



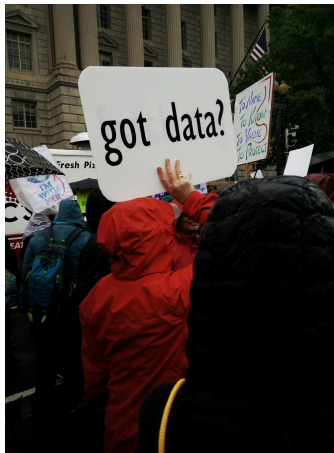
The Integrity of Public Access to Federal Data

Evaluating Disruptions to Open
Government Data, 2025-2026

Christopher Steven Marcum, Ph.D.

April 2026

The Integrity of Public Access to Federal Data © 2026 by Christopher Steven Marcum, Ph.D., is licensed under [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/) 



Funding support for this project was provided by *Funders for the Future of Public Data (3FPD)*.

Acknowledgments :

The author is grateful to the following people for their reviews which improved this report: Claire McKay Bowen, Andrew Gerard, Hyon Kim, Daniel A. Lee, Jacob Pasner, Meghan Stuessy, and two additional anonymous reviewers.

The author is also grateful to the many people who provided insights into the federal data landscape through anonymous interviews and conversations in support of this project.

Special thanks are also owed to Janet Freilich for sharing code with the author.

Recommended Citation :

Marcum, Christopher Steven. 2026. "The Integrity of Public Access to Federal Data: Evaluating Disruptions to Open Government Data, 2025-2026." DOI: <https://www.doi.org/10.5281/zenodo.19556076>

Table of Contents

[Executive Summary](#)

[Introduction](#)

[The Federal Data Catalog](#)

[Administrative Risks to Data](#)

[The Richardson Waiver Rescission](#)

[Paperwork Reduction Act Exemptions](#)

[Information Collection Discontinuation and Revision](#)

[Resourcing and Staffing](#)

[Data Tools](#)

[Agency Case Studies](#)

[Office of Management and Budget](#)

[United States Agency for International Development](#)

[U.S. Department of Veterans Affairs](#)

[Auditing Open Government Data Assets](#)

[References](#)

[Glossary](#)

Executive Summary

This report provides an analysis of the integrity of public access to federal open government data assets during the disruptions to the federal data ecosystem during 2025 and early 2026. Here, data integrity is [defined](#) as the “maintenance of, and the assurance of, data accuracy and consistency over its entire life-cycle”, of which public access to open government data assets is assumed to be essential. The report clarifies the scale and mechanisms of data disruptions, discusses specific threats to data integrity, highlights exemplar cases of data disruption from federal agencies, and delivers a transparent methodology for reproducing auditing routines used in this assessment. The evidence used in this report was derived from multiple sources, including news reports, academic literature, materials from civic society and government oversight organizations, interviews with experts, archives, and government sources.

Primary findings

1. As widely reported in the press and by advocacy organizations, there were 3,000 to 4,000 open government data assets removed from public access in the last year. However, the data relied upon by popular reporting (changes to the Federal Data Catalog (FDC) as provided by Data.gov) are not reliable. Rather, the findings of this report point to significant datasets that were removed that were not widely reported, including many that were not indexed in the FDC.
2. The Trump Administration has engaged in large-scale dataset discontinuations using appropriate, lawful routes to end information collections. Between January 21st, 2025 and January 20th, 2026 this Administration had discontinued at least 562 information collections using the Paperwork Reduction Act (PRA) processes, which is 65% more than the Biden Administration had discontinued in the same interval the year prior. While documenting changes to information collections involving data on humans and organizations (i.e., subject to the PRA) is easily facilitated through the public docket and reporting by the Office of Management and Budget, there is no such equivalent docket to track changes to datasets not subject to the PRA (such as climate observation data, for example).
3. High-value data tools were taken down by the Administration while public access to their underlying data were largely retained. While many of the data tools, which often provide greater utility to the general public than raw datasets, were restored by civic society groups or government contractors that retained access to them after contract cancellation, their removal from public access by the government adds friction to data users and breaks downstream workflows and applications that relied on the tools for information.

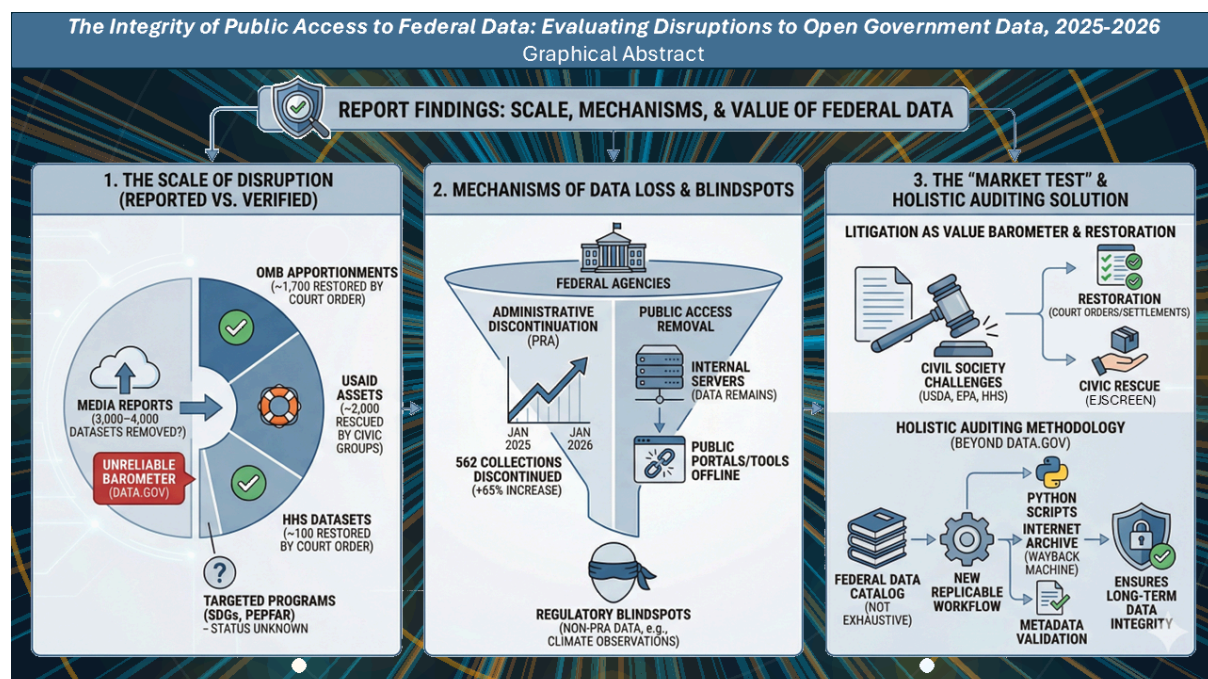


Figure: This figure summarizes the report’s findings on the scale, administrative mechanisms, and verification of disruptions to open government data. The leftmost panel describes the scale of public access loss and recovery of federal data, the central panel describes mechanisms and information gaps in the disruptions to federal data, and the third panel describes both litigation and forensic auditing. The figure is a custom modification of an image generated using Gemini Pro’s Nanobanana.

These findings are discussed in more detail below.

Actual Public Access Removal of Open Government Data

A primary finding of this report is that, while broadly reported estimates of 3,000 to 4,000 datasets being removed from public access are generally accurate in their total count, the evidence cited to support these figures is often technically flawed. Many observers relied on the topline dataset count displayed on Data.gov as a barometer for data integrity. The FDC is not a data repository and the topline count is not sufficient to surmise changes to underlying datasets. The dataset count fluctuates routinely due to normal harvesting cycles. A [forensic audit](#) conducted for this project indicates that the net change in the Data.gov topline count of assets in the [FDC](#) during the initial months of 2025 was approximately zero despite the observed fluctuations. There are almost 100,000 more datasets indexed in the FDC as of the publication

of this report than there were at the height of reports of the changes to the Data.gov topline. Many of the new additions indexed in the FDC have high-value, such as the updated Census geographic shapefiles which represented more than a third of the growth of the catalog.

The estimate of roughly 3,000 to 4,000 datasets removed from public access is more accurately derived by aggregating documented actions across specific cases:

- **Office of Management and Budget (OMB):** The revocation of public access to the [apportionments database](#) resulted in the loss of approximately 1,700 individual apportionment files. These data assets have all been restored by court-order.
- **United States Agency for International Development (USAID):** The elimination of the [Development Data Library](#) removed public access to more than 2,000 data assets. These data assets are largely available through rescue efforts but their federal records lifecycle is currently unknown.
- **Department of Health and Human Services (HHS):** Approximately 100 datasets had been identified as removed from public access based on lawsuit filings and external monitoring. These data assets have been restored by court-order.
- **Targeted Program Removals:** Several dozen evaluation datasets removed due to shifts in administration priorities (e.g., Sustainable Development Goals, PEPFAR evaluation studies). The statuses of these data assets are unknown.

Distinction Between Data Tools and Source Data

The report identifies a significant distinction between the removal of public-facing [data tools](#) and the removal of the underlying data assets. Throughout 2025, several high-profile data tools and dashboards were eliminated from agency websites. These removals significantly reduced the accessibility and utility of the information for the general public and data users. However, forensic auditing confirms that much of the source data provided by these tools remained accessible to the public through the agency servers or through federal contractors that hold the data on behalf of the originating agencies. The loss of these tools represents a degradation of federal data utility, interpretability, and accessibility rather than data loss. Because of the value these data tools bring to the public good, many of them have been functionally restored by civic society organizations.

Risks to Public Access and Data Integrity

While the direct removal of federal open government data from public access was rare between 2025 and 2026, the findings underscore that the primary risks involved a combination of political intervention and diminished administrative capacity. These risks include the removal of public access to existing assets and tools, errors that lead to metadata, resource constraints, the discontinuation, change, or expiration of information collections through the Paperwork Reduction Act (PRA) process. While certain political and administrative actions that threaten the integrity of federal open government data have been challenged in court (typically, as violations of the Administrative Procedures Act (APA)), much of the disruptions are likely fully legal and represent immutable exercise of Executive authority. When it comes to administrative process to modify information collections under the PRA, for example, it's likely that such changes were entirely proper, albeit unusual in scale: such as how the Trump Administration discontinued 562 information collections between January 2025 and January 2026, which is a 65 percent increase over the preceding year. Moreover, the lack of a transparent process for adequate public notice of changes to data collections that are exempt from the PRA or the APA (such as many non-human subjects research datasets, clinical research from the National Institutes of Health, and practically all other datasets from the Department of Health and Human Services associated with rulemaking or administrative procedures given the [Richardson Preyer rescission](#)) mean that oversight into future actions involving those federal data is potentially limited.

Need for Improved, Holistic Auditing

There are significant blindspots in the federal data ecosystem that obscure monitoring of open government data assets. While the FDC provided by Data.gov has nearly 500,000 data assets, there are [challenges in using that resource](#) for timely auditing. Moreover, despite the large number of indexed data assets, the FDC is not exhaustive of all open government data assets (for instance, the OMB apportionments data and many NIH datasets are not listed in the FDC). Accompanying this report is a replicable methodology for [auditing open government data assets and metadata](#) from both agency and FDC sources, [including code](#) and [associated data](#). While not an enterprise solution scalable for the entire federal data ecosystem, the auditing workflow provides a framework that could be generalized for such cases.

Recommendations

The report findings highlight that while direct data loss was relatively rare, the integrity of federal information was frequently compromised through administrative discontinuations, public access removals, and severe resourcing constraints. Moreover, forensic auditing revealed significant gaps in both the ability to holistically track federal data assets because of limitations in the FDC and a highly federated ecosystem. To safeguard the role of open government data as a vital public good, proactive measures are necessary across the legislative, executive, and civic sectors. The following recommendations provide a framework based on the findings of this report for Congress to strengthen statutory oversight, for federal agencies to improve transparency and metadata practices, and for outside stakeholders to adopt more rigorous monitoring and archival strategies.

Recommendations for Congress

- **Reform Repository Requirements:** Amend the *Foundations for Evidence-Based Policymaking Act* (Evidence Act) to require that the Federal Data Catalog (FDC) or a successor system supports the actual acquisition and storage of high-value datasets rather than merely indexing metadata links, which are currently susceptible to link rot.
- **Address Oversight Blindspots:** Close the Paperwork Reduction Act (PRA) exemption for the National Institutes of Health (NIH) and similar research entities to ensure all federal data collections adhere to government-wide standards, such as those regarding race and ethnicity data.
- **Stabilize Statistical Agency Funding:** Provide multi-year, protected funding for the 13 principal statistical agencies (such as BLS, NCES, and EIA) to prevent irrecoverable gaps in historic data series caused by government shutdowns and mass staff departures. These are the data that make the economy run.
- **Mandate Transparency for Non-PRA Data:** Establish a statutory requirement for a public docket to track changes, removals, or discontinuations of datasets not currently subject to the PRA or APA, such as climate observation and environmental monitoring data.

Recommendations for Federal Agencies

- **Prioritize Metadata Fidelity:** Ensure that Comprehensive Data Inventories (CDIs) follow DCAT-3.0 metadata schema and include individual downloadURL properties for all distributions rather than just landing pages. This practice reduces user friction and preserves access when website structures change. Mint DOIs for all datasets and ensure that their resolveable, persistent URLs link directly to the asset.
- **Adhere to Notification Standards:** Follow the guidance in OMB Circular A-130 to provide “adequate notice” before terminating any significant information product, even if the action is not strictly required by the Administrative Procedures Act or the Paperwork Reduction Act.
- **Protect Confidential Data Access:** Prioritize the retention of specialized staff who manage CIPSEA-protected data and secure enclaves (FSRDCs) to ensure that the paralysis of the restricted data ecosystem seen in 2025—which cut-off critical research—is not repeated.
- **Standardize Data Asset Classification:** Implement the guidance in M-25-05 to correctly distinguish between actual data assets and information products like infographics or reports, thereby improving the accuracy of the Federal Data Catalog.

Recommendations for Other Stakeholders

- **Utilize Replicable Auditing:** Employ and fund independent, code-based auditing workflows. Support and leverage tools like the Internet Archive’s Wayback Machine to verify agency data holdings rather than relying on the unreliable contemporary topline dataset counts provided by Data.gov.
- **Engage in “Market Test” Advocacy:** Continue using targeted litigation as a “market test” for data value; court rulings have proven to be an effective mechanism for restoring high-value data assets removed from public access.
- **Coordinate Data Rescue Efforts:** While redundancy and overlap is desirable, the ecosystem can suffer from competition and lack of coordination between data rescue and preservation efforts. Funders might consider conditioning projects on coordination and collaboration to ensure that redundant efforts work towards synergies and learn from one another.
- **Formalize Public Comment:** Use monitoring platforms (like dataindex.us) to keep track of changes to information collections that create federal data assets to organize and submit public comments during revision proposals.

Conclusion

Estimates of the scale of federal data loss and manipulation that occurred after the 2025 US Presidential Inauguration relied on largely on observing changes to the topline dataset count within the FDC on Data.gov or on changes to metadata associated with those data. By mid-2025, observers resolved on a figure of around 3,000 to 4,000 datasets removed from public access. However, the topline count on the FDC is not an appropriate barometer of disruptions to the integrity of federal data assets. Still, the estimate is approximately accurate when accounting for data losses resulting from political interference and infrastructural losses. Some of the data assets removed from public access were never indexed by the FDC and most have since been restored by court-order though the status of other assets is unknown. Moreover, information collection discontinuation and revisions to existing collections to conform to Administration priorities and directives were done at unprecedented scale through the PRA processes. Data tools, which provide data users accessibility and useability of underlying data assets, were significantly affected by disruptions during the last year. Because of the value of these resources to the public, many data tools have been the focus of restoration efforts by civic society and other special interest groups.

While political interference risks to the integrity of both data assets and data tools are substantial and real, the lawfulness of these activities is likely consistent with Executive authority - most lawsuits in this space focus on the administrative process involved in removal, manipulation, cessation or deresourcing, rather than the authorities for such actions themselves. Some legal challenges have been successful on their merits, while others have failed; no single case has wholly resolved the risks to public access to open government data. What both the lawsuits and the data tool restoration efforts reveal, however, is the magnitude of community-value of those public goods. Congress, the federal agencies, and civic society groups should work to improve the regulatory, administrative, and infrastructural durability of sustainable public access to federal data.

Introduction

Open government data is a cornerstone of transparency, productivity, accountability, and evidence-based policymaking in the United States (US) and abroad. Open government data enables researchers, journalists, policymakers, and the public to monitor government performance, accelerate discovery, evaluate programs, and safeguard democratic institutions. These data represent an important, non-excludable and non-rivalrous, public good. Yet over the past decade, open government data has faced growing threats, including program disruption from funding shortfalls, political interference, and an erosion of trust government.

Background

In the last year alone political interference led to the manipulation, suppression, or outright removal of federal data assets on topics ranging from climate change, to economics, to LGBTQ+ issues, to public health. This interference threatens to undermine public trust in federal data by limiting the ability of civil society to hold government accountable, eroding the trust businesses hold in federal statistical data, and revoking access to one-of-a-kind data resources to researchers for innovation and discovery.

At the same time, chronic under-funding, lack of interoperability across agencies, and outdated technical infrastructure have compounded the risks to open government data. Even absent political interference, federal data is often fragile and subject to degradation, disappearance, or diminished accessibility due to poor stewardship and a lack of ecosystem sustainability. This issue transcends any single Administration; however, it is compounded during the current moment when program cuts are an ever-present threat to the federal data landscape.

Since shortly after the 2025 US Presidential inauguration, many observers tried to quantify the scale of federal data loss and manipulation. Some suggested only a few hundred datasets have been affected by the actions of the current Administration, while others claimed thousands. As a result, estimates of how many federal data assets were affected by disruptions are uncertain. This is, in part, because of a lack of consistent, transparent methodologies — including definitions — used by each stakeholder. This report seeks to resolve that confusion by producing a rigorous, transparent analysis and a replicable methodology that stakeholders can use to continually track the state of open government data.

While the primary focus of this report is on open government data assets, which are data made freely available to the public with few-to-no encumbrances as defined in Title II of the *Foundations for Evidence-Based Policymaking Act of 2018* known as the OPEN Government Data Act ([United States Congress, 2019; Stuessy & Knoedl, 2026](#)), disruptions to restricted federal data assets (such as confidential statistical data) and data tools (such as EJSscreen) are also briefly discussed due to their importance in the federal data ecosystem. However, challenges to federal data integrity that involved reported violations of internal agency policies on restricted-access data ([Fowler et al., 2026](#)), inappropriate use of data ([Schilling & Slowey, 2026](#)), or other violations of the privacy and confidentiality of data providers is out of scope. As articulated in the *Federal Framework for Scientific Integrity* ([National Science and Technology Council, 2023](#)) there is a distinction between inappropriate interference and appropriate political influence in the production of data by the government. Therefore, proposals to modify data collections or methodologies based on Administration priorities that follow appropriate procedures within the context of statute or regulation—including those required by the Paperwork Reduction Act or the Administrative Procedures Act—are also out of scope (examples include testing a U.S. Citizenship question on the Decennial Census ([Wang, 2026](#)) or suspending foreign researcher access to NIH data repositories ([Stone, 2025](#))). The exception to this latter exemption is data collection discontinuation, which is discussed as a risk to the integrity of federal data (see below and the [related chapter](#)).

A Summary of Reporting on the Integrity of Federal Data from 2025 to 2026

Across 2025 and into early 2026, reporting and analysis detailed actual and perceived disruptions to how the federal government collects, maintains, and publishes data products ([Dayak & Kramer, 2026](#)). These accounts spanned multiple domains, including public health surveillance ([Rabin & Mandavilli, 2025](#)), economic and statistical information ([Heckman, 2025; Kiersz, 2026](#)), climate and environment resources and tools ([Brady, 2025](#)) ([Hirji, 2025](#)), and administrative records ([Hartman, 2025](#)). Disruptions manifested in various ways, ranging from datasets disappearing from public view to delayed statistical series, rewritten or removed webpages, altered metadata, modified survey questions, discontinued collections, and halted disclosure and research proposal reviews for restricted data access ([Jones, 2025; Levenstein & Kubale, 2025; Dayak & Kramer, 2026](#)).

Reports frequently attempted to quantify the scale of the disruption, though reporters and researchers acknowledged the immense difficulty of tracking exact figures ([Dayak & Kramer, 2026](#)). Publications cited large numeric estimates to illustrate the breadth of the problem, frequently relying on changes to the topline dataset count reported on Data.gov to demonstrate the decline, even as experts cautioned that the dynamic nature of the portal made this an imperfect metric ([Data Foundation, 2025](#)). Within the first two weeks of the new administration, observers noted a sudden drop of over 2,000 datasets from the portal, falling from approximately 307,854 to 305,564 ([Koebler, 2025](#)). By early February, legislative statements cited a reduction of 1,055 datasets ([Rep. Don Beyer, 2025](#)), while independent trackers noted the decline had grown to 3,379 entries later that month ([Kutz, 2025](#)). By June 2025, reports estimated that over 3,000 taxpayer-funded datasets had been removed across various agencies ([Palmer, 2025](#)). Audits of specific agencies revealed severe operational interruptions, with an analysis finding that nearly half of the frequently updated databases in the public health sector had been paused without explanation, and at least 146 specific files were documented as removed or modified to replace terminology ([Robbins, 2025](#)).

Furthermore, the scale of the threat to data was often framed by the sheer volume of material within the purview of data preservation efforts as independent groups rushed to archive information. Journalists routinely noted that an exact census of the losses remained elusive because many cuts occurred quietly without public announcement, forcing the press to rely on approximations and independent monitoring groups to estimate the total impact ([Dayak & Kramer, 2026](#)). To preserve public information, Harvard Law School's Library Innovation Lab [preserved 311,000 datasets](#) by systematically crawling agency dataset download links from Data.gov, totaling 16 terabytes of data ([Satter, 2025](#)). Other large-scale efforts like the [Data Rescue Project](#) archived more than 1,200 data products originating from over 80 distinct government agencies ([O'Leary, 2025](#)). Specialized data rescue initiatives also focused on scraping and archiving environmental justice data directly from federal sites to ensure continued public access ([Mandel, 2026; Willson, 2025](#)).

