

Data Rescue Project

Updated 2025-02-20

These suggestions come from various sources, including [IASSIST](#), [RDAP](#), [Data Curation Network](#), BlueSky, LinkedIn, and others.

WE HAVE A WEBSITE!

[About Data Rescue Project](#)

The Data Rescue Tracker: <https://www.datarescueproject.org/data-rescue-tracker/>

We will be phasing out the use of this google doc (it has been fun). Updates will be on [our website](#) moving forward. Feel free to subscribe to get those posts.

For more current information on Current Efforts go to:

<https://www.datarescueproject.org/current-efforts/>

**WE ARE NO LONGER USING THIS GOOGLE
DOC.**

**The information below is no longer updated.
PLEASE REFER TO OUR WEBSITE FOR
CURRENT INFORMATION.**

Ways to get involved:

- You can email us at datarescueproject@protonmail.com
- If you want to stay in touch with us about future efforts, fill out [this form](#). We will only use it to communicate efforts.
- **Follow us at bluesky** <https://bsky.app/profile/datarescueproject.org>

Information below is available on our website:
<https://www.datarescueproject.org/about-data-rescue-project/>

Larger and Established Data / Website Efforts

- [End of Term Crawl](#)
 - The main coordinated effort to archive websites
 - Datasets have been more of a challenge, especially data embedded in databases.
- [EDGI](#)
 - They have been focused on environmental data and a good organization to follow for updates.
 - They work with [Public Environmental Data Project](#) (see below)
- [Public Environmental Data Project](#)
 - A coalition committed to preserving and providing public access to federal environmental data.
 - February 7, 2025 - [EPA's EJScreen 2.3](#)
 - January 31, 2025 - [CDC's Social Vulnerability Index and Environmental Justice Index](#)
 - January 24, 2025 - [Council on Environmental Quality EJScorecard](#)
 - January 24, 2025 - [Climate and Economic Justice Screening Tool](#)
- [Harvard's Library Innovation Lab Team](#)
 - They have been focusing on data.gov and released their data on Feb 6, 2025.
<https://lil.law.harvard.edu/blog/2025/02/06/announcing-data-gov-archive/>
 - #SafeguardingResearch is in contact with them to mirror data on servers not in US-jurisdiction
- [ICPSR](#)
 - Overview of ICPSR's data rescue activities to date:
 - Downloaded ~2800 files from various sources requested by researchers; all the files ICPSR collected will soon be available via a dropbox link.
 - Examining CDC data dump from [archive.org](#) to assess what might be missing.
 - Ideally will also be a resource for those looking for data to see what is/isn't available.
 - ICPSR staff and allies are generating metadata for each of the datasets we have so that we can make them available through an existing archive at ICPSR (DataLumos, openICPSR, or the Resource Center for Minority Data, depending on our timeline and some technical issues we're working out)
 - [ICPSR Data Lumos](#) - They have the older version of a lot of major data, including a recent addition from the CDC.

- [IPUMS](#)
 - They have data and have been working on cataloging efforts
 - Notification went out yesterday that they will share more soon.
- [Dryad](#)
 - Generalist repository available to help with data publication, storage, and preservation.
- [Silencing Science Tracker](#)
 - Joint initiative of the Sabin Center for Climate Change Law and the Climate Science Legal Defense Fund.
 - Tracks government attempts to restrict or prohibit scientific research, education or discussion, or the publication or use of scientific information.
- [OSF](#)
 - Generalist repository for archiving, sharing, and storing all types of research outputs, not limited to preprints or only data.
 - OSF is available as an option for pre-prints of articles if, for some reason, they cannot be [posted on official sources](#).
 - Many universities also have institutional repositories where research (articles, data, dissertations, etc) from that institution can be posted. They also have preservation mandates. An example is [Penn's ScholarlyCommons](#).
- [The Climate Mirror Project](#)
 - Has NOAA data pulled during the 2017 data rescue.
- [Open Energy Data Initiative](#)
 - A volunteer has pointed out that “key equity data” is missing from the Dept of Energy. Says they were able to find it on this site. Includes additional data from DOE.
 - Note that OEDI is run by NREL which is part of DOE, it is not an independent organization. Several energy efficiency consultants have received tips that NREL has been ordered to start taking datasets offline. It's not clear if OEDI data may be removed or not.
- [Wayback Machine](#)
 - The Wayback Machine is an initiative of the Internet Archive, a 501(c)(3) non-profit, building a digital library of Internet sites and other cultural artifacts in digital form. Other projects include Open Library & archive-it.org.
- [Roper Center for Public Opinion Research at Cornell University](#)
 - Roper Center has collected over 50,000 files (datasets and documentation) from 22 federal survey projects. Efforts to this point have been focused on acquiring the files and ensuring backup copies are preserved on multiple servers.
- [PANGAEA - Data Publisher for Earth & Environmental Science](#)
 - [The information system PANGAEA is operated as an Open Access library aimed at archiving, publishing and distributing georeferenced data from earth system research. PANGAEA is ready to rescue data - just contact us.](#)

