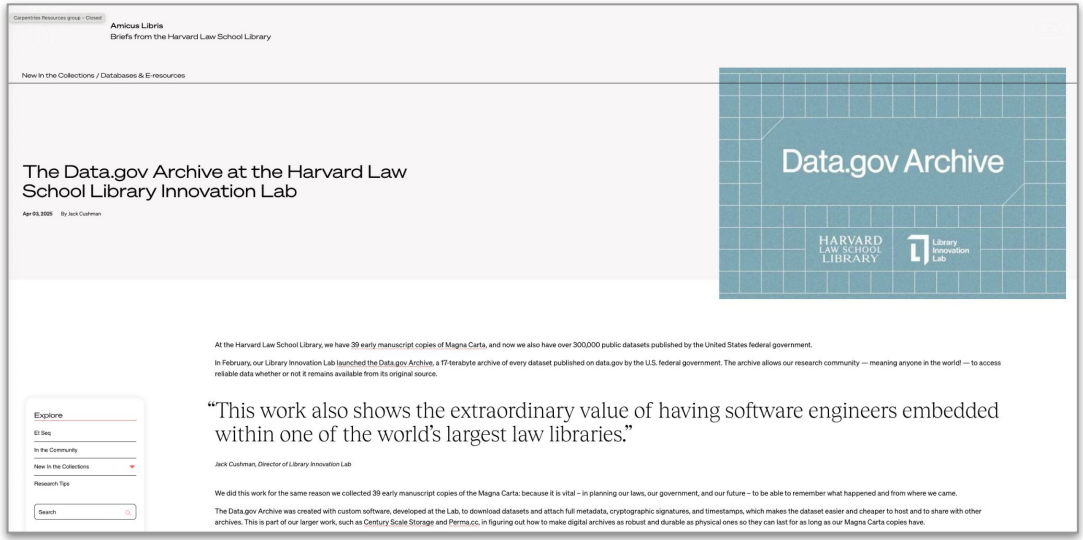
Extracted Data Use Case

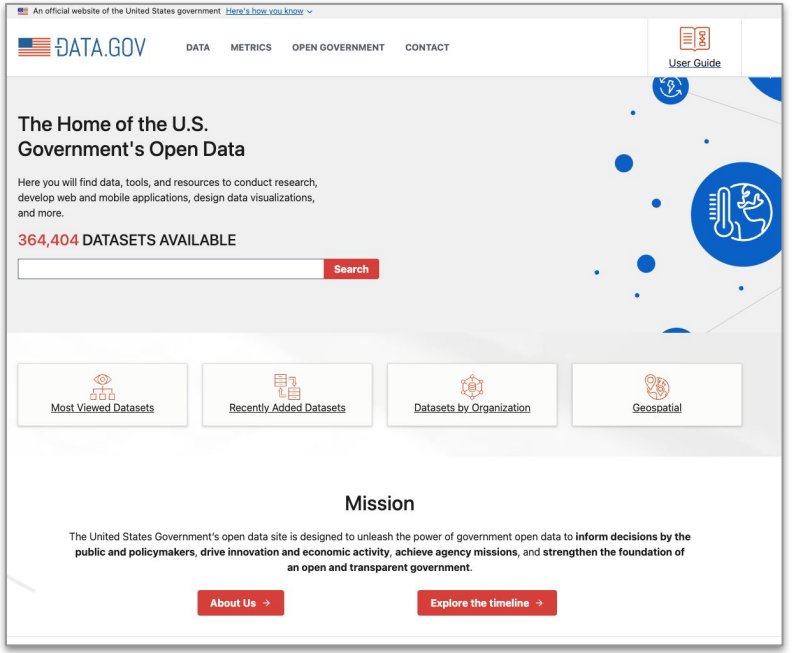




In the context of **data rescue**, extracted data sharing involves identifying, recovering, and reformatting valuable information from at-risk or inaccessible sources—such as outdated systems, legacy formats, or deteriorating media—and making curated subsets of that data available for reuse. These extracted 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





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.