Litigation also emerged as a critical mechanism for data restoration. In response to the widespread removal of federal climate and environmental justice resources, advocacy groups launched targeted legal challenges against the administration to force the restoration data and data tools. In February of 2025, a coalition representing farmers and environmental organizations sued the U.S. Department of

Agriculture (USDA) for unlawfully purging climate-related agriculture resources and the interactive “Climate Risk Viewer” from its websites ([Garza, 2026](#)). The USDA subsequently agreed to restore the webpages, and a March 2026 legal settlement required the agency to share the underlying raw datasets to ensure permanent public access, even if the government websites were taken offline again. Separately, environmental and consumer watchdog groups filed a federal lawsuit challenging the administration’s sudden deletion of several key environmental justice mapping tools, including the EPA’s EJScreen and the Council on Environmental Quality’s Climate and Economic Justice Screening Tool (CEJST), arguing that the unannounced removals violated administrative procedures and unlawfully deprived vulnerable communities of crucial pollution data ([Noor, 2025](#)). This suit was subsequently dismissed, with the presiding judge ruling that the groups did not have standing to sue ([Clark, 2026](#)). Another major lawsuit resulted in a legal settlement requiring the Department of Health and Human Services to restore over 100 specific datasets, webpages, and tools that had been removed from public access ([Alder, 2025](#)) (HHS ended up restoring more than 300 such resources). Additionally, federal accountability watchdogs determined that the takedown of a key budget apportionment website violated federal law, leading to a court order that forced the administration to republish the spending transparency data ([Hill, 2025](#); [Katz, 2025](#)). Subsequent appeals court rulings reinforced that such clamp-downs on spending data defied congressional authority ([Cheney & Gerstein, 2025](#)). Finally, outside groups also went to the courts when news of orders to destroy classified and personnel records at USAID came to light ([Beitsch, 2025](#)).

These lawsuits reveal insights into the value of federal open government data. While many data rescue efforts focus on data preservation at scale largely under a value system of archiving data not just for its use but as a cultural artifact, plaintiffs in lawsuits exercising a private-right-of-action operate out of immediate, tangible necessity for specific data. The willingness to endure the grueling, costly process of federal litigation serves as a *de facto* “market test” for a dataset’s value. It reveals exactly which data civil society relies on to hold the government accountable, ensure equitable access to resources and benefits, or protect public health and safety. Ultimately, while bulk preservation safeguards the existence of the data, the significant investment required to litigate underscores its active, indispensable role in the functioning of democracy.

Research Methods

A mixed-methods approach for gathering and reporting evidence was used in this report. This approach included statistical analysis, literature review, forensic auditing, and confidential interviews with key actors. Most statistical analyses were conducted in R and all programming for data collection and forensic auditing was done in Python. Unless otherwise disclosed, the data and scripts for reproducing this work (and to support future auditing of the federal data ecosystem) are provided in the [GitHub repository supporting this work](#). No interview notes or transcripts were retained to protect the identities of trusted confidants and the information obtained during those interactions was largely confirmatory (of evidence gathered or otherwise publicly reported) in nature. Efforts were made to rely on government documents and publicly accessible (or open access) reference and source materials. However, since much of the reporting about disruptions to federal data occurred in the press or on proprietary blogs, paywalls or other barriers may be encountered in attempting to access references listed in [the bibliography](#). A list of more than 150 sources from news media, scholarly publications, and civil society websites, can be found in the project data repository in the file named: [data-integrity-news.csv](#).

The research and forensic auditing supporting this report makes substantial use of the [Internet Archive’s Wayback Machine \(WBM\)](#). The Wayback Machine (WBM) functions by deploying automated “crawlers” that traverse the web, downloading publicly accessible pages and documents, processing and archiving them, and making them publicly accessible on their website. These captures (called snapshots) are timestamped and organized into an index of a website’s *in situ* history, allowing users to enter a URL and navigate through a calendar of such snapshots to see how a site looked during specific points-in-time. The WBM creates a permanent, searchable record of the internet’s history, preserving content that would otherwise be lost to link rot, page changes and removals, or server shutdowns. Critically, the WBM snapshots often capture datasets during its crawls. The Application Programming Interfaces (APIs) that the Internet Archive provides for the WBM make it accessible and adaptable to data auditing routines (see the [chapter on auditing](#), for instance). This is an invaluable resource and the work could not have been done without it.

In addition to the source code and data provided in the GitHub repository, the chapter titled [Auditing Open Government Data Assets](#) includes additional details on the methods used to assess changes in federal data. These methods were used throughout the development of this report, providing much of the data and evidence described in the substantive chapters and forming the basis of questions asked to key actors during those interviews and conversations. This chapter also includes a full workflow and a use-case, which aims to assist others in applying these same routines to future monitoring efforts.

Defining Disruption: Deletion, Access Removal, and Discontinuation

To accurately assess the federal data ecosystem, one must precisely define the mechanisms of data loss and disruption. Public discourse frequently conflates different administrative actions under the umbrella term “deletion.” However, federal data disruptions typically fall into three distinct categories with vastly different implications for preservation and recovery. These are deletion, access removal, and discontinuation. Moreover, there is a complex of statutes and policies that create a regulatory framework around which the federal data ecosystem is supposed to operate, including those that establish lawful methods to effectuate each of these aspects of data management. This regulatory framework, and its limitations, is discussed in detail by a new 2026 Congressional Research Service report ([Stuessy & Knoedl, 2026](#)).

Data Deletion Data deletion refers to the actual destruction or erasure of underlying records from federal servers and databases. True deletion is rare due to federal records retention laws, but when it occurs, it represents a permanent loss of historical information. If raw data files are permanently purged from an agency database without prior archiving, the fundamental integrity of that historical record is destroyed and the agency has likely run afoul of their obligations under the Federal Records Act.

There is very little evidence of actual data deletion - one example, however, may have occurred when USAID was ordered to destroy records during the chaotic dismantling of the agency ([Malesky, 2025](#); [Beitsch, 2025](#)). Many USAID open government data assets were removed from public access (see below), but the extent of actual data deletion is unknown (and, according to sources familiar with the subject, unlikely).

Public Access Removal Public access removal occurs when data continues to exist on internal federal servers but the public facing portals, dashboards, or download links are taken offline. In these instances, the agency retains the data for internal operational use or archiving, but external researchers, journalists, local governments, and the general public lose visibility and access. When this occurs, data are not deleted,

but their utility as a public good is reduced. One of the transparency tools that the public has at its disposal with respect to public access removal is [OMB Circular A-130](#) which requires an agency to provide:

“...adequate notice when initiating, substantially modifying, or terminating dissemination of significant information that the public may be using.”

There are many examples of public access removal during 2025. For example, the Homeland Infrastructure Foundation-Level Data dataset was pulled from public access, altering how non-federal actors could map infrastructure and plan disaster responses ([Dayak & Kramer, 2026](#)). Another critical example was the unlawful removal of apportionment data by OMB ([Hill, 2025](#)) that was later restored by court-order ([Cheney & Gerstein, 2025](#)).

Discontinuation Data collection discontinuation involves halting the ongoing gathering of new information that supports the growth or revision of exiting data assets. Historical data may remain perfectly intact and publicly accessible after discontinuation, but the pipeline for new data is severed. While the Paperwork Reduction Act (PRA) provides a framework for public transparency into collection discontinuation for a large tranche of federal data, there are blindspots including collections not associated with a federal rulemaking and any scientific or programmatic data that are not subject to the PRA.

Like public access removal, discontinuations were widespread during 2025. An example of this is the Department of Agriculture terminating a long running report on household food security ([Smith, 2025](#)). The Department amplified its decision in a [press-release](#) and justified it as an exercise in eliminating wasteful spending on redundant collections. However, this justification was largely decried by advocacy groups as a red-herring and, frankly, inaccurate ([FitzSimons, n.d.](#)).

Structure of this report

This report was created using a custom GitHub pages deployment based on [JustTheDocs](#) and [jekyll-scholar](#). This allows for a more dynamic reading experience and facilitates collaborative updating in the future. A version of the entire report suitable for viewing and printing is [available here](#). Each chapter can be read separately - there are no linear dependencies between the chapters of this report.

Chapter Contents

This Introduction and the [Executive Summary](#) are intended to provide sufficient overview of this project. Additional substantive chapters lavish more detail into specific topics relevant to the research conducted in support of this report. They may be read in any order without loss of context. These chapters contain information specifically on:

- **The Federal Data Catalog:** This section provides a history of Data.gov and the Federal Data Catalog (FDC). It details information policy history from the 2009 Presidential Memorandum on Transparency to the Evidence Act of 2018 to the release of the Open Government Data Act implementation guidance in 2025. It also discusses limitations of using Data.gov for evidence of data integrity issues: such as the harvesting model, incomplete metadata, and the distinction between a metadata catalog and a data repository. [Read chapter.](#)
- **Administrative Risks to Data:** This group of chapters examines several administrative risks to federal data and information collections.
 - **The Richardson Waiver Rescission:** This chapter provides a specific case study related to data integrity challenges that *could* result from the recent decision from the US Department of Health and Human Services to avoid the certain Administrative Procedures Act processes. [Read chapter.](#)
 - **The Paperwork Reduction Act Exemptions:** This chapter explains the legal requirements for federal information collection and how certain exemptions to those requirements pose a significant risk to federal data oversight and integrity. [Read chapter.](#)
 - **Information Collection Discontinuation and Revision:** This chapter analyzes the processes and impacts of ending specific data collections through the standard Paperwork Reduction Act and Administrative Procedures Act processes. [Read chapter.](#)
 - **Resourcing and Staffing:** This chapter focuses on the impacts of funding and staffing changes at federal agencies necessary to maintain data assets. [Read chapter.](#)
- **Data Tools:** This chapter reviews removal of the various tools and platforms used by agencies to disseminate data and information to the public. [Read chapter.](#)
- **Agency Case Studies:** These chapters provide detailed examinations of cases of data integrity issues at three specific agencies:
 - **Office of Management and Budget:** Describes the federal apportionments data takedown and restoration timeline by OMB. [Read chapter.](#)
 - **United States Agency for International Development:** Describes how the dismantling of USAID resulted in a large tranche of public access removal to data assets. [Read chapter.](#)
 - **Department of Veterans Affairs:** An examination of VA metadata and dataset changes that replicates (and augments) the findings of previous research. [Read chapter.](#)
- **Auditing Open Government Data Assets:** This chapter describes the workflow and methods used to audit the Federal Data Catalog, individual datasets, agency comprehensive data inventories, and provides a replicable use-case. [Read chapter.](#)
- **References:** This section includes a comprehensive list of references cited throughout the report [View references.](#)
- **Glossary:** The section provides a comprehensive glossary of terms either directly referenced within the report or otherwise relevant to federal data. [View glossary.](#)

Artificial Intelligence Use Disclosure

No artificial intelligence system (AI) was used in the writing of the text included in this report. However, AI large-language models were used in several other ways that contributed to the quality of this report. These included:

- Google Gemini: Generating bibtex entry citations from URLs and uploaded documents using a custom-build agent with Google Lab's Gem/Opal [available here](#).
- ChatGPT: Monitoring RSS and news feeds for new developments in the press on data integrity issues relevant to this report using OpenAI's "Pulse" with the following prompt: "Search for newly published (January 2026 onward) news stories, blog posts, academic papers, and Federal Register notices related to federal data, including statistical data, scientific data, privacy, removal, deletion, or integrity concerns, and notify me with a concise summary."

- Claude Code and Google Gemini Pro: Assisting with code generation, debugging coding errors, and designing API navigation routines linked to the project's GitHub repository. All AI-generated code is disclosed in the comments of relevant scripts.
- Google Gemini Fast: Creating figures from presentation slides using Google Slides's 'beautify this slide' functionality. All images created this way are disclosed in the captions.
- Google Gemini Pro: Creating a graphical abstract of the whole project for the [Executive Summary](#). This was done by granting Google Gemini Pro Model access to the repository and instructing with the prompt to: "carefully read the report and pay particular attention to the /report/_chapters/execsum.md. Create an infographic that could be used as a graphical abstract of the whole project to include in the executive summary."

References

1. Alder, S. (2025). *HHS Settlement Requires Restoration of 100+ Health Datasets and Tools*. <https://www.hipaajournal.com/hhs-settlement-lawsuit-restore-critical-health-information-federal-websites/>
2. Beitsch, R. (2025). USAID order to delete classified records sparks flurry of litigation. *The Hill*. <https://thehill.com/homenews/administration/5191064-usaids-document-destruction/>
3. Brady, J. (2025). More environmental data is deleted in Trump's second term. *NPR*. <https://www.npr.org/2025/08/08/nx-s1-5495338/climate-change-environment-websites-trump>
4. Cheney, K., & Gerstein, J. (2025). Appeals court rules Trump clamp-down on spending data defies Congress' authority. *Politico*. <https://www.politico.com/news/2025/08/09/appeals-court-rules-trump-clamp-down-on-spending-data-defies-congress-authority-00501348>
5. Clark, L. (2026). Judge dismisses lawsuit over feds' climate data erasure. *E&E News*. <https://subscriber.politicopro.com/article/eenews/2026/03/13/judge-dismisses-lawsuit-over-feds-climate-data-erasure-cw-00824893>
6. Dayak, S., & Kramer, A. (2026). Federal Data Is Disappearing. *NOTUS*. <https://www.notus.org/trump-white-house/federal-data-is-disappearing>
7. Fowler, S., Joffe-Block, J., & Bond, S. (2026). The government is investigating new claims that DOGE misused Social Security data. *NPR*. <https://www.npr.org/2026/03/11/nx-s1-5745153/doge-social-security-data-whistleblower-investigation>
8. Garza, F. (2026). After a lawsuit, USDA agrees to share climate risk data with farmers. *Government Executive*. <https://www.govexec.com/management/2026/03/after-lawsuit-usda-agrees-share-climate-risk-data-farmers/411848/>
9. Hartman, M. (2025). Federal data has been disappearing under Trump. *Marketplace*. <https://www.marketplace.org/story/2025/07/28/federal-data-has-been-disappearing-under-trump>
10. Heckman, J. (2025). 'Bedrock' federal data sets are disappearing, as statistical agencies face upheaval. *Federal News Network*. <https://federalnewsnetwork.com/big-data/2025/12/bedrock-federal-data-sets-are-disappearing-as-statistical-agencies-face-upheaval>
11. Hill, C. (2025). GAO says OMB takedown of apportionments website violates federal statutes. *Fedscoop*. <https://fedscoop.com/gao-omb-takedown-apportionments-website-federal-statutes/>
12. Hirji, Z. (2025). Six Environmental Mapping Tools the White House Doesn't Want You to See. *Bloomberg*. <https://www.bloomberg.com/news/articles/2025-05-07/six-environmental-mapping-tools-the-white-house-doesn-t-want-you-to-see>
13. Jones, L. A. (2025). A Shortlist of Federal Data the Trump Administration Has Tamed or Destroyed. *Talking Points Memo*. <https://talkingpointsmemo.com/news/a-shortlist-of-federal-data-the-trump-administration-has-tamed-or-destroyed>
14. Katz, E. (2025). Spending transparency data posted by Trump budget office after court order. *Government Executive*. <https://www.govexec.com/management/2025/08/spending-transparency-data-posted-trump-budget-office-after-court-order/407548/>
15. Kiersz, A. (2026). America's economy is 'driving through the fog.' *Business Insider*. <https://www.businessinsider.com/disappearing-economic-data-bad-economy-recession-unemployment-bls-jobs-report-2026-2>
16. Koebler, J. (2025). Archivists Work to Identify and Save the Thousands of Datasets Disappearing from Data.gov. *404 Media*.
17. Kutz, A. (2025). How much federal data has Trump really purged? *NewsNation*. <https://www.newsnationnow.com/politics/trump-federal-datasets-websites/>
18. Levenstein, M., & Kubale, J. (2025). Data that taxpayers have paid for and rely on is disappearing – here's how it's happening and what you can do about it. *The Conversation*. <https://theconversation.com/data-that-taxpayers-have-paid-for-and-rely-on-is-disappearing-heres-how-its-happening-and-what-you-can-do-about-it-251787>
19. Malesky, E. (2025). *2025 Letter from the Director - Duke Center for International Development*. <https://dcid.sanford.duke.edu/2025-letter-director/>
20. Mandel, K. (2026). The Women Saving America's Climate Data. *TIME*. <https://time.com/7344773/women-saving-federal-climate-data/>
21. Noor, D. (2025). Green groups sue Trump administration over climate webpage removals. *The Guardian*. <https://www.theguardian.com/us-news/2025/apr/15/trump-climate-webpage-removal-lawsuit>
22. O'Leary, M. (2025). Data Rescue Project Thwarts Government Censorship Wave. *Information Today*. <https://www.infotoday.com/IT/nov25/OLeary-Data-Rescue-Project-Thwarts-Government-Censorship-Wave.shtml>
23. Palmer, K. (2025). Preserving the Federal Data Trump Is Trying to Purge. *Inside Higher Ed*. <https://www.insidehighered.com/news/government/science-research-policy/2025/06/10/preserving-federal-data-trump-trying-purge>
24. Rabin, R., & Mandavilli, A. (2025). CDC Web Pages and Data Vanish Following Trump's DEI and Gender Orders. *The New York Times*. <https://www.nytimes.com/2025/01/31/health/trump-cdc-dei-gender.html>
25. Robbins, R. (2025). *STAT is backing up and monitoring CDC data in real time: See what's changing*. *STAT News*. <https://www.statnews.com/2025/02/14/tracking-cdc-data-changes-trump-executive-order-targets-gender/>
26. Satter, R. (2025). Harvard Law Library acts to preserve government data amid sweeping purges. *Reuters*. <https://www.reuters.com/world/us/harvard-law-library-acts-preserve-government-data-amid-sweeping-purges-2025-02-06/>
27. Schilling, E., & Slowey, E. (2026). IRS Improperly Shares Immigrants' Data with ICE: Explained. *Bloomberg Tax*. <https://news.bloombergtax.com/daily-tax-report/irs-overshares-thousands-of-immigrants-data-with-ice-explained>
28. Smith, J.-M. (2025). USDA cancels survey tracking how many Americans struggle to get enough food. *NPR*. <https://www.npr.org/2025/09/22/nx-s1-5549115/usda-food-insecurity-survey-hunger>
29. Stone, R. (2025). *Researchers from China and five other 'countries of concern' barred from NIH databases*. *Science*. <https://www.science.org/content/article/researchers-china-and-five-other-countries-concern-barred-nih-databases>
30. Stuessy, M. M., & Knoedl, T. R. (2026). *Availability of Federal Data: Policy Considerations for Disclosure, Preservation, and Governance*. Congressional Research Service, Library of Congress. <https://www.congress.gov/crs-product/R48889>
31. Wang, H. L. (2026). The Trump administration is adding a citizenship question to the 2030 census. *NPR*. <https://www.npr.org/2026/03/09/nx-s1-5613878/us-census-citizenship-question-redistricting>
32. Willson, M. (2025). Groups archive environmental justice data scrapped by Trump. *E&E News*. <https://www.eenews.net/articles/groups-archive-environmental-justice-data-scrapped-by-trump/>

33. Data Foundation. (2025). *Taking Stock of Federal Open Data in 2025*. Data Foundation Blog. <https://datafoundation.org/news/blogs/707/707-Taking-Stock-of-Federal-Open-Data-in->
 34. FitzSimons, C. USDA's Decision to End 30-Year Food Security Report Will Hide the Struggle of Millions of Families to Put Food on the Table. In *Food Research and Action Center*. Retrieved March 4, 2026, from <https://frac.org/news/foodsecuritysurveyterminationsept25>
 35. National Science and Technology Council. (2023). *A Framework for Federal Scientific Integrity Policy and Practice*. White House Office of Science and Technology Policy. <https://bidenwhitehouse.archives.gov/wp-content/uploads/2023/01/01-2023-Framework-for-Federal-Scientific-Integrity-Policy-and-Practice.pdf>
 36. Rep. Don Beyer. (2025). *79 U.S. Representatives Demand the Restoration of Public Access to Federal Data Sets Purged by the Trump Administration*. Official Website of U.S. Representative Don Beyer. <https://beyer.house.gov/news/documentsingle.aspx?DocumentID=6384>
 37. United States Congress. (2019). *Foundations for Evidence-Based Policymaking Act of 2018* (No. P.L. 115-435; Issue P.L. 115-435). United States Congress.
-

The Federal Data Catalog

The Federal Data Catalog (FDC) is the United States Federal Government's primary data portal for the public through [Data.gov](#). Data.gov provides the FDC, as well as additional resources that support federal agencies and the public regarding data governance of federal public data assets (principles and practices involved in the management and sharing of data). The site was launched by the General Services Administration (GSA) on May 21, 2009 at the direction of Vivek Kundra, the first Federal Chief Information Officer ([Kundra, 2009](#)) in response to the January 2009 Presidential Memorandum on Transparency and Open Government. The memorandum directed federal agencies to harness technology to promote transparency, participation, and accountability in government ([Obama, 2009](#)). The Office of Management and Budget's (OMB) M-10-06 "Open Government Directive" that followed later that year went further, requiring that all federal agencies post at least three high-value datasets online and register them on Data.gov within 45 days ([Office of Management and Budget, 2009](#)). By the end of 2010, most federal agencies had published data on the platform, and by 2012 Data.gov's holdings were regularly drawn upon by civil society organizations, researchers, and private businesses ([U.S. General Services Administration, 2024](#)). In 2013, the Obama Administration subsequently expanded its open government data policies with the back-to-back release of Executive Order 13642, which declared open and machine-readable data "the new default for government information," and the Office of Management and Budget (OMB) implementation guidance known as M-13-13, "Open Data Policy - Managing Information as an Asset."

M-13-13 required federal agencies to create enterprise data inventories (the precursor to what would later be called comprehensive data inventories (CDIs)), publish public data listings from those inventories at [agency.gov/data.json](#), and have GSA populate those assets in Data.gov. The Foundations for Evidence-Based Policymaking Act of 2018 (The Evidence Act) signed into law on January 14, 2019 by President Trump codified much of the policy guidance articulated in M-13-13 ([United States Congress, 2019](#)) in its Title II, the Open, Public, Electronic, and Necessary Government Data Act (OPEN Government Data Act or OGDAs). Under this law, federal agencies are required to maintain data assets as open data using standardized, machine-readable, non-proprietary formats, and the associated metadata for all of their data assets must be included in the FDC ([Congressional Research Service, 2022](#)). The Evidence Act required each agency to create and maintain a CDI that accounts for all data assets the agency "creates, collects, controls, or maintains." It also required the OMB to promulgate additional implementation guidance within a year of enactment of the law.