Data Rescue Events

- [University of Washington-based Data Rescue](#)

- Hosted by the University of Washington Center for Advances in Libraries, Museums, and Archives (CALMA), series of data rescues followed the model from 2017. The spreadsheet of data reviewed at the events is available:
 - ✕ [Data Tracking List - Data Rescue 2025 \(Responses\).xlsx](#)
 - It is unclear if they are hosting more.
- Healthy Regions Policy Lab at UIUC
 - <https://emails.illinois.edu/newsletter/02/615978402.html>
 - Includes CDC, EPA, and HRSA Data
- Stanford's Big Local News
 - They are running [Federal data collection collaborative](#)

Smaller/Ad Hoc Rescue Efforts/ Data Archiving Activists

- UCSB LSIT Data Mirroring
 - Mirrored and archived public data on locally hosted git server
 - Includes retrieved data sets from CDC, NIH, and NOAA
- CDC Page on Internet Archive
 - A special archive created on IA of all CDC datasets publicly available as of January 28, 2025
 - uploaded by [DataHoarders](#) (we think)
- Datasets in Dataverse
 - Includes CDC's Social Vulnerability Index data.
 - Most of what's being placed here is data focusing on health and the environment.
 - **DataRefuge from 2017** [DataRefuge](#) is open for deposits of individual datasets or collection (it's a subcollection within CAFE)
 - Please note **if you need TBs of support**, reach out to sbarbosa@g.harvard.edu
 - [NOAA \(fisheries\)](#) is working with Harvard Dataverse to upload over 4TB of data and likely support the efforts via CAFE
- NOAA heat-index files: <https://archive.org/details/noaa-heat-index-2025-01-17>
- NIH NIEHS Climate Change and Human Health Literature Portal (CCHHL) data dump: https://archive.org/details/cchhl_2025-02-10
- Safeguarding Research
 - Organizer is Henrik Schönemann; <https://fedihum.org/@lavaeolus>
 - There is a forum: <https://safeguarding-research.discourse.group/> (admin = Henrik)
 - Based in EU, USA and global - got access to Update 1-2 PB (and more on the way) of storage & people willing to seed
 - Currently, we've got around 1TB of data backed up
 - Including >100.000 PDFs from academia.edu ("transgender", "Queer Studies", "intersex", "nonbinary" etc. - see the forum for the full list)

- 350GB web archive of CDC, including all 30.000 files from archive.cdc.gov And much more
 - “We're working on providing a central index of archives, with metadata about who archived what, when, to be disseminated widely alongside torrent files and act as both a central point of coordination for archivers to assess what new work is needed, and a mass distribution channel.”
- Possible contact to CERN, will update asap
- [Data Hoarder](#)
 - A reddit community that is coordinating efforts to rescue data.
- [Data Hoarding](#)
 - index of resources and archives related to data hoarding, web archival and self hosting.
- [ArchiveTeam Warriors](#)
 - They run a distributed crawler. Anyone can install it to help contribute.
 - [US Federal Data page](#)
 - Data is uploaded to Archive.org by volunteers
- [Data Liberation Project](#)
 - Note: It looks like the project may have stalled in September 2024. Send info if you know more about them.
 - Run by [BigLocalNews](#) and [MuckRock](#), which are good groups to follow.