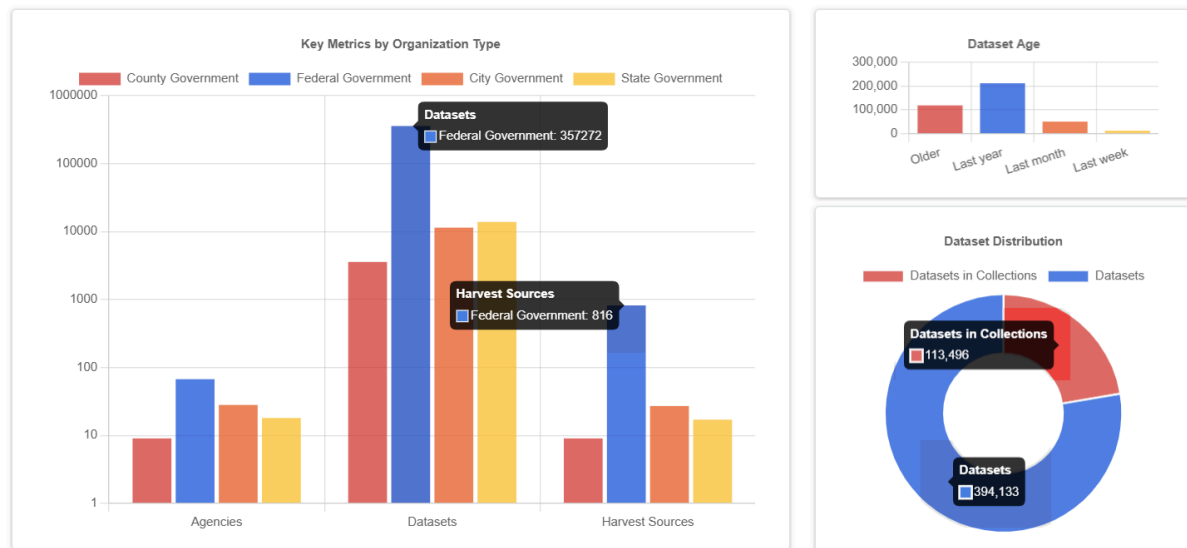
That implementation guidance was slow to materialize. Six years and a day after the law's enactment, as one of the final policy actions of the Biden Administration, OMB finally released M-25-05 "Phase 2 Implementation of the Foundations for Evidence-Based Policymaking Act of 2018: Open Government Data Access and Management Guidance" ([Office of Management and Budget, 2025; Alder, 2025](#)). The guidance requires all federal agencies and GSA to adopt an updated metadata schema for their CDIs and the FDC (known as [DCAT-US 3.0](#)), sets minimum metadata requirements required by the Evidence Act, provides guidance on how to prioritize data asset classification and review for public release as open government data assets, and clarifies what constitutes a data asset, public data asset, and open government data asset under the statute ([Office of Management and Budget, 2025](#)). Critically, M-25-05 rescinds and replaces M-13-13 while carrying forward its open-by-default philosophy and giving Chief Data Officers (a role that was statutorily created by the Evidence Act) specific responsibilities they did not have in 2013. Nearly all of the requirements in M-13-13 were preserved in M-25-05, including how agencies should publish CDIs in a common location. Among responsibilities articulated to the CDOs is ensuring that their agencies publish an open data plan and that the CDIs are updated and complete.

These CDIs form the harvest sources that populate the FDC, meaning that the quality, completeness, and timeliness of Data.gov is directly dependent on each agency's diligence in maintaining its own inventory. The lengthy gap between the law's passage and the issuance of implementation guidance had practical consequences: agencies proceeded with inconsistent metadata practices and varying interpretations of what belonged in a CDI, contributing to many of the challenges in using the FDC as a barometer for federal data integrity described below.

Challenges with using the Federal Data Catalog in monitoring open government data

The technical architecture of Data.gov has evolved considerably since its inception. The original portal relied on a custom-built catalog system, but in 2014 it was relaunched on CKAN (Comprehensive Knowledge Archive Network, which turned 20 years old this year ([Popova, 2026](#))), an open-source data management platform developed by the Open Knowledge Foundation ([U.S. General Services Administration, 2014](#)). The CKAN-based catalog introduced a robust public API, federated search across geospatial and non-geospatial datasets, and a harvesting infrastructure that allows individual agencies to maintain their own metadata sources while GSA automatically aggregates them into a central catalog on a scheduled basis ([U.S. General Services Administration, 2024](#)). Rather than requiring agencies to manually submit records to a single curator, the harvesting model means that any additions, modifications, or deletions made to an agency's metadata inventory are reflected in Data.gov at the next scheduled harvest - a design that has significant implications for interpreting fluctuations in the platform's reported dataset count.

Below, you can view breakdowns of the type, age, and distribution of [catalog.data.gov](#) datasets. For information on top dataset page views, file downloads, and external link clicks broken down by agency, see [data.gov/metrics](#). To learn more about the Metrics section, see the [Digital.gov blog post](#).



Last Updated: Mon, 09 Mar 2026 05:00:29 GMT

Figure: A composite of screenshots of the statistics on the [Data.gov](#) landing page from 03/09/2026. The statistics can be recovered using the CKAN API for live values or through snapshots captured by the WBM.

Inaccurate Assumptions

A clear understanding of what the FDC is and is not is a necessary precondition for using it to monitor open government data. Two common misunderstandings are particularly consequential.

The first is the assumption that the FDC is a data repository. It is not. As the platform's own documentation states, the Data.gov catalog contains only metadata about datasets — including URLs, descriptions, and other descriptive information — but not the actual data assets themselves ([U.S. General Services Administration, 2024](#)). The underlying data continue to reside on agency servers, portals, and websites, with Data.gov providing only a pointer to where they can be found. The confusion about the FDC being a data repository may have resulted from language in the Evidence Act itself:

“The Administrator and the Director shall ensure that agencies can submit public data assets, or links to public data assets, for publication and public availability on the interface.” (P.L. 115-435)

Because of the history with how Data.gov was initially rolled-out as an index of metadata information about datasets only, then reinforced by OMB in M-13-13, the infrastructure of the FDC never supported acquisition of actual datasets. That the Evidence Act allows for either data or links to the data provides a pathway for the FDC to add this functionality at a later date (or to replace it with a successor catalog in the future, a possibility that is also articulated in M-25-05). This distinction matters enormously when the FDC is used as a proxy for monitoring data availability: a dataset can be listed in the FDC while the underlying data are inaccessible, modified, or even deleted from the agency's own systems, and conversely, a dataset can be taken down from an agency website without any corresponding change in the FDC's count if the agency's CDI metadata record remains intact.

The second, and arguably more consequential, misunderstanding concerns the dynamic nature of the FDC's dataset count. Because Data.gov is continuously and automatically harvesting metadata from agency sources on varying schedule the topline number of datasets displayed on Data.gov's landing page fluctuates routinely as part of normal operations ([U.S. General Services Administration, 2024](#)). The number reflects the state of agency CDIs at the time of the most recent harvest, and it changes whenever agencies add, update, or remove entries from their own metadata inventories. These fluctuations are expected in the normal operations of the FDC. Some agencies sources are harvested daily, others weekly or monthly.

These two points were at the center of significant confusion during early 2025, when multiple press outlets and monitoring organizations cited changes in Data.gov's total dataset count as a primary measure of data removals by the second Trump Administration ([Koebler, 2025](#); [Kutz, 2025](#); [Mauran, 2025](#)). Beginning in late January 2025, following a series of executive orders directing federal agencies to remove websites and data products deemed inconsistent with the new administration's policy priorities, the total number of datasets listed on Data.gov's homepage did decline noticeably from roughly 308,000 at the time of the inauguration to approximately 304,600 by late February 2025, a reduction of some 3,400 entries ([Kutz, 2025](#)). News coverage, social media, and advocacy organizations amplified these figures as evidence of a systematic data purge, leading members of Congress to respond with inquiries about why Data.gov was reporting such losses ([Rep. Don Beyer, 2025](#)). However, owing to the dynamic nature of the Data.gov harvesting routines, a [forensic audit of the system done around](#) that time in conjunction with this project reveals that the true change netted out to about zero. Ironically, all of the reported accounts were *technically* relying on the wrong count as the Data.gov topline does not include datasets that are part of collections

(there are at least another 100,000 of those). The next iteration of the FDC will have a more accurate count once the site is fully released (see the [beta-version here](#), which lists 515,111 datasets as of the date of this report).

Despite likely inaccuracies, attention to those figures was not entirely misplaced. Real and consequential removals of federal datasets from public access did occur in 2025. These have been documented in detail by health policy researchers ([KFF, 2025](#)), journalists ([Robbins, 2025](#); [Palmer, 2025](#)), and advocacy organizations ([National Security Archive, 2025](#)). However, experts cautioned against treating the topline Data.gov count as a reliable or complete measure of those removals. At a public webinar convened by the Data Foundation in October 2025, panelists that included a former GSA Data.gov team lead and a former OMB senior statistician explained that the dataset count “is normally going to change all the time” due to routine harvesting activity, and that relying on it as a barometer of data integrity is fundamentally problematic for several compounding reasons ([Data Foundation, 2025](#)). The FDC may list a dataset as available while the link to that dataset points to an error page, and conversely, significant datasets that have never been properly indexed in agencies’ CDIs will not appear in (or perhaps even disappear from) the catalog at all, regardless of what happens to the actual data ([Data Foundation, 2025](#)). As one panelist summarized, the catalog represents information *about* data, not the data themselves, and the integrity of that information is only as strong as the metadata practices of the agencies that feed it.

Incomplete Catalog

There are currently over 500,000 data assets inventoried in the FDC (including datasets that are part of aggregated collections). While this is a non-trivial figure, it represents just a tiny fraction of all the data assets that federal agencies possess. Most data assets that federal agencies hold have not gone through the prioritization and review process required by M-25-05 and the Evidence Act. Even when data assets are publicly available through internet download they may be not be included in a CDI.

One example of publicly available data assets held by a federal agency but not indexed in the FDC comes from the Department of State’s President’s Emergency Plan for AIDS Relief (PEPFAR) [dataset website](#), which was previously administered by USAID. The website currently lists seven datasets available for download in two programs. None of those datasets are listed in either the Department of State’s current or historical data inventories nor those formerly maintained by USAID or CDC (which also collected data through PEPFAR). The absence of these datasets in the FDC means that when 17 previously available datasets were removed by this Administration in 2025 from the website, those changes would not be reflected on Data.gov and its summary statistics.

Incomplete coverage of agency data holdings is not solely the result of agency negligence or deliberate omission. Under the harvesting model, Data.gov can only index what agencies have documented in their CDIs. For agencies whose CDO and CIO offices have been historically under-resourced — a condition that the Phase II guidance for the Evidence Act explicitly acknowledged as a systemic problem ([Office of Management and Budget, 2025](#)) — maintaining a complete and current inventory of all data assets is a substantial operational and resource burden. Moreover, the Evidence Act’s definition of a data asset is broad enough to encompass a vast range of information products, many of which agencies have never systematically catalogued. The statutory requirement to account for *all* data assets created, collected, or maintained by an agency subject to an assessment of whether the costs and benefits to the public of converting a data asset into a machine-readable format are favorable. In practice, Congress’s intent of openness-by-default describes an aspirational standard that no federal agency has yet achieved.

The incompleteness of the FDC was further exacerbated in 2025 by workforce reductions affecting the GSA team responsible for maintaining Data.gov itself. At the beginning of 2025, that team comprised five government employees along with contractor support; by October of the same year, only two government employees remained in the office, with reduced contractor support and a heavier workload as a result of the departures of their colleagues ([Data Foundation, 2025](#)). These staffing changes have had direct implications for GSA’s ability to fulfill its own statutory obligations under M-25-05, including maintaining a current and accurate FDC.

Inaccurate Metadata

Incomplete metadata information is a major issue with both individual agencies’ CDIs and, as a result, the FDC. The Evidence Act requires that all agencies publish a CDI and that those entries appear aggregated in the FDC (i.e., as provided by Data.gov).

The Evidence Act and the subsequent OMB guidance in M-25-05 lay out the minimal requirements for such metadata. Moreover, the updated DCAT-US 3.0 metadata schema, required by OMB per M-25-05 to be used by all federal agencies and GSA in promulgating the inventories and the FDC, standardizes the formatting of both optional and required metadata ([Office of Management and Budget, 2025](#)). However, significant gaps remain as agencies’ CDO and CIO offices have been historically under-resourced with respect to maintaining their CDIs. Among those most pertinent to understanding the integrity of open government data assets is the failure of agencies to include, or maintain, the exact URLs where individual datasets can be downloaded by the public. Under the DCAT-US schema, the distribution property in the JSON files should include the following information in its array:

```
"distribution": [
  {
    "@type": "dcat:Distribution",
    "downloadURL": "",
    "mediaType": "",
    "title": ""
  }
]
```

This is in addition to the optional inclusion of a landing page that describes a dataset or set of datasets in a separate field:

```
"landingPage": ""
```

In some cases, an agency will provide a landing page URL that provides additional links to download individual datasets. Again, the best practice — and requirement — is to include those individual URLs in the CDI itself under the distribution property in the JSON file. For example, the VA’s National Center for Veterans Analysis and Statistics has a landing page describing the [Geographic Distribution of VA Expenditures \(GDX\)](#) data. The landing page includes links to the annual Expenditures Tables, which are provided as Excel files through separate downloadable links (note that, technically, Excel x1s files are proprietary and out of compliance with the Evidence Act as they are not in an open format). The VA’s CDI, however, only partially indexes each of these files.

Misclassifications of Data Assets

As discussed above regarding the aggregating of datasets into collections, the total number of entries in the FDC, which is used to populate the totals reported on the Data.gov landing page, is not an accurate estimate of the total number of data assets in the FDC. However, additional errors are introduced into this count simply because agencies often misclassify information as a data asset.

Many agencies include information products, such as reports, infographics, documentation, data tools, software, and other documents in their CDIs. These do not meet the statutory definition of a data asset under the Evidence Act (e.g., “a collection of data elements or data sets that may be grouped together”). An entire section of the OMB implementation guidance for M-25-05 (Section 4) was dedicated to helping agencies understand this definition of a data asset so they can more accurately update their CDIs ([Office of Management and Budget, 2025](#)).

For example, the CDC includes an infographic in their CDI “Going Smokefree Matters - In Your Home Infographic.” This document is clearly not a dataset — however, the metadata entry in the CDI that is harvested by Data.gov wrongly indicates that it is a dataset in the *@type* field:

```
{
  "@type": "dcat:Dataset",
  "accessLevel": "public",
  "bureauCode": [
    "009:20"
  ],
  "contactPoint": {
    "@type": "vcard:Contact",
    "fn": "OSHData Support",
    "hasEmail": "mailto:nccdosinquiries@cdc.gov"
  },
  "description": "Explore the Going Smokefree Matters - In Your Home Infographic which outlines key facts related to the effect",
  "distribution": [
    {
      "@type": "dcat:Distribution",
      "downloadURL": "https://data.cdc.gov/download/k4xj-uge6/application/pdf",
      "mediaType": "application/pdf"
    }
  ],
  ...
}
```

Following the CDC’s data inventory API [endpoint for this entry](#), reveals that the infographic is entered as a “file” in the *assetType* field:

```
{
  "id" : "k4xj-uge6",
  "name" : "Going Smokefree Matters - In Your Home Infographic",
  "assetType" : "file",
  "averageRating" : 0,
  "blobFilename" : "Going Smokefree Matters - In Your Home Infographic.pdf",
  "blobFileSize" : 552143,
  "blobId" : "ac7da77d-4178-4ab1-951e-d7b58d03c01a",
  "blobMimeType" : "application/pdf; charset=binary",
  ...
}
```



Figure: An example of an infographic by the Centers for Disease Control and Prevention inaccurately accessioned as a data asset in the FDC. Available for download at: <https://data.cdc.gov/download/k4xj-uge6/application/pdf>

Compare that with the CDC's API endpoint of their "[NNDSS - Table II. West Nile virus disease](#)", which is also listed [as a dataset in Data.gov](#), albeit correctly:

```
{
  "id" : "r7hc-32zu",
  "name" : "NNDSS - Table II. West Nile virus disease",
  "assetType" : "dataset",
  "attribution" : "Division of Health Informatics and Surveillance (DHIS), Centers for Disease Control and Prevention",
  "averageRating" : 0,
  "category" : "NNDSS",
  ...
}
```

Errors are present

The process of adding, modifying, or removing entries into the inventory can be error-prone. Even with strong data governance principles, such as using controlled vocabularies and standard operating procedures for naming conventions and metadata, mistakes happen. Whether these are the result of one-off human data entry mistakes, or systematically encoded into automation pipelines, the errors manifest across the FDC. Both the EPA and VA use-cases described above demonstrate examples of typos in the titles and descriptions of CDI entries. However, more consequential errors are also present in the FDC.

Take, for instance, the US International Trade Commission's (USITC) highly regularized releases of the [Harmonized Tariff Schedule of the United States](#) (HTSUS). This high-profile dataset is used by federal agencies and industry to understand the tariff rates and statistical

categories of all goods imported into the United States throughout a given calendar year. Typically, the USITC follows a naming convention for this dataset of [TITLE]([YEAR]), where TITLE is the name of the dataset and YEAR is the year of coverage. The URLs generated by GSA when creating landing pages in the FDC use the same information. In 2025, however, an error was introduced somewhere in the processing pipeline and the URL for the 2025 release has an FDC landing page URL on Data.gov that implies it is the 2024 release:

<https://catalog.data.gov/dataset/harmonized-tariff-schedule-of-the-united-states-2024>

This URL was previously the landing page of the actual 2024 release, evident from an early 2024 WBM snapshot:

<https://web.archive.org/web/20240528054332/https://catalog.data.gov/dataset/harmonized-tariff-schedule-of-the-united-states-2024>

Thankfully, the Data.gov harvester and landing page generation routines have reasonable fail-safes. As a dynamic site however, rather than preventing the overwriting of an existing URL, the system generates landing page URLs as it encounters them in its FDC source, which is sorted from most recent to least by default. As a result, the true 2024 release has a new active landing page URL with a random suffix appended to the end:

<https://catalog.data.gov/dataset/harmonized-tariff-schedule-of-the-united-states-2024-41c71>

This type of error is consequential because it limits the ability of users to construct reliably persistent URLs from source information. It also breaks the intended endpoint of links within existing content, directing users to potentially wrong sources. With this specific case, the error appears to have induced a data duplication issue with at least one data preservation initiative. For instance, there are [two entries](#) in Harvard's Data.gov archive search results for the 2024 HTSUS, both appearing to archive the same 2024 data.

Perhaps the most egregious error from a data integrity perspective present in the FDC is the pervasive problem of invalid URLs that purport to link to data assets. These errors can arise from typos and other input errors (i.e., filename encoding issues) in the harvest sources. However, one more insidious cause is link rot — that is, the deprecation of a valid URL to an invalid URL over time as files are moved, websites are restructured, et cetera. As an example of link rot, consider the Office of Management and Budget's [Public Budget Database - Governmental receipts 1962-Current](#) entry in the FDC. When the metadata for this entry was last updated on 3/22/2024, the download URL pointed to: https://www.whitehouse.gov/wp-content/uploads/2024/03/receipts_fy2025.xlsx. Because the White House website was taken down and slowly rebuilt after the new Trump Administration took office—and because the Administration moved to excise [spending data from public view](#)—this URL was lost to link rot. Rather than suffer this avoidable fate, the URL should have been updated in the OMB metadata to point to the National Archives' copy of the Biden White House website during the transition: https://bidenwhitehouse.archives.gov/wp-content/uploads/2024/03/receipts_fy2025.xlsx.

Conclusion

Data.gov and the FDC are critical national data infrastructure, providing the public with a significant resource to discover data supported by taxpayers. However, misunderstandings about how that resource can be appropriately used and the nature of its limitations can lead to confusion over how it should be considered in the context of federal data integrity issues experienced in 2025 (and into early 2026). Importantly, the topline count of federal data assets reported on Data.gov is a signal only of what is indexed at a moment in time and subject to both mundane and irregular changes to its harvest sources.

This mismatch between the FDC's signals and underlying data reality runs in both directions. Some of the most significant documented data removals of 2025 involved datasets and websites that were never indexed in the FDC in the first place. These included [PEPFAR datasets](#), [OMB apportionments data](#), and many [data tools](#). Their removal was invisible to any analysis based on Data.gov's dataset count. At the same time, many entries in the FDC contained broken or stale URLs pointing to resources that had been reorganized, taken offline, or deleted long before 2025, a phenomenon known as link rot. The net effect is that the FDC's total dataset count is an unreliable single-number proxy for the actual availability of federal data, capable of both overcounting (by including misclassified assets, stale records, and entries without working URLs) and undercounting (by missing datasets that are publicly available but never properly inventoried).