Tools for Data Rescues

- [DCN Curating Data for Data Rescues](#)
 - Provides key insights for curating data and the types of questions that need to be asked.
- [Data Management Checklist For Data Rescues](#) (from MIT)
 - Checklist to assist with curating data rescue efforts.
- [#RStats](#) package from [@ropensci.org](#)
 - [gitcellar](#) downloads and archives all repos, issues, and PRs from a GitHub organization in one shot: docs.ropensci.org/gitcellar/
- [WebRecorder.net](#)
 - According to an email: has archived 8TB+ of government sites, some from the End-of-Term-Archive seed list, some from EDGI Slack requests, and many sites independently
- [ArchiveBox.io](#)
 - According to an email: has also archived government datasets from [data.gov](#), CIBP, USCIS, NOAA, NASA, NSIDC, and more
- [Awesome-datahoarding](#)
 - Provides a list of tools for web harvesting, etc.
- [Awesome Web Archiving](#)
 - Another curated list of web archiving tools
- [DataRescue Workflow](#)
 - This is the workflow from the original data rescue/DataRefuge project in 2017.

- Many of the tools are no longer working, but the workflow is still useful. UW used this to create their workflow above.
- The challenge with the original project was where to store and how to make discoverable the large amounts of data captured.
- Part of this effort is also housed in the [Harvard Dataverse Repository](#) and can be opened for more data deposits
- There is a [CKAN instance](#) with some of the 2017 data.
- <https://govdiff.com/>
 - Tool created by [Jerome Paulos](#) to show side-by-side changes in government websites.
- [How You Can Help Archive U.S. Government Data Right Now: Install Archive Team Warrior](#)
 - This is a reddit post, but it lists instructions for how to archive and the tools needed to be able to contribute. Figured it would best be categorized here.

Existing Alternative Data Sources

Thanks to Brianne Dosch for suggesting the section and some of the bullets.

- [PolicyMap](#) – offers a free tier that can be used to view basic information down to the tract-level, but more detailed data and functionality requires a subscription; available at some universities
 - [Purged Federal Agency Data Available](#)
- [GEM](#) - contains 46 demographic, energy, health, and housing indicators down to the tract-level with free access for community-based organizations; others require a subscription.
- [FRED](#) - They have some demographic data as well; free and open source
- [Census Reporter](#) – is a free, open-source platform focused on making American Community Survey (ACS) data more accessible, including the recent upload of the 2022 1-Year ACS data
- [Esri](#) – for mapping users, the GIS vendor publishes several U.S. Census Bureau data sets, [including the ACS](#), through its [ArcGIS Online Platform](#)
- [IPUMS](#) – Even when the government operates normally, many analysts turn to Minnesota Population Center products to access [ACS](#), [Current Population Survey](#) microdata and [Decennial Census](#) data
- Social Explorer - historical Census data and more; available at some universities
- SimplyAnalytics - has internally processed American Community Surveys; available at some universities
- [American College of Obstetricians and Gynecologists](#) - Hosting copies of immunization schedules and contraceptive use guidance from the CDC

[Economic Indicators](#)

- National League of Cities: Federal Grant Navigation [Equity Dashboard](#)

- This tool aggregated data from many sources – it seems to be still able to categorize disadvantaged communities (by environmental and economic standards), as well as other critical data denotations that are increasingly hard to access
- ALICE Economic Vitality [Dashboard](#) and [Report \(2022 w/ 2024 update\)](#)
 - This resource specifically provides data on work, housing, and community resources for households below the ALICE threshold (Asset Limited, Income Constrained, Employed). The data is provided by the U.S. Census Bureau's Public Use Microdata Sample (PUMS, 2021)
- National Equity Atlas [Dashboards](#)
 - A data and policy tool that provides a detailed report card on racial and economic equity – this tool can provide a holistic Racial Equity Index snapshot of communities. The Atlas draws its data from a unique regional equity indicators database developed and maintained by two private institutions: PolicyLink and USC Equity Research Institute ERI.
- Economic Policy Institute's [State of Working America Data Library](#)
 - provides researchers, media, and the public with easily accessible, up-to-date, and comprehensive historical data on the American labor force. It is compiled from Economic Policy Institute analysis of government data sources. Use it to research wages, inequality, and other economic indicators over time and among demographic groups.