References

1. Alder, M. (2025). White House finalizes OPEN Government Data Act guidance, restarts CDO Council. *FedScoop*. <https://fedscoop.com/white-house-open-government-data-act-restarts-cdo-council/>
2. Koebler, J. (2025). Archivists Work to Identify and Save the Thousands of Datasets Disappearing from Data.gov. *404 Media*.
3. Kundra, V. (2009). *Data.gov Launch Announcement*. U.S. General Services Administration, Technology Transformation Services. <https://data.gov/timeline/>
4. Kutz, A. (2025). How much federal data has Trump really purged? *NewsNation*. <https://www.newsnationnow.com/politics/trump-federal-datasets-websites/>
5. Maura, C. (2025). Thousands of datasets from Data.gov have disappeared since Trump's inauguration. What's going on? *Mashable*. <https://mashable.com/article/government-datasets-disappear-since-trump-inauguration>
6. Obama, B. (2009). *Memorandum on Transparency and Open Government*. Presidential Memorandum. <https://obamawhitehouse.archives.gov/the-press-office/transparency-and-open-government>
7. Palmer, K. (2025). Preserving the Federal Data Trump Is Trying to Purge. *Inside Higher Ed*. <https://www.insidehighered.com/news/government/science-research-policy/2025/06/10/preserving-federal-data-trump-trying-purge>
8. Popova, Y. (2026). CKAN Turns 20: Two Decades of Open Data Infrastructure. *CKAN Blog*. <https://ckan.org/blog/ckan-turns-20-two-decades-of-open-data-infrastructure>
9. Robbins, R. (2025). *STAT is backing up and monitoring CDC data in real time: See what's changing*. STAT News. <https://www.statnews.com/2025/02/14/tracking-cdc-data-changes-trump-executive-order-targets-gender/>
10. Congressional Research Service. (2022). *The OPEN Government Data Act: A Primer* (No. IF12299; Issue IF12299). Congressional Research Service. <https://www.congress.gov/crs-product/IF12299>
11. Data Foundation. (2025). *Taking Stock of Federal Open Data in 2025*. Data Foundation Blog. <https://datafoundation.org/news/blogs/707/707-Taking-Stock-of-Federal-Open-Data-in->
12. KFF. (2025). *A Look at Federal Health Data Taken Offline*. KFF Policy Watch. <https://www.kff.org/policy-watch/a-look-at-federal-health-data-taken-offline/>
13. National Security Archive. (2025). *Disappearing Data: Trump Administration Removing Climate Information from Government Websites*. National Security Archive, George Washington University. <https://nsarchive.gwu.edu/briefing-book/climate-change->

- [transparency-project-foia/2025-02-06/disappearing-data-trump](#)
14. Office of Management and Budget. (2025). *Phase 2 Implementation of the Foundations for Evidence-Based Policymaking Act of 2018: Open Government Data Access and Management Guidance*. <https://www.whitehouse.gov/wp-content/uploads/2025/01/M-25-05-Phase-2-Implementation-of-the-Foundations-for-Evidence-Based-Policymaking-Act-of-2018-Open-Government-Data-Access-and-Management-Guidance.pdf>
 15. Office of Management and Budget. (2009). *OMB Memorandum M-10-06 - Open Government Directive*. <https://obamawhitehouse.archives.gov/open/documents/open-government-directive>
 16. Rep. Don Beyer. (2025). *79 U.S. Representatives Demand the Restoration of Public Access to Federal Data Sets Purged by the Trump Administration*. Official Website of U.S. Representative Don Beyer. <https://beyer.house.gov/news/documentsingle.aspx?DocumentID=6384>
 17. U.S. General Services Administration. (2024). *Data.gov Program Timeline*. Data.gov. <https://data.gov/timeline/>
 18. U.S. General Services Administration. (2014). *Data.gov CKAN Catalog Launch*. Data.gov Blog. <https://data.gov/announcements/datagov-ckan-catalog/>
 19. U.S. General Services Administration. (2024). *Data.gov User Guide*. Data.gov. <https://data.gov/user-guide/>
 20. United States Congress. (2019). *Foundations for Evidence-Based Policymaking Act of 2018* (No. P.L. 115-435; Issue P.L. 115-435). United States Congress.
-

Administrative Risks to Data

Federal data face a set of administrative risks that are often more subtle than efforts that result in direct interference, discontinuation, or data erasure but are often more consequential in practice. These risks emerge from the day-to-day decisions, priorities, and resource constraints that shape how agencies steward information. Even when legal authorities to collect and publish data remain intact, administrative choices, such as shifting leadership priorities, reorganizing offices, or reallocating staff, can weaken data governance structures and reduce an agency's capacity to maintain high-quality, well-documented, and publicly accessible datasets. Inconsistent implementation of government-wide policies, uneven adoption of data standards, and gaps in internal oversight can compound these vulnerabilities, creating an environment where data integrity can degrade.

Some examples of administrative risks include budget reductions, hiring freezes, and contract eliminations. These can slow or halt essential data maintenance activities, from collection to metadata updates to system modernization. Agencies may also deprioritize their data programs, quality assurance routines, or public dissemination mandates when confronted with competing programmatic demands, shifting Administration priorities, or a lapse in resources needed to maintain their data assets. These pressures can lead to outdated systems, incomplete documentation, and diminished institutional memory - especially when key personnel depart without structured knowledge transfer. Over time, such administrative erosion can make it difficult for agencies to comply with open-data requirements, respond to oversight inquiries, or support evidence-building activities. (Bowen et al., 2025)

These administrative risks interact with, and are sometimes amplified by, legal mechanisms used by agencies to collect, use, maintain, and disseminate federal data such as the Paperwork Reduction Act (PRA) processes. This chapter reviews several administrative risks poised to open government data during 2025.

References

1. Bowen, C. M. K., Citro, C., Crosby, M., Pierson, S., Potok, N., & Seeskin, Z. (2025). *The Nation's Data at Risk: 2025 Report*. The American Statistical Association. <https://www.amstat.org/policy-and-advocacy/the-nations-data-at-risk-2025-report>

The Richardson Waiver Rescission

One of the more subtle administrative threats to HHS data integrity comes from an obscure exemption of the Administrative Procedures Act in [5 U.S.C. § 553\(a\)\(2\)](#) that states that the notice-and-comment process does not apply to:

“a matter relating to agency management or personnel or to public property, loans, grants, benefits, or contracts.”

While this exemption is broad and applies to all agency rulemakings, the impact of the exemption on HHS activities is perhaps the largest given their outsized role in generating data through grants and benefits programs. For this reason, since 1971 HHS (previously HEW at the time), has waived its exercise of this exemption through an internal policy memorandum known as the Richardson Waiver. The waiver committed HHS to use notice-and-comment procedures for certain categories of rulemaking that the Administrative Procedure Act (APA) does not require to undergo notice and comment, particularly rules involving public property, loans, grants, benefits, or contracts.

On February 28th, 2025, HHS [rescinded the Richardson Waiver](#) “effective immediately,” stating that HHS would follow notice-and-comment procedures only when required by statute and that the APA's exemptions would be applied according to their text ([U.S. Department of Health and Human Services, 2025](#)). Although HHS components retain discretion to use notice-and-comment in particular cases, the default presumption in favor of broader public participation for these categories of actions has been removed.

Potential Consequences on Data Integrity

The Richardson Waiver was not itself a data-specific policy. It did not directly govern any other information policy derived from other statutes such as the Freedom of Information Act, Paperwork Reduction Act, Public Health Act, or the Evidence Act. Nevertheless, many public-facing data practices are shaped by program rules, grant conditions, and contracts at HHS. Observers have characterized the rescission as part of a broader set of administrative changes that may reduce transparency and participatory governance at HHS ([Reiss, 2025](#)).

This procedural shift does not directly repeal any data-specific policy or regulation. However, it changes the process by which many HHS program rules that structure how data are collected, standardized, reported, and disclosed, are developed. As a result, it may have downstream implications for data collection continuity, data integrity, public data access, and public oversight into data collection activities at HHS. As the American Bar Association pointed out, with the waiver rescinded, HHS may implement changes in these areas without notice-and-comment periods unless otherwise required by law ([Mys, 2025](#)). Many stakeholders, including [contributors to dataindex.us](#) ([Maury & Marcum, 2025](#); [Maury & Ross, 2026](#)), have consistently argued that public comment is essential for the maintenance of federal data integrity and the loss of it could impact data collection in several ways, including:

- **Data standards:** Limited public review of revisions to definitions, coding and data collection standards, or collection instruments.
- **Scope of collected elements:** Policy shifts may alter which variables are collected, retained, or prioritized.
- **Feasibility and compliance issues:** Notice-and-comment historically allowed external stakeholders to identify operational barriers that might otherwise degrade data completeness or reliability.
- **Removal of data tools:** Some data generated from grants and from benefits programs have derivative use in data tools and those become at-risk of removal without notice.

Data integrity, specifically, benefits from public input. Without a presumption of notice-and-comment as provided for by the APA, for exempt rule categories that generate data, certain HHS actions may proceed without public notice or input. As a result, the rescission of the Richardson Waiver could implicate data integrity risks, including:

- **Specification errors:** Faster implementation timelines may increase the likelihood of technical inconsistencies.
- **Implementation variability:** Reduced opportunity for pre-issuance clarification may lead to inconsistent interpretation across implementing entities.
- **Reduced transparency of methodological change:** Notice-and-comment generates a formal administrative record explaining why changes were made. Without it, downstream data users may have less documentation to interpret discontinuities or anomalies.

To be clear, these risks are contingent rather than automatic. HHS may still voluntarily use notice-and-comment in particular circumstances. However, the structural incentive toward public input has shifted since the rescission and provides a larger runway for skulduggery by HHS agencies.

Conclusion

The rescission of the Richardson Waiver does not directly eliminate datasets, suspend statutory reporting obligations, or repeal disclosure laws. Its primary effect is procedural: it restores the APA's exemptions for certain categories of HHS rulemaking without a voluntary overlay requiring public input via notice-and-comment.

Because many HHS data systems are shaped through grants, benefits, contracts, and program administration rules, this procedural shift may:

- Increase the likelihood of faster, less publicly vetted changes to data collection frameworks;
- Heighten risks of technical discontinuities or documentation gaps;
- Indirectly affect public data access practices; and,
- Narrow the administrative record available for public oversight.

The full impact will depend on how frequently HHS components elect to use discretionary notice-and-comment going forward and how changes in program rules interact with statutory data requirements. Nonetheless, the rescission alters a substantial portion of HHS information policy, much of which has structured federal health data systems for decades and represents a non-trivial threat to its integrity.

References

1. Maury, M., & Marcum, C. (2025). How You Can (and should) Shape Federal Data Collections. *America's Data Index*. <https://dataindex.us/newsletter/article/6cfecae3-3c89-487e-b8f7-cfed85ded6c7>
2. Maury, M., & Ross, D. (2026). *Take Action: How to Write a Public Comment on Federal Data*. America's Data Index. <https://dataindex.us/events/Take-Action-How-to-Write-a-Public-Comment-on-Federal-Data>
3. Mys, A. (2025). HHS Rescinds Richardson Waiver, Reducing Public Input in Rulemaking. *American Bar Association*. https://www.americanbar.org/groups/health_law/news/2025/3/hhs-rescinds-richardson-waiver-reducing-public-input-in-rulemaking/
4. Reiss, D. (2025). Administrative Changes That Decrease Transparency at HHS. *The Regulatory Review*. <https://www.theregview.org/2025/03/24/reiss-administrative-changes-that-decrease-transparency-at-hhs/>
5. U.S. Department of Health and Human Services. (2025). *Policy on Adhering to the Text of the Administrative Procedure Act*. Federal Register. <https://www.federalregister.gov/documents/2025/03/03/2025-03300/policy-on-adhering-to-the-text-of-the-administrative-procedure-act>

Paperwork Reduction Act Exemptions

When Congress enacted the [21st Century Cures Act of 2016](#), it included a targeted administrative change affecting the National Institutes of Health (NIH): Section 2036 carved out a limited exemption from the Paperwork Reduction Act (PRA) for certain research-related information collections. It provides that certain information collections conducted during the course of biomedical research are not subject to the PRA's clearance requirements or process. The exemption was designed to reduce delays that researchers sometimes faced when initiating surveys or other data collection instruments tied to scientific studies ([Riley & Blizinsky, 2017](#)).

As a result, the NIH does not need to seek OMB clearance for qualifying research collections. Because qualifying NIH research collections are exempt from PRA clearance, they are not subject to that structured OMB review and lack OMB's assessment of whether such collections align with government-wide data standards—such as [Statistical Policy Directive Number 15 \(SPD-15\)](#) regarding race and ethnicity data—or whether they duplicate existing federally-funded data efforts. There is also no notice-and-comment period required for these collections, reducing public oversight and input into how NIH conducts research surveys.

The exemption does not eliminate other safeguards. NIH research remains subject to Institutional Review Board oversight, human subjects protections under the Common Rule, privacy protections, and other statutory constraints. But the PRA's centralized, government-wide coordination function—particularly OMB's role in reviewing redundancy and standards compliance—does not apply to exempt collections.

Conclusion

In short, the NIH exemption in the 21st Century Cures Act was intended to accelerate biomedical research by reducing administrative friction. The tradeoff is a narrower layer of cross-agency oversight into whether qualifying research data collections are harmonized with federal data standards or unnecessarily duplicative. The exemption, despite the benefits it provides to accelerate biomedical research, poses an administrative risk to federal data integrity by removing a mechanism for federal-wide standards conformity and public input into survey collections.

References

1. Riley, W. T., & Blizinsky, K. D. (2017). Implications of the 21st century cures act for the behavioral and social sciences at the national institutes of health. *Health Education & Behavior*, 44(3), 356–359. <https://doi.org/10.1177/1090198117707964>
-

Information Collection Discontinuation and Revision

The Paperwork Reduction Act (PRA) aims to minimize the public paperwork burden while maximizing the utility of federal data. Under this framework, the Office of Information and Regulatory Affairs (OIRA) within the Office of Management and Budget (OMB) review and approve information collection requests (ICR) that agencies use to generate data from individuals and organizations. Under the PRA, agencies must notify the public of their intent to collect new data or modify an existing collection obtain OIRA approval and an OMB Control Number for new collections. The OMB Control Number is valid for a maximum of three years. While public notice and approval from OIRA is required for nearly all (non-exempt) ICRs by a federal, discontinuation of a collection - and thus the stream of data it generates - may follow a different path depending on the nature of the ICR. Discontinuation represents a real, albeit mundane, administrative risk to the integrity of federal data.

Discontinuation

Agencies typically discontinue collections through one of three mechanisms. The most formal discontinuation mechanism involves ICRs associated with a specific federal regulation. Because these collections are mandated by the Code of Federal Regulations (CFR), agencies cannot unilaterally stop collecting the data. Under 5 CFR Part 1320, they must seek explicit OMB permission to discontinue the ICR and publish a notice in the Federal Register. This process is usually accompanied by a proposed rulemaking to formally remove the requirement, providing high public visibility and an opportunity for stakeholders to comment on the loss of data.

For ICRs not tied to a regulation, the discontinuation process requires fewer steps. An agency can formally discontinue the collection at its own prerogative. If leadership decides the data is no longer relevant to their mission, they can officially close out the ICR. In these instances, the agency posts a notice of discontinuation on the OMB tracking portal, reginfo.gov. This creates an official administrative record but lacks the broader public notification and formal comment period of a Federal Register notice.

A third mechanism is passive expiration. Because every OMB Control Number has a maximum lifespan of three years, an agency can terminate a collection simply by allowing the approval to expire. This can happen due to the natural end of a temporary survey, reduced administrative capacity, or shifting agency priorities. Consequently, the data collection ends quietly without any formal public notification.

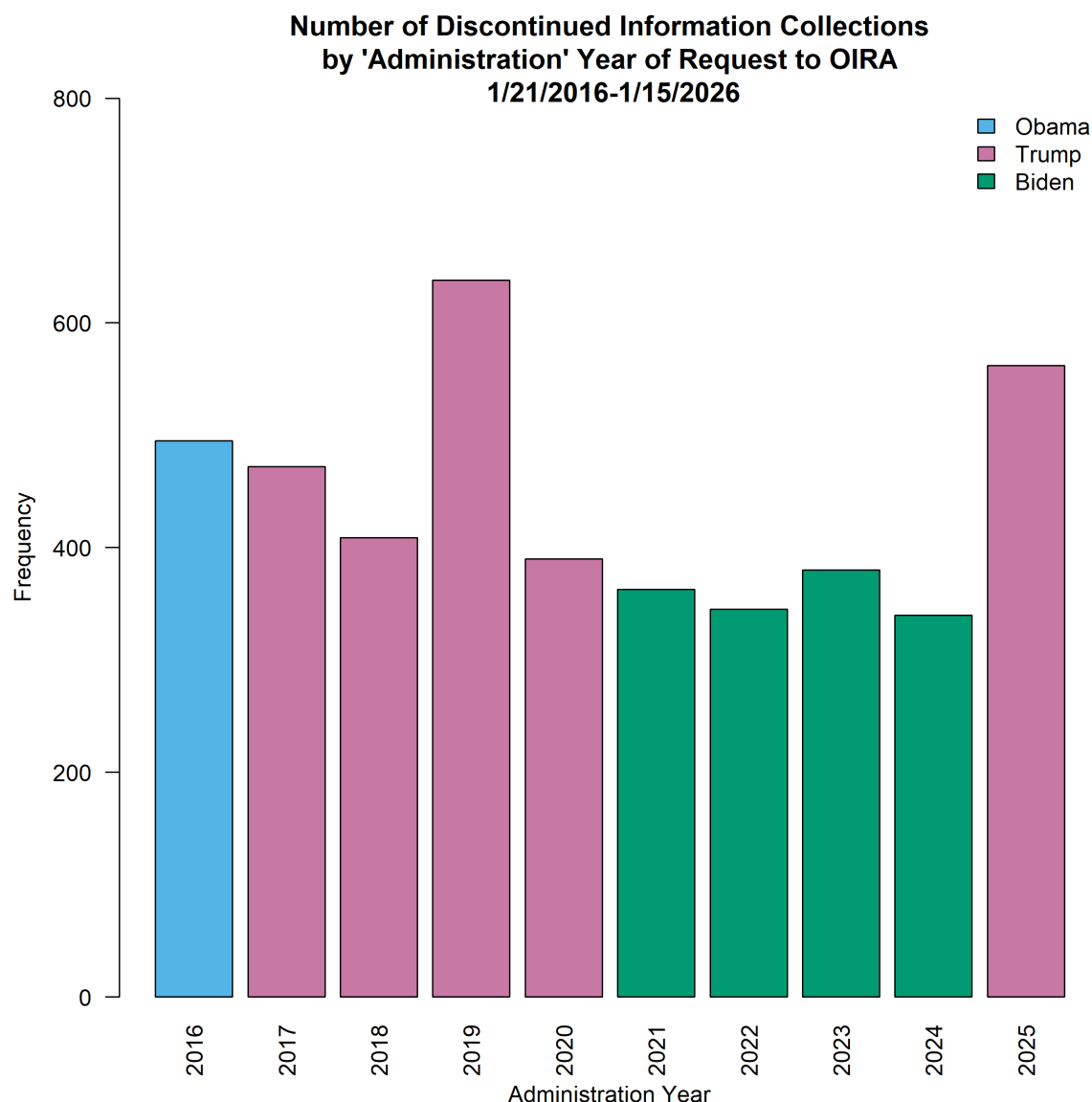


Figure: The [pra-icr-tools package](#) was used to extract data from [OIRA's ICR search tool](#) on all discontinued information collections between 01/21/2016 and 1/20/2026 in ten annual steps. The annual steps were mapped to *Administration Years* such that they initiate on the 21st of January and conclude on the 20th of January in the subsequent year. This ensures that discontinuation actions were requested by federal agencies correcting for Presidential inaugurations. The *Date Type* field was set to discontinued for each of the ten steps.

As Gretchen Gehrke points out in a 2026 editorial for The Hill ([Gehrke, 2026](#)), discontinuation represents one of the most commonly used tools by the Trump Administration in its efforts to reshape the federal data landscape. This is, in part, due to the connection between ICRs and deregulatory actions taken by agencies in fulfillment of both the 2-for-1 and 10-for-1 deregulation executive orders in the first and second Trump Administrations. Rescinding regulations reduces the number of information collections conducted by federal agencies. Data on information collection discontinuations is available through OIRA's [reginfo.gov](#) website. Between January 21st, 2025 and January 20th, 2026 this Administration had discontinued at least 562 information collections which is 65% more than the Biden Administration had discontinued in the same interval the year prior. In broad comparison to the 1923 collections discontinued in five Democratic Administration years (Obama 2016, Biden 2021-2024), the Trump Administration has discontinued nearly 2471, or about 30% more.

Revision

Of course, information collection discontinuation results in the complete loss of future data collections. However, the PRA process is also regularly used to modified continuing collections as well, especially by removing or changing survey items. In the last year alone, Executive Order 14168 (EO 14168) required agencies to make significant changes to information collections to gender identify resulting in changes to at least 360 individual surveys ([Bouton & Redfield, 2026](#); [Klein & Medina, 2026](#)). As discussed in the chapter reviewing data integrity issues at the [US Department of Veteran's Affairs](#) many of these changes involved response-option and variable name revisions and were sanctioned by OIRA despite the office's historical resistance to dramatic changes to survey collections ([Office of Management and Budget, 2025](#)). Such hasty, unscientific, collection revision should be considered an additional risk to data integrity.

While there are limited avenues for public comment to voice concerns over data discontinuation (or, relatedly, passive expiration of the OMB Control Number that authorizes a collection to continue), there are often ample opportunities for such comment during revision collection process. In response to the quiet disappearance of data and the opaque nature of government tracking systems like reginfo.gov and the Federal Register, independent monitoring by dataindex.us emerged to increase transparency and advocate for public input through the notice and comment process during information collection request revisions. The dataindex.us platform actively tracks ICRs and aggregates proposed changes to federal surveys and forms to help data users stay informed. By simplifying the complex information published by OIRA, the site makes otherwise invisible administrative shifts visible to policymakers, journalists, and researchers. Crucially, dataindex.us plays a vital role in amplifying the use of public comment by highlighting specific opportunities for stakeholders to weigh in on data collections, organizing these opportunities. Through [weekly newsletters](https://dataindex.us/newsletter), periodic blog posts, and detailed ICR tracking, the organization mobilizes the public to submit comments in an effort to support good governance and democratic principles in how the government collects data ([Maury & Marcum, 2025](#)). Of course, public comment is only effective when the opportunity exists - in the case of EO 14168, OIRA circumvented the notice and comment process by authorizing agencies to invoke “non-substantive” changes justifications which do not require the full PRA review process.

Conclusion

In general, agencies are not allowed to delete datasets generated through an information collection that has been discontinued or revised. Discontinuation ceases collection only and does not de-obligate federal agencies (or OIRA) from following law and policies on data retention, protection, and sharing: agencies are still obligated to assess the data assets produced by those collections, inventory them in their comprehensive data inventories, make those datasets meeting certain requirements publicly accessible, and follow the Federal Records Act disposition processes. While public notice and comment is typically available for requests for revision to continuing collections subject to the PRA, the same transparency is only partly available for discontinuations. For all datasets that rise to the level of being “significant information”, however, agencies are supposed to follow the guidance in [OMB Circular A-130](#) when making changes to collections resulting in high-value data assets (even if they are not subject to the PRA). This circular governs federal information management and requires agencies to provide “adequate public notice” before substantially modifying or terminating “significant information products” ([Office of Management and Budget, 2016](#)). The policy is designed to ensure stakeholders, including other federal agencies, who rely on the data are informed and can provide input prior to its removal. When agencies let the OMB control numbers for these significant products expire passively or discontinue them on reginfo.gov without broad announcements, they bypass the transparency safeguards outlined in Circular A-130. As a result, valuable government data collections can disappear quietly, directly undermining the continuity and integrity of the federal data ecosystem.

References

1. Bouton, L. J. A., & Redfield, E. (2026). *Removal of Sexual Orientation and Gender Identity from Federal Data Collections: January 2025 to January 2026*. Williams Institute. <https://williamsinstitute.law.ucla.edu/publications/sogi-data-collection-removal/>
2. Gehrke, G. (2026). The Trump administration is disappearing climate change data. *The Hill*. <https://thehill.com/opinion/energy-environment/5756951-federal-data-threatened-trump-era/>
3. Klein, M., & Medina, C. (2026). One Year In: The Cost of Rolling Back Federal LGBTQ Data. *America's Data Index*. <https://dataindex.us/newsletter/article/f69108a5-7346-4257-9fbc-e2ef784bd96b>
4. Maury, M., & Marcum, C. (2025). How You Can (and should) Shape Federal Data Collections. *America's Data Index*. <https://dataindex.us/newsletter/article/6cfecae3-3c89-487e-b8f7-cfcd85ded6c7>
5. Office of Management and Budget. (2025). *Guidance on Implementing Section 3(e) of Executive Order 14168 in Accordance with the Paperwork Reduction Act and the Privacy Act*. Office of Information and Regulatory Affairs.
6. Office of Management and Budget. (2016). *Circular No. A-130: Managing Information as a Strategic Resource*. Executive Office of the President. https://www.whitehouse.gov/wp-content/uploads/legacy_drupal_files/omb/circulars/A130/a130revised.pdf

Resourcing and Staffing

Actions taken by the Administration to reduce the size of the federal workforce and restructure data-generating programs in 2025 posed some of the most significant risks to the integrity of federal data in 2025. These reductions, initiated through hiring freezes, deferred retirement and other separation incentives, budget cuts, and program restructuring, led to a substantial loss of institutional capacity to sustain prior levels of data collection, curation, and dissemination. In the federal data ecosystem, resourcing and staffing risks were significant both inside and outside of the federal statistical system ([Bowen et al., 2025](#); [Gehrke, 2026](#); [Smith, 2026](#)). This chapter provides a broad overview of both theoretical and realized risks to federal data from administrative changes to staffing and program support. More comprehensive reporting is available in the references.

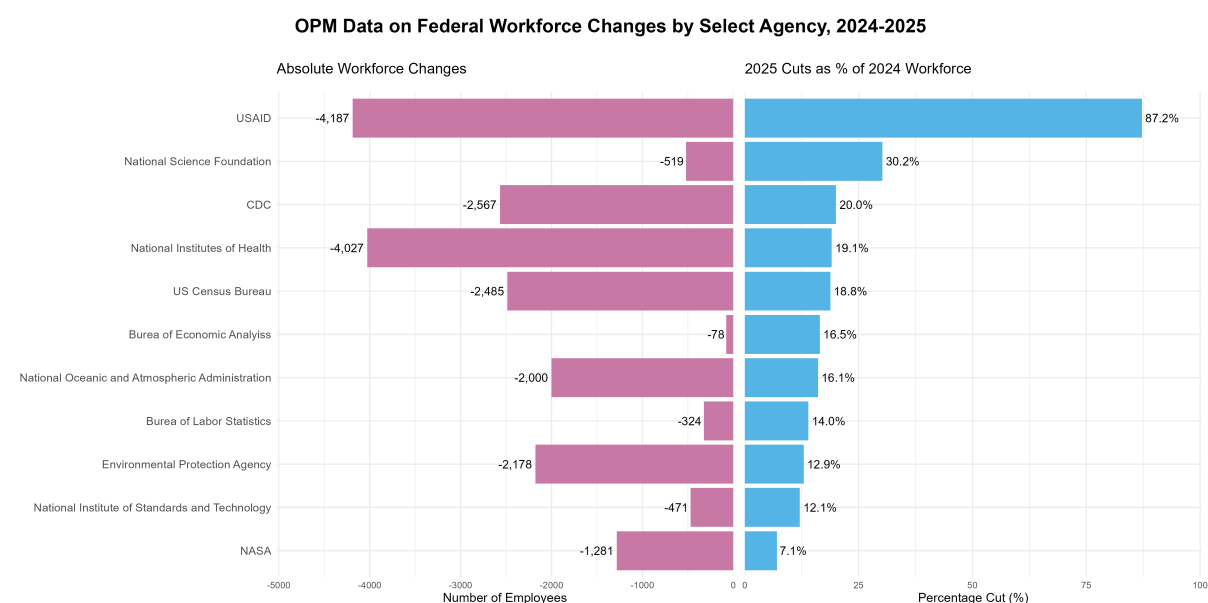


Figure: Data from the [OPM EHRI](#) 12/31/2025 release was used to illustrate workforce changes in select federal agencies between 12/31/2024 and 12/31/2025. Agencies were selected for their role in generating a significant volume of open government data. Both the absolute change in the number of agency employees and the turnover as a percent of the 2024 workforce size are represented in the graph. The data differ slightly from that reported by the American Statistical Association (ASA) in the *Nation's Data At Risk* because these figures represent adjustments made by OPM since the ASA report was published.

Impact on the Federal Statistical System

In 2025, the federal statistical system experienced widespread workforce contractions and resourcing constraints, a pattern of under-resourcing that has plagued the system for years ([Auerbach et al., 2024](#)). The American Statistical Association (ASA) reported that all 13 principal statistical agencies lost staff during this period and data-generating programs were cut as well ([Bowen et al., 2025](#)). Most statistical agencies lost between 20 percent and 30 percent of their professional staff. Statistical products were discontinued or removed from public access because data collection pipelines ended when contracts were cut, funding lapsed, or technical staff were unavailable to process the data ([Heckman, 2025](#)). For example, the Energy Information Administration, the Department of Energy's statistical agency, lost approximately 40% of its staff—over 100 out of approximately 350 employees. As a result, the EIA canceled the publication of its International Energy Outlook for 2025, a 70-page statistical report detailing global trends, specifically citing the “loss of key resources” and staff ([Dayak & Kramer, 2026](#); [Elkind, 2025](#)). The disruption at EIA led agency leadership to request “all hands-on-deck” to preserve institutional knowledge for future capacity before the mass departure ([LaRose, 2025](#)).

Perhaps hardest hit of all federal statistical agencies, The Bureau of Labor Statistics (BLS) faced compounding pressures from political turmoil, staff reductions, budget cuts, and pressures to modernize in the context of long declining survey response rates ([Auerbach et al., 2024](#)). A 43-day government shutdown in late 2025 prevented the execution of employment and inflation surveys for the first time in nearly 80 years (missing reference). This historic shutdown produced an irrecoverable gap in the modern unemployment rate series dating back to 1948 and halted the publication of the Consumer Price Index (CPI) in its modern form. The BLS also experienced severe headcount declines. By the time Commissioner Erika McEntarfer was fired by President Trump on August 1, 2025, BLS had lost up to 22 percent of its staff and left numerous leadership roles unfilled ([American Statistical Association, 2025](#)); though, some positions have apparently since been restaffed according to updated OPM data. The agency also suspended CPI data collection entirely in Lincoln, Nebraska; Provo, Utah; and Buffalo, New York. It reduced the sample size in 72 other areas by 15 percent ([U.S. Bureau of Labor Statistics, 2025](#)). Furthermore, the BLS discontinued the calculation and publication of approximately 350 specific Producer Price Index (PPI) series ([U.S. Bureau of Labor Statistics, 2025](#)).

Effect on Confidential Statistical Data

While public attention heavily focused on the impact these disruptions had on open government data, a quieter and equally damaging disruption occurred within the secure data ecosystem. Federal statistical agencies house repositories of confidential statistical data (such as individual tax records, health behavior data, business establishment data, and student records) that cannot be published openly due to statistical confidentiality laws ([Levenstein & Kubale, 2025](#)). Confidential statistical data are protected by the Confidential Information Protection and Statistical Efficiency Act of 2018 (CIPSEA), which makes it a class-E felony for unlawful disclosure of data collected under a pledge of confidentiality. Those protections, however, are not automatic. They require oversight and implementation by highly skilled staff at the statistical agencies. When the staff who run those processes are lost - or the programs supporting those staff are defunded - it is increasingly likely that agencies will both limit release of open government data and limit access to confidential statistical data products. Neither outcome is desirable for a public statistical system that depends on both trust in confidentiality and public usability.

In another extreme example from the federal statistical system, The National Center for Education Statistics (NCES) staff was reduced to approximately three employees ([Bowen et al., 2025](#)) in 2025. This reduction severely limited the agency's ability to manage complex datasets, notably the Integrated Postsecondary Education Data System (IPEDS) and the National Assessment of Educational Progress (NAEP). Under-staffing was so bad at NCES, that the Office of Management and Budget (OMB) required the agency to remove CIPSEA protections from the NAEP as the agency could not meet its obligations under the law to protect those data ([U.S. Department of Education, 2025](#)), resulting in considerable [push-back from stakeholders](#).

Usually, researchers and other data users gain access to confidential statistical data through secure enclaves by submitting project proposals that are rigorously reviewed by federal staff. The Federal Statistical Research Data Center (FSRDC) program serves as the primary secure environment for accessing restricted federal data requested through the Standard Application Process (SAP). Under the Foundations for Evidence-Based Policy Making Act of 2019 (Evidence Act), federal statistical agencies are charged with simultaneously protecting confidentiality while simultaneously increasing public access to statistical data. The FSRDC and SAP programs work hand-in-hand to accomplish that dual charge. However, the administrative disruptions to statistical agencies described here significantly jeopardized the integrity of this system.

For example, NCES has been unable to process access requests to their confidential statistical data since at least May of 2025. Even before the full impact of the agency's upset would be fully appreciated, the Bureau of Labor Statistics (BLS) abruptly stopped participating in the FSRDC program very early in 2025. This decision immediately cut off academic and institutional researchers from analyzing restricted establishment and labor microdata necessary for complex economic modeling.

Consequently, agencies stopped reviewing new data access requests through the SAP and paused disclosure review of existing projects ongoing in the FSRDC program had to be halted as agency staff were not available to provide disclosure review of manuscripts and aggregated data extracts intended for publication. The paralysis of the confidential statistical data ecosystem was formalized through notices posted to the [SAP portal](#), the centralized federal system for requesting access to confidential data. The SAP portal and program staff (housed at the National Center for Science and Engineering Statistics (NCSES) at the National Science Foundation) explicitly notified researchers of massive programmatic shutdowns throughout 2025 per OMB guidance. As of the publication of this report, the portal still indicated the following notices:

"Note: the Internal Revenue Service Statistics of Income did not have a 2025 Joint Statistical Research Program application window and is not accepting new proposals, with a re-evaluation of the program's status to occur at a date to be determined. This action does not impact requests to access commingled Federal Tax Information through the Federal Statistical Research Data Centers."

"Note: the National Center for Education Statistics is not accepting new proposals or reviewing proposals."

"Note: the SAMHSA Center for Behavioral Health Statistics and Quality is accepting applications for the National Survey on Drug Use and Health."

Administrative disruptions to restricted data assets outside of the federal statistical system were also prevalent in 2025 and continue into 2026. The CDC's popular *Pregnancy Risk Assessment Monitoring System* (PRAMS), for example, was widely reported to be at risk of discontinuation ([Hamad et al., 2026](#)). While the CDC [PRAMS website](#) providing access to reports using its data was restored as a result of a lawsuit ([Alder, 2025](#)), review of data access requests is still on pause as indicated by this disclosure on its website:

"PRAMS ARF data requests are not currently being processed. Researchers wanting to analyze data can contact each site separately to request access to their data. Please email the point of contact or visit the website for each of these sites for more information. (Participating PRAMS Sites)"

Impact on Other Agencies

A similar pattern of staffing and resourcing disruption to federal data also appeared outside the federal statistical system in 2025. Program disruptions across US federal agencies through separations, contract cancellations, and other administrative hazards have been widely reported elsewhere. However, it's useful to examine an illustrative example that is representative of system-wide effects on federal data. What happened at the National Oceanic and Atmospheric Administration (NOAA), for example, provides one such exemplar.

Reuters reported that NOAA lost about 1,000 employees, or roughly 10 percent of its workforce, in the early months of 2025 and that at least six National Weather Service offices had stopped routine twice-daily weather-balloon launches that feed weather models ([Reuters, 2025](#)). The total NOAA staff departures would rise to about 2,000 by the end of 2025, according to updated OPM data. Reuters also reported that the Administration's FY 2026 budget proposal would eliminate NOAA's Office of Oceanic and Atmospheric Research and proposed cutting funding for regional climate data, research labs, and cooperative institutes ([Reuters, 2025](#)).

Even though Congress has repudiated the Administration's budget request in the final appropriations bills for most data-generating programs ([Zimmermann, 2026](#)), the loss of institutional capacity resulting from staff departures can have sustained effects on the integrity of federal data. For example, former U.S. Chief Data Scientist Denise Ross warned that the loss of staff and contractors at NOAA could lead to the physical decay of sensors that collect much of the climate data used in forecasting, such as the ground-based radar systems that protect rural areas from tornadoes, because there is no one left to fix the equipment if it malfunctions ([Kim, 2025](#)).

Conclusion

The drastic and widespread reductions in federal workforce and resourcing throughout 2025 fundamentally undermined the integrity of both open and confidential data in the US Federal government. By hollowing out subject matter experts and other critical staff across agencies the Administration reduced data integrity capacity in a systemic manner. Ultimately, this systemic disruption created lasting deficits in the nation's ability to reliably collect, protect, and disseminate the vital data necessary for informed policymaking, economic forecasting, and scientific research. Notably, the long-standing under-resourcing of the federal statistical system that existed before the current Administration came into power set the conditions for the losses experienced in 2025. A more resource-rich federal statistical system would have inherent resilience to budget cuts, staff reductions, and other capacity challenges that affect the ability of the agency to protect the integrity of the data assets they collect from, and disseminate to, the public.

References

1. Alder, S. (2025). *HHS Settlement Requires Restoration of 100+ Health Datasets and Tools*. <https://www.hipaajournal.com/hhs-settlement-lawsuit-restore-critical-health-information-federal-websites/>
2. Auerbach, J., Bowen, C. M. K., Citro, C., Pierson, S., Potok, N., & Seeskin, Z. (2024). *The Nation's Data at Risk: Meeting America's Information Needs for the 21st Century*. American Statistical Association. <https://www.amstat.org/docs/default-source/amstat->

- [documents/the-nation's-data-at-risk—report.pdf](#)
3. Bowen, C. M. K., Citro, C., Crosby, M., Pierson, S., Potok, N., & Seeskin, Z. (2025). *The Nation's Data at Risk: 2025 Report*. The American Statistical Association. <https://www.amstat.org/policy-and-advocacy/the-nations-data-at-risk-2025-report>
 4. Dayak, S., & Kramer, A. (2026). Federal Data Is Disappearing . *NOTUS*. <https://www.notus.org/trump-white-house/federal-data-is-disappearing>
 5. Elkind, P. (2025). The Latest Trump and DOGE Casualty: Energy Data. *ProPublica*. <https://www.propublica.org/article/the-latest-trump-and-doge-casualty-energy-data>
 6. Gehrke, G. (2026). The Trump administration is disappearing climate change data. *The Hill*. <https://thehill.com/opinion/energy-environment/5756951-federal-data-threatened-trump-era/>
 7. Hamad, R., Warren, M., Ross, D., Maury, M., & Jarosz, B. (2026). *Rapid Response Data Briefing: Pregnancy Risk Assessment Monitoring System (PRAMS)*. Webinar hosted by dataindex.us, Association of Public Data Users, Population Reference Bureau, March of Dimes, and Population Association of America. <https://dataindex.us/events/Rapid-Response-Data-Briefing-Pregnancy-Risk-Assessment-Monitoring-System-PRAMS>
 8. Heckman, J. (2025). 'Bedrock' federal data sets are disappearing, as statistical agencies face upheaval. *Federal News Network*. <https://federalnewsnetwork.com/big-data/2025/12/bedrock-federal-data-sets-are-disappearing-as-statistical-agencies-face-upheaval>
 9. Kim, A. (2025). Federal Data Are Disappearing. *Washington Monthly*. <https://washingtonmonthly.com/2025/11/26/federal-data-are-disappearing/>
 10. LaRose, A. (2025). *Internal Memorandum* [Email sent to all staff]. U.S. Energy Information Administration, Office of Energy Analysis. <https://www.documentcloud.org/documents/25928821-larose-memo/>
 11. Levenstein, M., & Kubale, J. (2025). Data that taxpayers have paid for and rely on is disappearing – here's how it's happening and what you can do about it. *The Conversation*. <https://theconversation.com/data-that-taxpayers-have-paid-for-and-rely-on-is-disappearing-heres-how-its-happening-and-what-you-can-do-about-it-251787>
 12. Smith, M. (2026). America's Statistical System Is Breaking Down. *Bloomberg*. <https://www.bloomberg.com/news/articles/2026-01-09/why-the-trump-administration-is-choosing-not-to-collect-some-us-data>
 13. Zimmermann, A. (2026). *FY 2026 R&D Appropriations: Final R&D Report*. AAAS. https://www.aaas.org/sites/default/files/2026-02/Final%20Report%202026_0.pdf
 14. American Statistical Association. (2025). *Bureau of Labor Statistics*. <https://www.amstat.org/docs/default-source/amstat-documents/the-nations-data-at-risk-2025/bureau-of-labor-statistics.pdf>
 15. Reuters. (2025). *NOAA "fully staffed" with forecasters, scientists, US commerce secretary says*. <https://www.reuters.com/world/us/noaa-fully-staffed-with-forecasters-scientists-us-commerce-secretary-says-2025-06-04/>
 16. Reuters. (2025). *White House aims to eliminate NOAA climate research in budget plan*. <https://www.reuters.com/sustainability/climate-energy/white-house-proposes-eliminate-noaa-climate-research-budget-proposal-2025-04-11/>
 17. U.S. Bureau of Labor Statistics. (2025). *Notice of CPI collection reductions*. <https://www.bls.gov/cpi/notices/2025/collection-reduction.htm>
 18. U.S. Bureau of Labor Statistics. (2025). *BLS to Discontinue Selected PPIs*. <https://www.bls.gov/ppi/notices/2025/bls-to-discontinue-selected-ppis.htm>
 19. U.S. Department of Education. (2025). *Agency Information Collection Activities; Submission to the Office of Management and Budget for Review and Approval; Comment Request; National Assessment of Educational Progress (NAEP) 2026*. In *Federal Register*. <https://www.federalregister.gov/documents/2025/05/15/2025-08602/agency-information-collection-activities-submission-to-the-office-of-management-and-budget-for>
-

Data Tools

Datasets are the foundational collections of information that contain raw observations or measurements in structured or unstructured formats. These resources usually exist as static files in formats such as XLS, shapefiles, CSV, or JSON or are stored in relational databases such as SQL and often require external software or specific technical skills to be useful. Raw data often remain difficult to use for individuals who do not possess specialized training in data science or statistics. For example, a federal agency might maintain a dataset containing millions of individual records of disease reports that serves as the base evidence for long-term research.

Data tools, on the other hand, provide accessible platforms for users to interact with and understand those underlying datasets. Data tools help translate complex data into formats that are more readily accessible to a wider audience. Examples of data tools include dashboards and mapping portals that provide a way to search or visualize information without the need for manual coding. Their value to the public is significant because they provide a way for general data users, including small business owners, students, and local officials to access federal evidence without needing to perform their own technical analysis. This accessibility empowers the public to make informed decisions and supports transparency in government operations.

Disruptions to Data Tools

Unsurprisingly, some of the most impactful disruptions to the federal data ecosystem occur when public access to a valued data tool is removed or their underlying data changes unexpectedly. In the last year, a coordinated shift in administrative priorities led to the removal or manipulation of several high-profile interactive data tools that previously served as cornerstones for research, public health, and national security analysis. Unlike the periodic archiving of individual datasets that occur on transparent and predictable timelines (either internally at an agency or with the National Archives), actions removing functional interfaces and analytical dashboards that improved public access to data were abrupt and disarming to many data users.

Climate and Environmental Justice Data Tools

One of the earliest and most significant removals occurred in January 2025 with the decommissioning of Council on Environmental Quality's Climate and Economic Justice Screening Tool (CEJST) and the Environmental Protection Agency's EJScreen ([CEQ, 2025; Archive, 2025](#)). These tools provided geospatial visualizations of environmental burdens and socioeconomic indicators contributing to inequality. While the raw data behind these projects remained technically accessible via federal servers, the integrated mapping interfaces that allowed local governments and community organizers to identify disadvantaged areas were taken offline.

Similarly, the National Climate Assessment (NCA) hub was temporarily removed from federal servers on June 30, 2025, when the GlobalChange.gov domain was deactivated following the elimination of funding for the U.S. Global Change Research Program (USGCRP). The NCA Atlas tool allowed users to explore temperature and precipitation variables across different global warming levels using an interactive mapping interface provided through a contract with ESRI. The site was later restored after a reorganization of websites by the National Oceanic and Atmospheric Administration and is [currently publicly accessible](#), although the data are likely to have been changed given the stated intent to do so by the Administration ([Frazin, 2025](#)). The original 5th NCA hub along with the NCA Atlas are also currently available through independent mirrors such as [nca5.climate.us](#) and a partial continuation of the original site provided by ESRI [through ArcGIS](#).

Because of the high-value of the information that CEJST, EJScreen, and the NCA Hub provided to the public, their removal prompted a rapid-response in the data rescue community. Independent organizations and civil society groups launched private restoration projects to maintain public access to these tools ([Willson, 2025](#)). Organizations and projects such as the Public Environmental Data Project (PEDP) have successfully restored versions of tools like the [EJScreen](#) and the data underlying the NCA hubs. PEDP has restored a number of related tools, which can be found on <https://screening-tools.com/>. These efforts rely on rapid preservation and archiving of the original federal code and data.