[Public Health](#)

- County Health Rankings & Roadmaps [\(CHR&R\)](#)
 - A program of University of Wisconsin's Population Health Institute, this data tool aims to highlight the symbiotic nature of health and equity by factoring in physical environment, social and economic indicators, clinical care, and health behaviors to health outcomes.
 - They also recommend these additional health data platforms:
 - [America's Health Rankings](#) report is a health assessment tool based on state-level health indicators.
 - [Congressional District Health Dashboard](#) pulls together local data on the health and well-being for each congressional district.
- City Health [Dashboard](#)
 - From NYU Langone Health, this platform provides 40+ measures of health and factors affecting health across five areas (Health Behaviors, Social and Economic Factors, Physical Environment, Health Outcomes, and Clinical Care) for 970+ cities across the U.S.

Library Guides to Data Rescues

- American Univ: https://subjectguides.library.american.edu/data_rescue (Now shared through Springshare)
- Univ of MN: <https://libguides.umn.edu/govpubs/admin>
- Salem State: <https://libguides.salemstate.edu/datapreservation>
- Butler: <https://libguides.butler.edu/archiveddatasources>
- Hamilton: <https://libguides.hamilton.edu/c.php?g=132443&p=10779226>
- Albany: <https://libguides.library.albany.edu/c.php?g=1450281&p=10779581>
- GODORT: <https://godort.libguides.com/c.php?g=1450475&p=10780944>
- Syracuse University: <https://researchguides.library.syr.edu/data>
- Ohio State: <https://guides.osu.edu/data/federal-data> - focused on a streamlined subset of info intended for researchers that are just trying to access data
- U of Notre Dame: <https://libguides.library.nd.edu/research-data-services/archived-federal-data-2025>
- UC Merced (shared through Springshare) <https://libguides.ucmerced.edu/US-Federal-Data>
- Trinity College: <https://courseguides.trincoll.edu/gov-data-archives>

Articles on current efforts

- [Preserving Data Access](#) - Duke University Libraries blog post by Joel Herndon
- [Call to arms: What government information librarians can do to help save critical federal information from being lost](#) - Blogpost from FGI (Free Government Information)
- [Why EDGI is Archiving Public Environmental Data](#) - blog post from EDGI
- [Preserving federal health data](#) - by The Journalist's Resource out of the Harvard Kennedy School
 - [As the US government removes health websites and data, here's a list of non-government data alternatives and archives](#) - by The Journalist's Resource
- [Archivists Work to Identify the Thousands of Datasets Disappearing from Data.gov](#) - by 404 Media; interviews with EOT and James Jacobs
- [The scramble to back up CDC.gov](#) - by Garbage Day; mentions some coordinating efforts by Health Professionals and Journalists to gather the CDC data
- [Lending a hand with EOT Crawl](#) - blog post from the PEGI Project.
- [As the Trump admin deletes online data, scientists and digital librarians rush to save it](#) - Salon Magazine. Talks about EOT.
- [Three Efforts to Preserve Government Data as a New Trump Administration Approaches](#) - Union of Concerned Scientists
- [What's at Stake if the Data at Federal Agencies Disappears?](#) - Union of Concerned Scientists
- [Researchers rush to preserve federal health databases before they disappear from government websites](#) from The Journalist's Resource

Articles for context

- [CDC Site Restores Some Purged Files](#) from NYT
- [Thousands of U.S. Government Web Pages Have Been Taken Down Since Friday](#) by Ethan Singer.
- [The Government Information Crisis Is Bigger Than You Think It Is](#) blog post by Free Government Information
- [CDC removes gender, equity references in public health material](#) from WaPo
- [BREAKING NEWS: CDC orders mass retraction and revision of submitted research across all science and medicine journals](#) from Inside Medicine
- [A Look at Federal Health Data Taken Offline](#) from KFF
- [As Data Goes Off-Line Under Trump, Environmental Researchers Are Uploading Backups](#) from Inside Higher Ed
- [The mad dash to protect environmental data from Donald Trump](#) from The Verge
- [Some federal health websites restored, others still down, after data purge](#) from VPM
- [Trump orders USDA to take down websites referencing climate crisis](#) from The Guardian