Geospatial Planning Data Tools

In late August 2025, the Department of Homeland Security discontinued the public-facing Homeland Infrastructure Foundation-Level Data (HIFLD) Open portal. This tool was a primary resource for emergency managers and urban planners, offering an interactive environment to visualize critical infrastructure layers like energy grids and medical facilities. The removal restricted this high-fidelity mapping capability to vetted government users through a secure portal, effectively ending public access to a centralized geospatial utility.

Leadership at the contractor company that hosted the data tool, ESRI, was alerted by concerned federal staff familiar with the plans to shutter HIFLD Open through contract cancellation prior to the takedown. Initially, ESRI agreed to provide continued free public access to the resource and vowed to backup the data. However, since December 2025 the ESRI site via an [ArcGIS endpoint](#) was no longer available. Connected endpoints to shared resources at other agencies, such as the [geospatial layers provided by NASA](#) to the tool, have also been discontinued. Thanks to efforts of the [Data Rescue Project](#), in partnership with ICPSR, all of the underlying data and metadata have been archived via [datalumos.org](#).

HHS' Large-scale Data Tool Takedown

Data tools were also wrapped up in the wholesale erasure of critical health information at HHS that led to the removal of approximately 8,000 web pages. Targeted tools provided insights on topics such as health equity, reproductive health, and LGBTQ+ health.

The Centers for Disease Control and Prevention's Social Vulnerability Index (SVI) dashboard was one of the most prominent tools eliminated during this period. This geospatial application allowed emergency managers to identify communities at risk during natural disasters by analyzing fifteen different census variables. Its removal in early 2025 forced local jurisdictions to rely on dated archives or alternative private indices to plan for disaster response. Similarly, the AtlasPlus and Youth Risk Behavioral Surveillance System dashboards and interactive tools were briefly taken offline.

These actions were challenged in federal court by a coalition of medical organizations, including the Washington State Medical Association ([Alder, 2025](#)). A legal settlement reached in September 2025 required the department to restore more than 100 of these removed tools and pages to their state as of January 29, 2025. While this mandated the return of the SVI and AtlasPlus, the restored versions often carry prominent headers indicating they are under review for future modification:

Per a court order, HHS is required to restore this website to its version as of 12:00 AM on January 29, 2025. Information on this page may be modified and/or removed in the future subject to the terms of the court's order and implemented consistent with applicable law. Any information on this page promoting gender ideology is extremely inaccurate and disconnected from truth. The Trump Administration rejects gender ideology due to the harms and divisiveness it causes. This page does not reflect reality and therefore the Administration and this Department reject it.

Despite the court-ordered restoration of the HHS resources, especially to CDC data tools, the disclaimers clearly indicate an intent to remove the tools once the order is lifted. To ensure long term access to the original analytical functions, civil society groups established independent mirrors at [RestoredCDC.org](#).

International Data Tools

The landscape for global data tools shifted significantly between early 2025 and 2026 as well. Primary federal resources for international comparative data were eliminated. These tools, which provided interactive interfaces for analyzing global health, geography, and political structures, served as essential utilities for researchers, diplomatic missions, and the general public.

The United States officially [rejected the United Nation's Sustainable Development Goals \(SDGs\)](#) on March 4th, 2025 in remarks by Edward Heartney, US Minister Counselor to the United Nation's Economic and Social Council. It's unclear whether Heartney had the authority to revoke US involvement in the SDGs as the responsibility for coordinating participation in international statistical activities is statutorily prescribed by the Paperwork Reduction Act to one of the roles of the US Chief Statistician in the Office of Management and Budget (OMB). The SDG website, a collaboration between OMB, the Department of State, and the General Services Administration (GSA), remained online until at least [May 1st, 2025](#), albeit with a disclaimer from GSA:

! This site is under review and content may change

The SDG reporting data had not been previously updated by OMB since October 11th, 2024. The eventual removal website in late Spring of 2025 marked a significant shift in how the United States reports its progress on international benchmarks. Previously, <https://sdg.data.gov> served as a centralized system tracked domestic indicators across the 17 global goals, including climate action, gender equality, and poverty reduction. While the raw data is available via the Internet Archive's [Wayback Machine] (https://web.archive.org/web/20250501022335/https://gsa.github.io/sdg-data-usa/en/zip/all_indicators.zip) the interactive data can still be found on the [UN Statistics Division website](#). Thankfully, in accordance with OMB policy, GSA retained public access to the open-source github repository where the [source code and data](#) are stored.

Ironically, one international data tool that was at significant risk but was not necessary to be recovered was the United States Agency for International Development (USAID) supported Demographic and Health (DHS) STATcompiler. The DHS Program STATcompiler allowed users to make custom tables based on hundreds of demographic and health indicators across more than 70 countries collected through USAID's long-standing Demographic and Health Survey program. This DHS dashboard was a primary tool for monitoring global population health by both federal agencies and the public. In early 2025, the interactive portal for these surveys was taken offline as part of a review of foreign assistance data. Throughout this period, the underlying survey data and documentation remained available to registered researchers through the program's primary contractor, ICF International. The entire [DHS program was preserved](#) via interim funding secured by the contractor.

Most recently, The Central Intelligence Agency officially sunset the interactive version of the World Factbook on February 6, 2026. For decades, this tool was the global standard for comprehensive, country-level profiles, providing searchable data on everything from population health to geopolitical conflicts to telecommunications infrastructure. The takedown of the World Factbook was sudden and absolute ([Marcum, 2026](#)) - all traffic to the original website domain is currently forwarded to [an anonymous blog post](#) and all official backups were removed from the website. The removal of the interactive portal initiated community-led restoration projects. These non-governmental initiatives, listed on [worldfactbook.us](#), collectively aim to preserve the static content, the searchable functionality, and historical continuity of the World Factbook. To date, there has not been a full restoration of the World Factbook similar to the restored functionality of some of the other data tools described in this chapter.

Data Tool Status Summary Table

The following table provides a summary of the status of these major tools as of February 2026. This is not a comprehensive list of all data tools removed, but rather focuses on major resources discussed in the news media as high-value. A more comprehensive, albeit mixed, list is available on [Wikipedia](#).

Data Tool	Original Agency	Current Functional Status	URL
AtlasPlus	CDC	Active; restored per court order	available at: https://gis.cdc.gov/grasp/ncshstpatlas/main.html
CEJST / EJScreen	White House / CEQ / EPA	Eliminated; available via non-gov mirrors	mirror at: https://pedp-ejscreen.azurewebsites.net/
DHS STATCompiler	USAID	Eliminated; available via non-gov mirrors	mirror at: https://www.statcompiler.com
Climate Risk Viewer	USDA	Active; restored per court order	available at: https://storymaps.arcgis.com/collections/87744e6b06c74e82916b9b11da218d28
NCA Hub	USGCRP	Active; modified per executive directive	partial mirror at: https://nca-atlas-nationalclimate.hub.arcgis.com/
Open HIFLD	DHS	Eliminated; available via non-gov mirror	mirror at: https://hifld.publicenvirodata.org/
sdg.data.gov	OMB	Eliminated; no functional replacement	data archived at: https://github.com/GSA/sdg-data-usa/tree/develop/data
SVI Interactive Map	CDC	Active; restored per court order	available at: https://www.atsdr.cdc.gov/place-health/php/svi/svi-interactive-map.html
World Factbook	CIA	Eliminated; no functional replacement	various restoration efforts listed at: https://worldfactbook.us

Conclusion

Users of federal data tools found many of their favorite dashboards and analytics platforms disrupted during 2025. Highly valued data tools have been largely restored by preservation initiatives. In part, this was possible because much of the underlying data were still publicly accessible (either at federal agencies or through non-governmental archives) and because the code supplying the tools was available open-source and deposited in the public domain (often, on GitHub). In some cases, government contractors who had originally been the stewards of the data tools on behalf of a government agency were able to maintain access to those tools.

Data tool restoration projects are also approachable from an external funding perspective because they have a tangible, time-bound, and discrete deliverable. Recreating specific canceled, deleted, or discontinued federal data collections that fuel these data tools, however, is more costly, more time consuming, and less easily executed due to transparency issues with underlying methodology.

References

1. Alder, S. (2025). *HHS Settlement Requires Restoration of 100+ Health Datasets and Tools*. <https://www.hipaajournal.com/hhs-settlement-lawsuit-restore-critical-health-information-federal-websites/>
2. Archive, N. S. (2025). *Disappearing Data: Trump Administration Removing Climate Information*. <https://nsarchive.gwu.edu/briefing-book/climate-change-transparency-project-foia/2025-02-06/disappearing-data-trump>
3. CEQ. (2025). *Climate and Economic Justice Screening Tool (CEJST)*. Harvard Dataverse. [10.7910/DVN/B6ULET](https://dataverse.harvard.edu/citation?persistentId=doi:10.7910/DVN/B6ULET)
4. Frazin, R. (2025). Energy Secretary Chris Wright says admin 'reviewing' past climate reports. *The Hill*. <https://thehill.com/policy/energy-environment/5441347-energy-department-chris-wright-national-climate-assessment-review/>
5. Marcum, C. S. (2026). The World Factbook was a Valuable Data Resource. In *Open Evidence*. <https://www.doi.org/10.59350/npaqr-r3682> <https://www.chrismarcum.com/marcum-blog/2026/02/06/The-World-Factbook-Was-A-Valuable-Data-Resource.html>
6. Willson, M. (2025). Groups archive environmental justice data scrapped by Trump. *E&E News*. <https://www.eenews.net/articles/groups-archive-environmental-justice-data-scrapped-by-trump/>

Agency Case Studies

While there were systemic, government-wide challenges to federal data integrity in 2025 that affected the ecosystem as a whole, there was also a lot of variation between Federal agencies. Federal agencies operate within vastly different statutory authorities, operational environments, and have different, agency-specific cultures around data governance. Those differences directly shape their exposure to data integrity risks and challenges.

Agencies also differ in the sensitivity of the information they steward, the scale and complexity of their holdings, and the governance cultures that shape how data is curated and protected. Some face intense public scrutiny and have mature open government and open data plans; others are less far along with their plans or operate with limited resources, fragmented systems, or competing operational priorities. These contextual factors influence exposure to risk, resilience to disruption, and the feasibility of resisting stressors on the integrity of their open government data. In this chapter, three cases of specific disruptions to the integrity of federal data between 2025 and 2026 at different agencies are profiled. These are: apportionments data takedown by the Office of Management & Budget; the collateral damage to data assets from the dismantling of the United States Agency for International Development; and reproducing research by Freilich & Kesselheim ([Data Manipulation within the US Federal Government, 2025](#)) on undocumented manipulation of metadata changes to datasets by the United States Department of Veteran Affairs.

The focus of this chapter is on profiles cases that occurred at non-statistical agencies; that is, agencies not designated as recognized statistical agencies and units (RSAUs) by the Office of Management and Budget. Comprehensive profiles of RSAUs, and the Federal Statistical System in general, are available in two reports by the American Statistical Association (see, Bowen et al. ([The Nation's Data at Risk: 2025 Report, 2025](#))).

References

1. Bowen, C. M. K., Citro, C., Crosby, M., Pierson, S., Potok, N., & Seeskin, Z. (2025). *The Nation's Data at Risk: 2025 Report*. The American Statistical Association. <https://www.amstat.org/policy-and-advocacy/the-nations-data-at-risk-2025-report>
2. Freilich, J., & Kesselheim, A. S. (2025). *Data manipulation within the US Federal Government*. [https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(25\)01249-8/fulltext](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(25)01249-8/fulltext)

Office of Management and Budget

On March 24, 2025, the OMB revoked public access to its [apportionments database](#). Apportionments are instructions issued by OMB that specify the timing and conditions under which federal agencies may use funds appropriated by Congress ([Fiorentino & Riccard, 2025](#)). Following the impoundment of more than \$250M in Congressionally mandated aid to Ukraine by OMB in 2019, Congress established a legal requirement for OMB to make all apportionment documents, including footnotes and written explanations, available on a public website within two business days of their issuance in the The Consolidated Appropriations Act of 2022 ([Heilweil, 2025](#)). This legislation was designed to provide Congress and the public with real-time insight into how the executive branch manages appropriated funds. Previously, OMB had voluntarily made apportionments data publicly accessible. A Protect Democracy initiative provides both historical and, when it is available, contemporaneous data available through [OpenOMB](#).

Based on [data from the OpenOMB API](#), OMB prevented access to more than 1700 individual apportionment files for FY 2024 and 2025 between March 24th and August 15th of 2025 (code is [also available in the repository](#)). The actions taken by OMB, prompted inquiries from civic-society organizations and the Government Accountability Office (GAO). Public access to apportionments data is the primary transparency tool used for accountability that OMB is appropriately spending funds appropriated by Congress. The GAO subsequently issued a finding that the removal of the website was inconsistent with the law. Loss of public access to the apportionments data represented one of the most significant examples of direct political interference into the integrity of federal data in 2025.

OMB Director Russell Vought defended the decision to restrict access to the database, outlining his reasoning in communications with the House Appropriations Committee and in subsequent court filings. Vought argued that the 2022 transparency requirements interfered with the executive branch's internal operations. He maintained that apportionments often contain pre-decisional and deliberative information that should be protected from immediate public release. According to Vought, making these documents public in real-time could inhibit candid discussions between OMB staff and agency officials regarding budgetary adjustments.

Furthermore, the OMB argued that the President possesses inherent Article II authority to manage the execution of the budget. From this perspective, the mandatory public posting of technical spending footnotes was viewed by the OMB as an encroachment on executive discretion. Vought also noted concerns that some spending data might inadvertently reveal sensitive administrative strategies, asserting that the OMB required a level of confidentiality to execute the President's policy priorities efficiently without premature public or legislative interference ([Emma, 2025; Katz, 2025](#)).

The removal of the data led to lawsuits from several organizations, including Citizens for Responsibility and Ethics in Washington (CREW). The plaintiffs argued that the OMB was in direct violation of the 2022 Act. In July 2025, U.S. District Judge Emmet Sullivan ruled that the OMB was legally required to maintain the database. The court found that because apportionments are legally binding directives under the Antideficiency Act, they constitute final agency actions rather than protected deliberative materials. Consequently, the court issued a permanent injunction requiring the OMB to restore the website ([Emma, 2025](#)).

Following this ruling, the OMB restored the database in August 2025. However, further disputes arose regarding the completeness of the data. Plaintiffs noted that the OMB was using "A" footnotes to refer to "spend plans" and other documents that remained private. This led to additional litigation regarding whether the statutory requirement for "all footnotes and written explanations" included documents incorporated by reference ([Citizens for Responsibility and Ethics in Washington, 2025; Katz, 2026](#)).

Timeline of Events

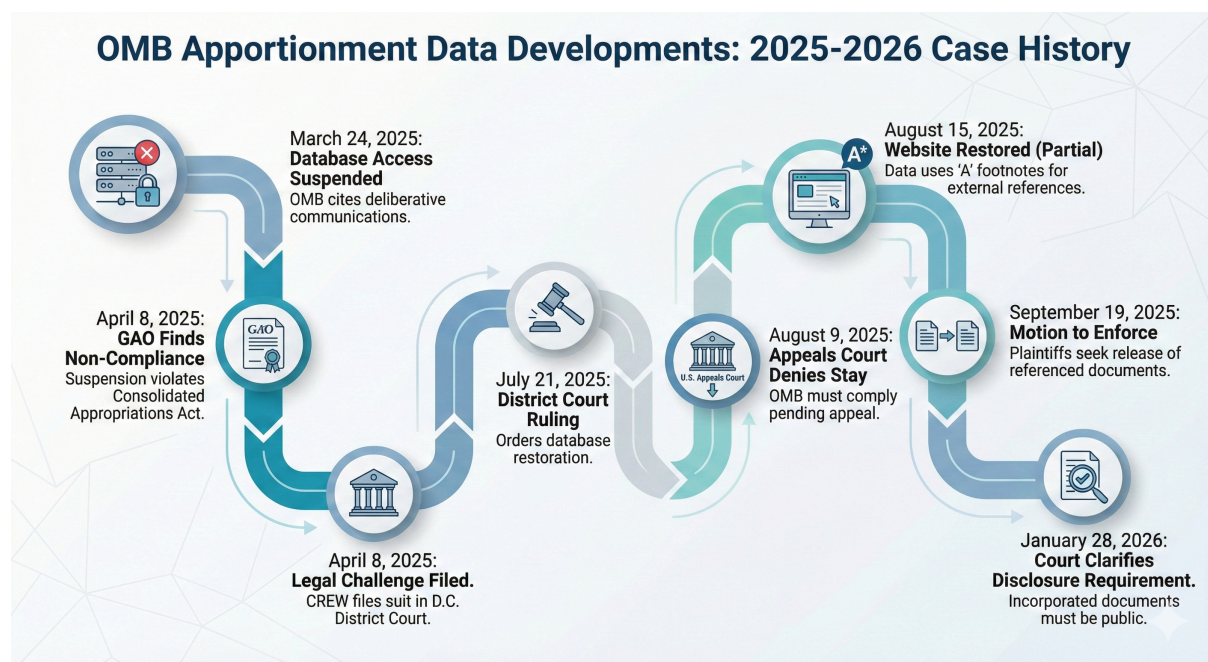


Figure: Apportionments dataset public access removal and restoration timeline, 2025-2026. The figure was generated with Google Slides AI 'beautify this slide' feature on a plain slide with the bulleted timeline text (below).

- **March 24, 2025:** OMB disables the Public Apportionments Database, citing the need to protect deliberative communications.
- **April 8, 2025:** The GAO confirms that the suspension of the website violates the Consolidated Appropriations Act.
- **April 8, 2025:** Legal challenges by Citizens for Responsibility and Ethics in Washington are filed in the U.S. District Court for the District of Columbia.
- **July 21, 2025:** The District Court rules against OMB and orders the restoration of the database.
- **August 9, 2025:** A federal appeals court declines to stay the lower court's order, requiring the OMB to comply during the appeals process.
- **August 15, 2025:** OMB restores the public website, though some data remains inaccessible through the use of external references known as 'A' footnotes.
- **September 19, 2025:** Plaintiffs file a motion to enforce the injunction, seeking the release of documents referenced in footnotes.
- **January 28, 2026:** The court clarifies that documents incorporated by reference in apportionments must also be made public.

Conclusion

OMB's 2025 removal of the public apportionments database represents a significant episode of political interference that eroded the integrity of federal data in 2025. Specifically, the loss of continuity of public access to these data for several months before being restored by court-order pitted unprecedented assertions of Executive power over deliberative information against congressional mandates for data access. Although legal challenges and judicial rulings successfully forced the reinstatement of the database, the OMB's subsequent use of external references to obscure complete spending details underscores an ongoing under-current of a desire to erode and obscure public access to these critical assets. Ultimately, this episode illustrates the persistent vulnerability of federal data integrity to political interference and the critical role of judicial oversight and civic action in enforcing public access to government records. Apportionments data likely represent the most critical data assets related to government transparency under threat by the present Administration and, despite restoration, will remain under threat as OMB appeals the decision.

References

1. Emma, C. (2025). Appeals court rules Trump clamp-down on spending data defies Congress' authority. *Politico*. <https://www.politico.com/news/2025/08/09/appeals-court-rules-trump-clamp-down-on-spending-data-defies-congress-authority-00501348>
2. Fiorentino, D. A., & Riccard, T. N. (2025). *Office of Management and Budget (OMB) Reporting on Apportionments* (CRS Insight No. IN12538; Issue IN12538). Congressional Research Service. <https://crsreports.congress.gov/product/pdf/IN/IN12538>
3. Heilweil, R. (2025). GAO says OMB takedown of apportionments website violates federal statutes. *FedScoop*. <https://fedscoop.com/gao-omb-takedown-apportionments-website-federal-statutes/>
4. Katz, E. (2026). Court orders OMB to publish more info about how federal funding is distributed. *Government Executive*. <https://www.govexec.com/management/2026/01/court-orders-omb-publish-more-info-about-how-federal-funding-distributed/411092/>
5. Katz, E. (2025). Spending transparency data posted by Trump budget office after court order. *Government Executive*. <https://www.govexec.com/management/2025/08/spending-transparency-data-posted-trump-budget-office-after-court-order/407548/>
6. Citizens for Responsibility and Ethics in Washington. (2025). *OMB's latest effort to conceal spending data*. CREW Investigations. <https://www.citizensforethics.org/reports-investigations/crew-investigations/ombs-latest-effort-to-conceal-spending-data/>

United States Agency for International Development

The shuttering of the United States Agency for International Development (USAID), a process that began on January 20th, 2025 with the issuance of [Executive Order 14169](#), resulted in a significant disruption to the public accessibility of a large corpus of federal data. While practically unreported in the news media relative to other data disruptions, the shuttering of USAID likely had an out-sized effect on overall data-loss during 2025. USAID housed significant data assets in its Development Data Library (DDL). The DDL primarily hosted structured, machine-readable datasets generated through USAID-funded programs, ranging from longitudinal health and education surveys, to local economic conditions in developing countries, to geospatial maps and images. The DDL also served as USAID's Research and Development (RAD) grant program's public access repository for research data funded by the agency.

Director of the Duke Center for International Development, Dr. Edmund Malesky, wrote about the value of the DDL and the RAD data it held in a letter attached to the center's [annual report](#) that he penned:

"Moreover, we were obligated to post all of our analysis and data on a publicly available website, so anyone could check our work. I know very few organizations that insist on this level of transparency. In fact, one tragic feature of the Trump administration's closure was the destruction of this very repository, sending thousands of USAID employees and contractors scrambling to recover and preserve their hard work." ([Malesky, 2025](#)).

The DDL housed 2,027 data assets and remained online until at least [January 28th, 2025](#) and according to an analysis of data supporting this project, around 81% of those were publicly accessible. While public access to the DDL site went dark on or about January 31st, 2025 USAID's comprehensive data inventory (previously available at <https://data.usaid.gov/api/views> and <https://data.usaid.gov/data.json>) remained listed as a harvest source for Data.gov until at least [August 19th, 2025](#). Despite this, no data could be downloaded from hyperlinks to the source, leading to a significant number of ghost datasets in the Federal Data Catalog up until that time. The DDL was among the data resources memorialized by [essentialdata.us](#) and the Federation of American Scientists for their Halloween 2025 "[Dearly Departed Datasets](#)" initiative.

Because the DDL's Socrata API did not provide direct download links for many of its resources to the Data.gov harvester, data rescue efforts that relied on the Federal Data Catalog had mixed success. For instance, the Harvard Law Library's archive of Data.gov contains [only metadata](#) for the formerly publicly accessible [Demographic and Health Surveys \(DHS\)](#). The last Wayback Machine snapshot of the Federal Data Catalog entry for the DHS was on [August 19th, 2025](#). All of the underlying data can [still be accessed](#) through the website of the contractor that has run the program for more than 40 years, thanks to funding they secured to shore-up the project for at least the next three years. A small number of other USAID data assets were preserved and archived by the Inter-university Consortium for Political and Social Research (ICPSR) via [datalumos.org](#). Fortunately, much of the remaining publicly accessible data held by USAID [was preserved](#) by the Library Innovation Lab at Harvard Law School's efforts to archive all of the assets linked from Data.gov between late 2024 and mid-2025 ([Satter, 2025](#)).

The future of USAID data

Conversations with individuals familiar with the data infrastructure at the Department of State, USAID, and the General Services Administration revealed that uncertainty remains around the future of access to the agency's datasets. For instance, while the Administration garnered considerable attention for the mishandling of classified documents during its haphazard dismantling of USAID ([Fischer, 2025](#)), the status of non-classified records and data is not currently known to the public. One person familiar with the matter suggested that the process of formal archiving of USAID data with the National Archives was underway at State. As of the date of publication of this report, neither State nor USAID has made a request for a change in the records schedule with the National Archives and the last records schedule request for the agency [occurred in 2023](#).

However, data restoration is very likely to involve an assessment of whether the release of specific USAID data assets is consistent with Administration priorities. For instance, the USAID-administered [President's Emergency Plan for AIDS Relief \(PEPFAR\)](#) website had previously [listed twenty program evaluation datasets](#) in five programs available for download; the Administration removed 13 of these datasets as inconsistent with Administration priorities. Among those datasets removed (and still available via the Internet Archive's WaybackMachine) were data on [voluntary circumcisions](#) (an HIV prophylactic procedure) and on [cervical cancer surveillance](#). Currently, [only two evaluation program summaries remain accessible](#) with a disclosure posted to the website indicating further data and program review is possible:

"A subset of historic Spotlight datasets through Fiscal Year 2024, compliant with Executive Orders, are available below. Additional datasets and FY25 results will be made available with the next Spotlight update. Note: Fiscal Year 2025 targets and budgets as approved through congressional notification procedures for FY25 are included in these datasets, however program interruptions, reporting challenges, and programmatic shifts have occurred since January 2025. Fiscal Year 2025 target and budget data should be analyzed and interpreted accordingly."

Conclusion

The sudden dismantling of USAID in early 2025 precipitated a massive, albeit underreported, disruption to global development and public health data access. While swift interventions by civic society groups and at least one USAID contractor managed to rescue critical assets like the Demographic and Health Surveys and portions of the Development Data Library, the broader data infrastructure remains severely compromised. The ongoing uncertainty surrounding records preservation (including final disposition in the National Archives), compounded by the politicized purging of specific datasets to align with new administration priorities, highlights the profound vulnerability of federally funded research to abrupt administrative shifts, policy reprioritization, and ideological censorship.

References

1. Fischer, W. (2025). *Letter to Mr. Christopher Colbrow, Agency Records Officer, US Agency for International Development*. <https://www.archives.gov/files/records-mgmt/resources/ud-2025-0047-usaid-open.pdf>
 2. Malesky, E. (2025). *2025 Letter from the Director - Duke Center for International Development*. <https://dcid.sanford.duke.edu/2025-letter-director/>
 3. Satter, R. (2025). Harvard Law Library acts to preserve government data amid sweeping purges. *Reuters*. <https://www.reuters.com/world/us/harvard-law-library-acts-preserve-government-data-amid-sweeping-purges-2025-02-06/>
-

U.S. Department of Veterans Affairs

Some agencies were overzealous in their compliance with [Executive Order 14168](#) by replacing the terms “gender” with “sex” in the titles, descriptions, and field labels of their datasets, and may have also considered several dataset titles to be inconsistent with [Executive Order 14151](#). For instance, the U.S. Department of Veterans Affairs made substantial changes to its comprehensive data inventory (CDI). The metadata for at least fifty-one datasets was altered between February 15th, 2025 and December 31st, 2026. Most of these alterations involved titles invoking gender, race, or both, and occurred between March 15th, 2025 and April 2nd, 2025. Some metadata changes were entirely innocuous and likely reflected mundane maintenance of the files. All entries with changes to titles between the two time points are listed below.

- VetPop2020_Gender_2000to2023
- Board of Veterans’ Appeals
- Use of VA Benefits and Services: 2021 (Part 2)
- VetPop2020 Urban/Rural by Period of Service FY2023
- Use of VA Benefits and Services: 2021 (Part 1)
- Number of Users by Program, FY2010-2021
- Trend in Use of Any VA Benefit, FY 2010-2021
- VETPOP2014 LIVING VETERANS BY RACE/ETHNICITY, GENDER, 2013-2043
- AIAN Veterans Report (2015)
- FY 2021 Total Number of Veterans, Veteran VA Users, and Veteran VA Healthcare Users by Gender and Age Group
- Veterans Utilization Profile FY18 - Fig. 9 - Use Rate of Genders within Era
- Korean War Veterans by State
- Rates of Use within Age Group by Sex, FY2021
- Rate of Use by Race/Ethnicity, FY2021
- Percentage Era Distribution of Female Users and Non-Users, FY 2021
- VetPop2020 Urban/Rural by Ethnicity FY2021-2023
- Use of VA Benefits and Services: 2021 (Introduction)
- Trend in Percent Health Care Enrolled Users, Enrolled Non-Users & Non-Enrolled among Service-Connected Disabled Veterans, FY2010-2021
- Profile of Veterans: (2017)
- Percentage Age Distribution of Male Users and Non-Users, FY 2021
- Users of VA Benefits by Program, FY2021
- VetPop2020 State Estimates 2000 to 2020
- Use of VA Benefits and Services: 2021 (Appendix)
- Percentage Age Distribution of Female Users and Non-Users, FY 2021
- Trend in Rate of Users by Sex, FY2010-2021
- VetPop2020 National Estimates by Race 2000 to 2020
- Analysis of Differences in Disability Compensation in the Department of Veterans Affairs
- Percentage of Service-Connected Disabled Veterans Who Used VA Health Care, by Race/Ethnicity, FY 2021
- Rural Veterans: FY2021-2023
- Percentage Distribution of Users by Era of Initial Service and Sex, FY 2021
- Rates of Use by Sex within Era of Initial Service, FY 2021
- Number of Users of One or More VA Programs by Sex, FY2010-2021
- Age Distribution of Users by Sex, FY2021
- VetPop2020_GenderRaceEthnicity
- Trend in Percent of Health Care & Disability Compensation Users vs Other Users, FY2010-2021
- VetPop Total Population by Gender
- vetpop_gender
- VetPop2020 Urban/Rural by Race FY2021-2023
- VetPop2020 National Estimates 2000 to 2020
- VetPop2020 Urban/Rural by Poverty & Disability FY2021-2023
- Percent of Veterans who Use VA Benefits by Program and Gender, FY2023
- Percentage of Service-Connected Disabled Who Did and Did Not Use Health Care, by Disability Rating, FY2021
- Percentage Era Distribution of Male Users and Non-Users, FY 2021
- Use of VA Benefits and Services: 2021 (Part 3)
- VetPop2020 Urban/Rural FY2021-2023
- FY 2020 Total Number of Veterans, Veteran VA Users, and Veteran VA Healthcare Users by Gender and Age Group
- FY10 Compensation and Pension by County
- VetPop2020_GenderPeriodOfService
- Take-up Rate by Race/Ethnicity and Gender
- Veterans Receiving Compensation Service Benefits On the Rolls by Period of Service and Residence FY22 and FY23
- Percent Change in Veteran Population by State from 2000 to 2022

Critically, many of these changes were not simply modifications to persistent records. Rather, VA apparently removed and replaced many of the updated the entries altogether. For example, the previous CDI entry for the dataset titled, “VetPop2020 estimate of Veterans by gender from 2000 to 2023”, had this entry:

```
{
  ...
  "identifier": "https://www.data.va.gov/api/views/2rci-xm64",
  "description": "VetPop2020 estimate of Veterans by gender from 2000 to 2023",
  "title": "VetPop2020_Gender_2000to2023",
  "programCode": [
    "029:086"
  ],
  "distribution": [
    {
      "@type": "dcat:Distribution",
      "downloadURL": "https://www.data.va.gov/api/views/2rci-xm64/rows.csv?accessType=DOWNLOAD",
      "mediaType": "text/csv"
    },
    ...
  ]
}
```

which appears to have been entirely deleted and replaced with a new entry:

```
...
"identifier": "https://www.data.va.gov/api/views/y4w4-egzx",
"description": "The Department of Veterans Affairs provides official estimates and projections of the Veteran population",
"title": "VetPop2023 National Estimates by Sex and Age Groups 2000 to 2023",
"programCode": [
  "029:086"
],
"distribution": [
  {
    "@type": "dcat:Distribution",
    "downloadURL": "https://www.data.va.gov/api/views/y4w4-egzx/rows.csv?accessType=DOWNLOAD",
    "mediaType": "text/csv"
  },
  ...
]
```

In part, this is likely due to the how the underlying data management system VA uses, Socrata, automatically generates API endpoints for every dataset uploaded to its platform through its Socrata Open Data API (SODA).

Manipulation of underlying data files

According to research by Freilich & Kesselheim ([Data Manipulation within the US Federal Government, 2025](#)), nearly half of the 79 datasets from the VA they examined were manipulated in response to EO 14168 between the Inauguration and March of 2025. Freilich graciously provided the python code used for their report and the acquisition routines and review methods were reproducible (and indeed comparable to those used in this project and complementary to those in [the auditing workflow](#)). They manually evaluated both metadata and underlying dataset changes by comparing live versions of VA datasets to those archived prior to January 20th, 2025 by the WayBack Machine. Their results are consistent with those presented above: VA replaced “gender” with “sex” in the title, general description, and column headers of the plurality of the data and metadata they evaluated. None of the changes were documented by VA and there were no disclosures on the public dataset landing pages that this was occurring. Further auditing of several of the datasets Freilich & Kesselheim examined was done by evaluating the checksums of version-pairs after stripping the column headers out of the datasets were the same (via the SHA256 algorithm). In these cases, even though the variable names differed, the underlying data remained unchanged.

Even so, there is evidence that VA’s efforts were not comprehensive. Even up to the date of publication of this report, there are still metadata entries in the CDI that include the word “Gender” in both the title and in the underlying datasets (this likely belies a less-than-systematic approach on their part). For example, as of 3/12/2026, the dataset titled “FY 2021_NCVAS Vet Pop Gender Over Time Data For State Summaries” was still included in the CDI and available for download.

```
...
"identifier": "https://www.data.va.gov/api/views/gidu-8zyi",
"description": "These data are based on the latest Veteran Population Projection Model, VetPop2020, provided by the Natick",
"title": "FY 2021_NCVAS Vet Pop Gender Over Time Data For State Summaries",
"programCode": [
  "029:000"
],
"distribution": [
  {
    "@type": "dcat:Distribution",
    "downloadURL": "https://www.data.va.gov/api/views/gidu-8zyi/rows.csv?accessType=DOWNLOAD",
    "mediaType": "text/csv"
  },
  ...
]
```

Conclusion

Arbitrarily changing variable and study names is against best practices for data governance. It can also contribute to link rot and code replication problems in the future. However, those violations of data integrity are far less impactful than manipulating the underlying data. Forensic auditing that compared pre- and post-Administration versions of several of these datasets revealed that data manipulation was unlikely. The materials reviewed here support the conclusion that the VA’s 2025 public-data integrity issues were primarily metadata and labeling changes on public dataset pages, often without clear documentation, rather than changes to the underlying data.

References

1. Freilich, J., & Kesselheim, A. S. (2025). *Data manipulation within the US Federal Government*.
[https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(25\)01249-8/fulltext](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(25)01249-8/fulltext)
-

Auditing Open Government Data Assets

This chapter outlines a reproducible process for auditing metadata and dataset changes of an agency's Data.gov harvest source and/or its comprehensive data inventory and individual data files. Additionally, non-inventoried information collection requests (IRCs) subject to the Paperwork Reduction Act (PRA) processes are also considered. The process makes use of several external resources, especially the Internet Archive's [Wayback Machine](#) (WBM) to (re)establish historical baselines should they not exist. The workflow relies on a set of convenience functions written in python that provides an audit package for federal open government data (mainly, [datagov-audit.py](#) and the [pra-icr-tools](#) package).

The entire workflow is described by the diagram, which outlines a temporal system designed to track changes to open government data. The workflow begins by establishing a baseline, referred to as T0, and comparing it against subsequent monitoring runs, or Tn, to detect content drift, link rot, metadata alterations, and availability issues. The workflow pulls metadata references from Data.gov, the specific agency's inventory, and (optionally) from the WBM. Nodes in the diagram prefixed with "CLI" (short for command-line interface) indicate specific helper functions provided by datagov-audit.py; for example, the workflow's list-harvest node is executed by the list-harvest CLI to enumerate active Data.gov streams. Similarly, the snapshot-org node utilizes the snapshot-org CLI to capture the raw Data.gov catalog of a specific agency, while the fetch-datajson node relies on the fetch-datajson CLI to download harvest sources from an agency. More details about Data.gov, including its limitations for monitoring changes to underlying datasets, are described in the [chapter on the Federal Data Catalog \(FDC\)](#). Parts of this workflow is similar to that used by Freilich & Kesselheim ([Data Manipulation within the US Federal Government, 2025](#)) in their forensic study of HHS data and metadata in early 2025.

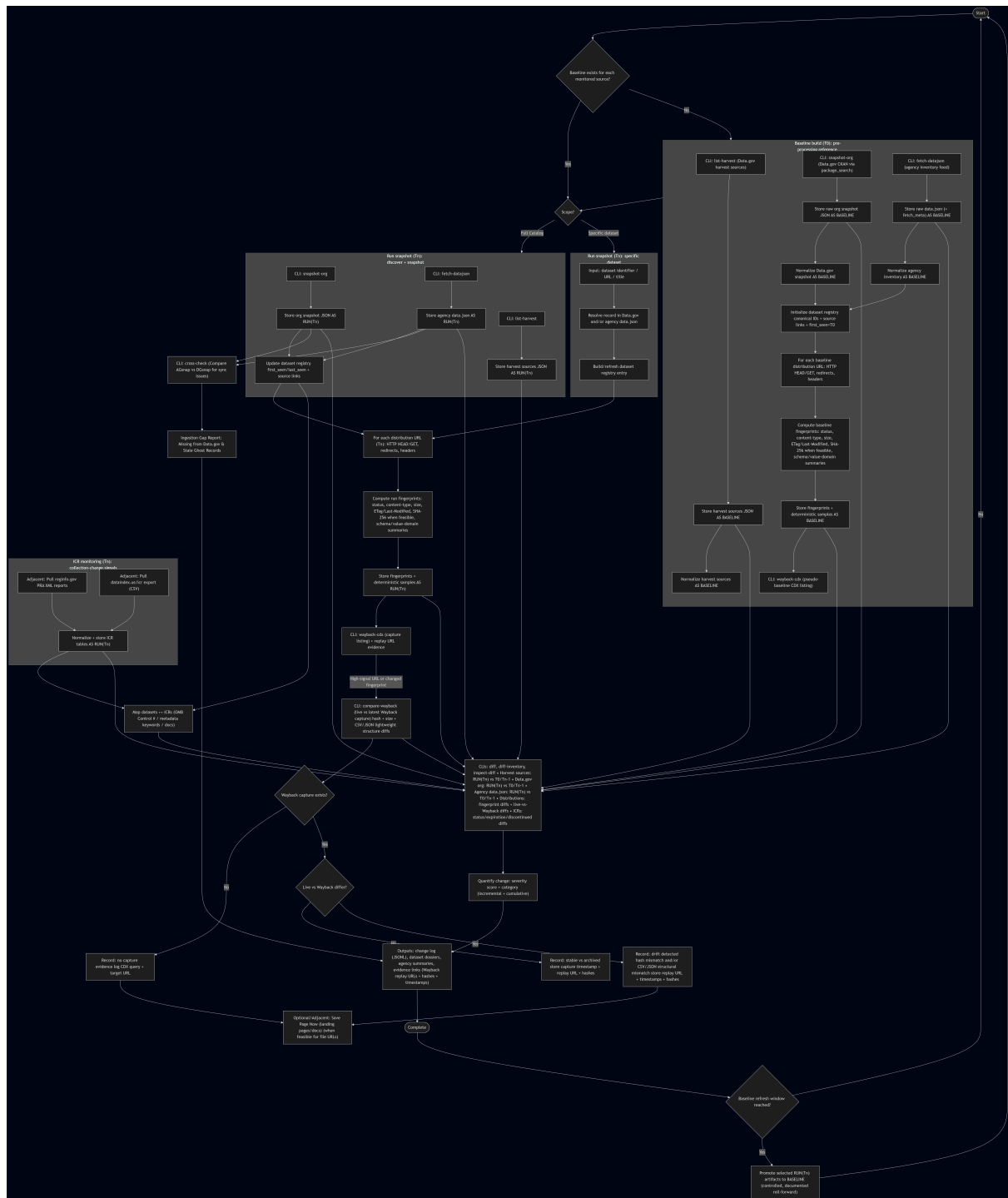


Figure: Workflow for Auditing Open Government Data Assets and Information Collections

When monitoring specific data distributions, such as CSV or JSON files, the workflow leverages the WBM to capture historical snapshots of the dataset's state. To determine if data available for download from an agency at a specific moment in time has drifted from this archived version, the compare-wayback node provides comparative statistics and information. Based on the file type detected, this step automatically evaluates the data for modifications using file size, SHA-256 hashes, and other structural differences. The function does not evaluate specific differences within a dataset per se, rather it simply evaluates whether differences exist between T0 and Tn instances.

Diff Engine and Related Subroutines

At the core of this monitoring system is the Diff Engine node, which is responsible for comparing the T0 baseline against the Tn run to quantify changes. For comparing Data.gov CKAN metadata over time, the workflow maps to the script's diff CLI. Conversely, diffing the agency's internal data.json files over time is handled by the diff-inventory CLI. The generation of human-readable dataset dossiers and

change logs from these comparisons is managed by the inspect-diff CLI, which iterates through the diff engine's output to reveal exact, field-level modifications.

Alongside the time-series Diff Engine, the workflow includes cross-check CLI, which compares two files from the exact same monitoring run (Tn): the agency's live data.json and Data.gov's live CKAN snapshot. By extracting and matching identifiers across the two different schemas, the cross-check CLI reports out whether the FDC has ingested the current version of the agency harvest source. This report feeds directly into the final outputs, revealing datasets the agency published but Data.gov failed to harvest. Ghost records lingering on Data.gov that the agency has already removed are also returned in this step (such as was the case for several months with more than 2000 datasets previously provided by the [United States Agency for International Development](#)).

FDC Harvest Sources

Data.gov harvests metadata that it populates in to the Federal Data Catalog across multiple sources at federal, state, and local agencies websites. Data.gov reports the list of all sources at: <https://catalog.data.gov/harvest>. According to the WBM snapshot of [January 15, 2025](#), which is closest snapshot immediately prior to the 2025 inauguration, there were 916 harvest sources available. As of 1/23/2026, there were 919.

The number of harvest sources can change on Data.gov for a few reasons. Harvest sources are sometimes combined or disaggregated and they can change formats (some harvest sources were updated in the last few years to be provided through an API for example) and are subject to outages like any other online resource.

Despite guidance from OMB and Congress that agencies are supposed to have a single comprehensive data inventory (CDI) made publicly available at *agency.gov/data.json*, which would serve as the harvest source for the Federal Data Catalog, many agencies have more than one source of harvest data due to the real-world complexities of managing a multiple information systems within agencies. The U.S. Census Bureau, for example, has [659 sources as of 1/23/2026](#), representing 70% of all the harvest sources used by Data.gov.

Historical data on the number of harvest sources can be obtained through WBM snapshots of <https://catalog.data.gov/harvest/>, a time-series of which is provided in the project data repository file: [datagov_harvest_counts-01282026.csv](#). The auditing workflow described here considers the general case of agency supplied harvest sources whether they represent a single CDI or are a collection of multiple inventories.

List of all datasets in the FDC

The Diff Engine node can take various, general, methods for comparing files. One useful purpose is to use the Diff Engine to evaluate changes to the total list of datasets indexed in the FDC. This requires an inventory as input at T0, which can be obtained through Data.gov's flexible CKAN API.

To generate a JSON object of all available (non-collection) datasets indexed in the Federal Data Catalog, the CKAN API has the following public baseurl:

```
https://catalog.data.gov/api/3/action/package_search
```

Alternatively, an API key can be [requested from GSA](#) and can be used with the following baseurl:

```
https://api.gsa.gov/technology/datagov/v3/action/package_search?api_key=API_KEY
```

Through either access point, the total number of assets is returned in the count element the top of the call:

```
{
  "help": "https://catalog.data.gov/api/3/action/name=package_search",
  "success": true,
  "result": {
    "count": 391082,
    "facets": {
    },
    ...
  }
}
```

Now, as documented on the [Data.gov user-guide](#), the collections of datasets are counted as a single dataset. This means that the total count displayed on the landing page, along with the figure returned in the count element above, is a significant undercount. The total number of collections can be found by counting up the datasets with field name "collection_package_id" present using the following API call:

```
https://catalog.data.gov/api/3/action/package_search?fq=collection_package_id:*&rows=0
```

which (as of 1/24/2026) returns:

```
{
  "help": "https://catalog.data.gov/api/3/action/help_show?name=package_search",
  "success": true,
  "result": {
    "count": 113495,
    "facets": {
    },
    "results": [],
    "sort": "views_recent desc",
    "search_facets": {
    }
  }
}
```

Adding the two numbers together provides the total number of datasets (at around the end of January 2026) of: 504,577. This value reflects the total, including non-federal data assets indexed by Data.gov. On January 23, 2026, GSA published a beta-version of the next iteration of the FDC at: <https://beta-catalog.data.gov>. In this version, each dataset contained within a collection is counted separately, reflecting a more accurate representation of the true number of data assets in the FDC. Of note, the calculation conducted here and the new beta-version of the FDC agree, sans a few hundred datasets. These statistics can also be recovered visually from the summary statistic graphs Data.gov website as evident in the screenshot compilation represented by the figure below:

By default the dataset results are limited to the first 20 responses of the API call (which can be increased to 1000 per 'page' using the API with a registered key). Two additional stand-alone scripts can be used to collect a complete inventory of all data assets indexed in Data.gov: [resume-data.catalog.py](#) for datasets not in a collection and [resume-datagov-collections.py](#) for datasets in a collection. These python scripts iterate through all results gracefully and handle interruption and have pause/resume capability. Both functions output a csv file that can then be concatenated together. For an auditing run done for this project, the output files are both available via the GitHub Large File Storage system in the project data repository: [datagov_inventory-01232026.csv](#) and [\[datagov_collections_inventory-01252026.csv\]](#) (https://www.github.com/cmarcum/data-integrity/tree/main/data/datagov_collections_inventory-01252026.csv), respectively.

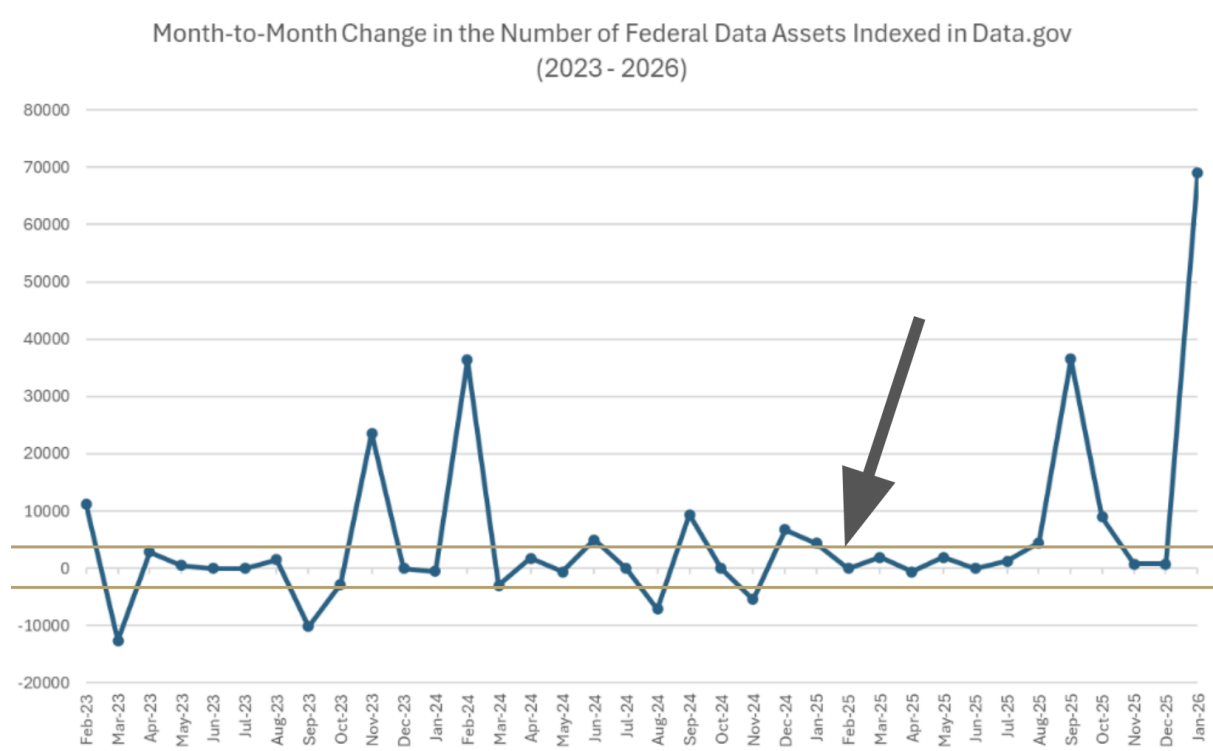


Figure: Month-to-Month Change in the Number of Federal Data Assets Indexed in Data.gov (2023 - 2026) with +/- 3000 lines to provide visual context to the magnitude of changes between January 2025 and February 2025 reported in the press (see arrow). Data were collected from WBM snapshots of Data.gov and are provided in the [project data repository file: datagov_collection_count-01262026.csv](#). The first successful WBM snapshot of Data.gov occurring each month between January 2023 and December 2025 (inclusive) were visited and the total number of datasets was recorded manually from information available in the static page. A major update to the format, layout, and presentation of information occurred in early March of 2023, which added a new presentation of metrics and statistics directly on the landing page, including statistics that disaggregate the number of datasets by source type. For snapshots visited prior to this change, corresponding snapshots of [catalog.data.gov/dataset](#) were also visited to collect the number of federal datasets indexed by Data.gov.

Using the WBM with the FDC and CDIs

Because the WBM does not take snapshots of the API URLs with much regularity, it is not a reliable source of status changes to the FDC via that route. There were just two snapshots of the API baseline URL in 2025 and the closest snapshot taken prior to the 2025 inauguration was done on [6/28/2024](#).

Instead, historical data on the number of reported datasets indexed into the FDC should be collected by scraping WBM snapshots of the relevant Data.gov pages (and the underlying source code). The WBM's CDX API provides a convenient way to access snapshots that were crawled and stored by the Internet Archive in the past. For instance, the following API call collects the snapshot timestamps from January 1st, 2023 to December 31st, 2025 and returns the result as a json object:

```
https://web.archive.org/cdx/search/cdx?url=data.gov&from=20230101&to=20251231&collapse=timestamp:6&output=json&fl=timestamp,or:
```

The associated CLI auditing scripts handle the fetching in python to write out a csv file containing each snapshot's WBM URL, basically like this:

```
import requests

url = "https://web.archive.org/cdx/search/cdx?url=data.gov&from=20230101&to=20251231&collapse=timestamp:6&output=json&fl=timestamp"
data = requests.get(url).json()

with open("snapshots.csv", "w") as f:
    for row in data[1:]:
        f.write(f"http://web.archive.org/web/{row[0]}/{row[1]}\n")
```

For one auditing run, the resulting list of URLs was saved in [snapshots_datagov-01012026.csv](#).

From the April 2023 to the December 2025 snapshots, the number of datasets can be extracted from the snapshot.html. Presumably, this will work on contemporaneous (live) versions of data.gov, assuming that GSA does not make updates to the landing page structure after the date of publication of this project:

```
import csv, requests, re
from bs4 import BeautifulSoup

# Input: 'urls.csv' (URLs in first column) | Output: 'results.csv'
with open('urls.csv', 'r') as f_in, open('results.csv', 'w', newline='') as f_out:
    writer = csv.writer(f_out)
    for row in csv.reader(f_in):
        try:
            html = requests.get(row[0], headers={'User-Agent': 'Mozilla/5.0'}).text
            text = BeautifulSoup(html, 'html.parser').find('div', class_='hero__dataset-count').get_text()
            count = re.search(r'(\d,)+', text).group(1).replace(',', '')
            writer.writerow([row[0], count])
        except:
            pass
```

Agency Comprehensive data inventory differencing

Individual agency CDIs, usually stored as json files on agency websites, are not archived with much regularity by the WBM. The WBM snapshots also inject noise into the json files because its archiving routines attempt to automatically add the snapshot's WBM URL prefix to each link in the file, even if the endpoints of those URLs do not lead to pages that were part of the snapshot. For example, a recent snapshot of EPA's comprehensive data inventory by the WBM is available here (taken on August 7th, 2025 at 5:34pm GMT):

[<https://web.archive.org/web/20240807053438/https://pasteur.epa.gov/metadata.json>] (<https://web.archive.org/web/20240807053438/https://pasteur.epa.gov/metadata.json>). The prefixing issue is apparent in many of the *accessURL* (the DCAT-US endpoint for where data can be accessed from) for the datasets even when those datasets are stored outside of EPA's own server. For instance, "*The Social Cost of Ozone-Related Mortality Impacts From Methane Emissions - Associated Model Data and Code*" dataset is stored on Zenodo, where it can be associated with related publications and other derivatives. The URL in the WBM archive, however, points to <https://web.archive.org/web/20240807053438/https://doi.org/10.5281/zenodo.8276748> not the correct record at <https://doi.org/10.5281/zenodo.8276748>.

To resolve this issue, any archived json files acquired from the WBM for auditing purposes should be modified using a simple regular expression that searches for "https://web.archive.org/web/[PREFIX]/" (where [PREFIX] references the WBM prefix associated with a specific snapshot) and replaces it with an empty string. This enables direct comparison between WBM archived versions of the json files and between a WBM archived version and the live version of the json files.

Monitoring Changes to Information Collection Requests

Finally, in addition to direct monitoring of open government data assets from the FDC and CDIs, the workflow incorporates an Information Collection Request (ICR) subroutine. This component acts as a fallback when the underlying datasets associated with federal collections are not properly indexed or published within an agency's CDI. To monitor these changes independently of the agency catalogs, the workflow utilizes external resources like [dataindex.us](#) and the [pra-icr-tools](#) package. By pulling Paperwork Reduction Act (PRA) XML reports, CSV exports, and collection documents, this subroutine normalizes and stores ICR tables during each monitoring run. It then maps these collections back to specific datasets using metadata keywords, documentation, or OMB Control Numbers.

What the routine misses

The auditing routine relies on catalogs of information related to federal data assets, including the Federal Data Catalog provided by Data.gov, agency comprehensive data inventories, and the inventory of information collection requests provided by reginfo.gov (or alternatively by dataindex.us). The workflow does not consider, however, non-inventoried sets of data that may be available to the public on discrete agency websites. For instance, the non-inventoried data assets that are publicly accessible via download from the [PEPVAR](#) website would not be captured in the process. However, since the WBM crawls most publicly accessible federal websites, it should be possible to monitor tracking and changes using the WBM CLI tools/steps.

The workflow could be modified to accommodate these edge-cases by adding a sub-routine that:

- inputs a list of websites that have data assets available for download
- archives each website
- runs the Diff engine comparing each website to snapshots archived in the WBM
- captures each dataset available on each website and a snapshot from the WBM
- runs the Diff engine on each dataset available on each website comparing to the WBM snapshot versions
- updates the list and continues

EPA ScienceHub Data.gov Inventory Audit Workflow Example

To illustrate the workflow in action, consider the Environmental Protection Agency’s (EPA) ScienceHub, which supplies a data inventory .json file that is used by Data.gov as a harvest source. ScienceHub is the agency’s central catalog for research datasets, models, code, and other scientific products generated across its laboratories. It was built to make EPA science easier to find, reuse, and evaluate, drawing heavily on the work historically produced by the Office of Research and Development (ORD), which was the program that coordinated EPA’s research programs, maintained specialized labs, and ensured scientific rigor across environmental and public-health studies.

That foundation was disrupted when the Trump administration eliminated ORD in 2025, replacing it with a new structure and reducing the agency’s independent research capacity. Because ScienceHub depends on the continuity of EPA’s scientific programs, the loss of ORD introduced uncertainty about future dataset production, long-term monitoring, and the stewardship of existing scientific resources, making it a good candidate for this exercise.

Step 0: Establish Baseline and Current Dates

Before pulling data, decide what you are comparing today’s data against (e.g., January 1, 2025 vs. Today).

Condition	Action	Next Step in Workflow
Local Baseline Exists	You already saved a metadata.json file from the agency at a previous date.	Skip to Step 5 (use your local file as the old input).
No Local Baseline	You do not have historical files saved.	Proceed to Step 1 , then extract a snapshot from the Internet Archive in Step 4.

For completeness, this workflow assumes no local baseline files exist.

Step 1: Capture Current List of All Harvest Sources

Download a master list of everything Data.gov is currently configured to harvest to get a landscape view of active sources.

```
python datagov-audit.py list-harvest --out harvest_sources.json

The first few lines of the output of this command saved in harvest_sources.json will look similar to this:

head harvest_sources.json -n 9
[
  {
    "id": "cc7df4cc-8036-4868-b422-5823c63957d7",
    "title": "Exim Data.json",
    "source_url": "https://img.exim.gov/s3fs-public/dataset/vbhv-d8am/data.json",
    "frequency": "WEEKLY",
    "source_type": "datajson"
  },
```

This file is useful for other purposes as well, as it can be used to monitor the changes to the Data.gov harvest sources.

Step 2: Confirm the Source Exists

Search the master list you just downloaded to verify the EPA Pasteur metadata source is present and active on Data.gov. While this can be done using regular expressions, the datagov audit package provides a convenience function for this purpose with fuzzy-matching:

```
python datagov-audit.py find-datajson harvest_sources.json pasteur.epa.gov

Which find the harvest source for epa’s ScienceHub and reports the entry to the terminal:

[
  {
    "id": "04b59eaf-ae53-4066-93db-80f2ed0df446",
    "title": "EPA ScienceHub",
    "source_url": "https://pasteur.epa.gov/metadata.json",
    "frequency": "DAILY",
    "source_type": "datajson"
  }
]
```

Step 3: Grab Live Version of Harvest Source Data Inventory

Download and validate the current, live metadata inventory directly from the EPA. The datagov audit scripts provide a convenience function for wrapping the CURL call to download the file from the agency website. This serves as the ‘current’ or ‘new’ file for comparison.

```
python datagov-audit.py fetch-datajson https://pasteur.epa.gov/metadata.json --out epa-pasteur-metadata.json
```


Step 4A: Find the Baseline Wayback Capture

Use the Internet Archive's Wayback Machine CDX query tool to list captures from a specific target date (e.g., 01/01/2025) to capture a snapshot with a timestamp closest to your baseline date for comparison to the live version.

```
python datagov-audit.py wayback-cdx https://pasteur.epa.gov/metadata.json --from-ts 20250101 --fetch
```

The output will print options closest to the specified timestamp which can be used to locate the exact timestamp from the output (e.g., 20250105123000) for the next step.

Step 4B: Grab and Clean the Baseline Wayback Snapshot

Using the timestamp found in 4A, download the raw JSON from the Wayback Machine (using the `id_modifier`) and clean it to strip out any Wayback-injected HTML or URL rewriting.

```
python datagov-audit.py clean-wayback "https://web.archive.org/web/20250105123000id_/https://pasteur.epa.gov/metadata.json" --i
```

Because WBM adds relay prefixes to URLs found in its snapshots that forward to other snapshots, this script automatically cleans them up so that comparisons with non-WBM snapshot equivalent files can be made.

Step 5: Compare WBM Baseline File to Current File

The `datagov-audit` script contains a useful differencing engine that can be used to compare the cleaned baseline against today's live inventory. This generates a machine-readable JSON report of all added, removed, and modified datasets.

```
python datagov-audit.py diff-inventory pasteur-wbm-20250105-metadata.json epa-pasteur-metadata.json --out pasteur-compare-diff
```

In addition to a machine-readable json list of added, removed, and modified datasets, the function prints a high-level summary to the terminal:

```
{
  "removed": 0,
  "added": 800,
  "modified": 29,
  "old_total": 3457,
  "new_total": 4257,
  "old_indexed": 3457,
  "new_indexed": 4257
}
```

Step 6: See Which Dataset Metadata Entries Were Modified

Convert the machine-readable JSON report into a human-readable text file that shows exactly which fields changed (e.g., description updates, modified dates, new contact emails) for every modified record.

```
python datagov-audit.py inspect-diff --report pasteur-compare-diff.json --out pasteur-changes.txt
```

In this case, inspecting the report reveals that the 29 modified metadata elements were mostly changed to update versions of files, new downloadURLs, or to remove the "Office of Research and Development" as the publishing office (see Step 9, below, for an example of five such cases).

Step 7: Evaluate Modification of a Specific Downloadable Dataset

Often, metadata changes alone are insufficient to reveal whether underlying data were also modified. The `datagov-audit` scripts can also do rough checks of whether the actual downloadable files (i.e., .zip, .csv, .json) were also modified since the baseline date (assuming that the WBM actually captured the file around that time). The script mathematically evaluates differences in the files using hashing (specifically, by using the SHA-256 algorithm to create a digital fingerprint of the files to compare).

As an example, we grab one of the files where the metadata had been modified. In this case, it's the file associated with advanced septic systems pilot data: <https://catalog.data.gov/dataset/performance-data-for-enhanced-innovative-alternative-i-a-septic-systems-for-nitrogen-remov>.

```
python datagov-audit.py compare-wayback "https://pasteur.epa.gov/uploads/10.23719/1529539/V1%20data%20release.zip" --from-ts 20
```

In this specific case, the underlying data have not been modified and the output clearly shows no changes in either the file sizes or the underlying structure per the hashing routine:

```
{
  "ok": true,
  "target_url": "https://pasteur.epa.gov/uploads/10.23719/1529539/V1%20data%20release.zip",
  "format": "bytes",
  "live": {
    "final_url": "https://pasteur.epa.gov/uploads/10.23719/1529539/V1%20data%20release.zip",
    "content_type": "application/zip",
    "bytes": 873201,
    "sha256": "d9772818ebee03f4e099e5d3d6b47dcf3266640f9df369f2f36fb580b7c6e2"
  },
  "wayback": {
    "capture": {
      "timestamp": "20240621201154",
      "original": "https://pasteur.epa.gov/uploads/10.23719/1529539/V1%20data%20release.zip",
      "statuscode": "200",
      "mimetype": "application/zip",
      "digest": "BA7N5KHZEYODCVGQBLSP2P43QS5BADTV",
      "length": "838168"
    },
    "replay_url_raw": "https://web.archive.org/web/20240621201154id_/https://pasteur.epa.gov/uploads/10.23719/1529539/V1%20dat",
    "bytes": 873201,
    "sha256": "d9772818ebee03f4e099e5d3d6b47dcf3266640f9df369f2f36fb580b7c6e2"
  },
  "match": {
    "sha256_equal": true,
    "bytes_equal": true
  }
}
```

This routine could easily be extended to recursively iterate over a set of datasets. Of course, diligence to assess the fidelity of all federal data assets by a single entity would be infeasible but stakeholders with special interests in specific datasets could be done with some regularity.

Step 8: Verify Data.gov Ingestion

Finally, verify that Data.gov's harvester has successfully ingested the agency's updates into the public catalog.

```
python datagov-audit.py snapshot-source "04b59eaf-ae53-4066-93db-80f2ed0df446" --out datagov-pasteur-snapshot.json
```

Running this function periodically will provide a mechanism to establish new baselines directly from what Data.gov ingests and uses in the Federal Data Catalog. Two such instances can be compared using the `diff` command, or they can be to ensure Data.gov stays in sync with the EPA's live file.

Two such snapshots can also be compared using the `diff-inventory` and `inspect-diff` functions, but with minor argument changes to ensure the json field mappings are comparable:

```
python datagov-audit.py diff-inventory datagov-old-snapshot.json datagov-new-snapshot.json --dataset-key packages --id-field id
python datagov-audit.py inspect-diff --report ckan-diff.json --id-field id --out ckan-changes.txt
```

Step 9: Compare Federal Data Catalog with Agency Inventory

```
python datagov-audit.py cross-check epa-pasteur-metadata.json datagov-pasteur-snapshot.json --out cross-check-report.json
{
  "agency_total": 4257,
  "ckan_total": 4258,
  "missing_from_datagov": 0,
  "extra_on_datagov": 1,
  "common": 4257
}
```

In this case, one dataset appears in the FDC that is not in the agency inventory, and that's because there was a one-day difference between the time the two files were downloaded (with the Data.gov version being more recent). That dataset was added on 3/2/2026: [Assessing Flooding from Changes in Extreme Rainfall: Using the Design Rainfall Approach in Hydrologic Modeling](#). No datasets appear to be missing.

However, some changes did occur. A closer inspection by looking at the difference in the names of the data assets reveals that were five additional entries that appear in the WBM archive of the EPA comprehensive data inventory that do not appear in the version live as of 1/1/2025. These are:

- *EnviroAtlas - 2010 Dasymetric Population for the Conterminous United States v3 (In Review)*
- *Chironomid nitrogen stable isotope data for NARS 2007, 2008, and 2009 surveys*
- *Performance data for enhanced Innovative/Alternative (I/A) septic systems for nitrogen removal installed in a field demonstration in Barnstable, Massachusetts (2021 - 2023)*
- *Convex Hulls for Species Richness and Invivudal Species*
- *Bioenergy Senario Model (BSM) data from Miller et al*

None of these datasets appear to have been removed due to undue political interference. The first entry, referencing an *in review* version of EnviroAtlas, has been merged into the new "*EnviroAtlas - 2010 Dasymetric Population for the Conterminous United States v3*" dataset. The second entry, has also been merged with new data and combined added as new dataset: "*Chironomid nitrogen stable isotope data for NARS 2007-2009, 2012-2014 surveys*". The third entry has been replaced with an updated dataset: "*Performance data for enhanced Innovative/Alternative (I/A) septic systems for nitrogen removal installed in a field demonstration in Barnstable, Massachusetts (2021 -*

2023). *Version 2*". The fourth and fifth entries were removed and replaced to correct the typo in their respective title fields: "*Convex Hulls for Species Richness and Individual Species*" and "*Bioenergy Scenario Model (BSM) data from Miller et al*".

Conclusion

This chapter provided a replicable workflow for auditing and evaluating the integrity of federal data, using the EPA's ScienceHub repository as a practical example. By combining historical baselines from the Wayback Machine and contemporary and historical snapshots of agency files, with the automated comparative routines, independent monitors can effectively track data disruptions, detect silent metadata alterations, and verify catalog synchronization. The audit routine can be adapted for dataset-specific, agency-wide, or system-wide monitoring. While this specific example focused on the EPA, these methods are highly generalizable. The cross-schema checks, diffing engines, and hashing routines should operate on almost any JSON-based inventory.

References

1. Alder, M. (2025). White House finalizes OPEN Government Data Act guidance, restarts CDO Council. *FedScoop*. <https://fedscoop.com/white-house-open-government-data-act-restarts-cdo-council/>
2. Alder, S. (2025). *HHS Settlement Requires Restoration of 100+ Health Datasets and Tools*. <https://www.hipaajournal.com/hhs-settlement-lawsuit-restore-critical-health-information-federal-websites/>
3. Archive, N. S. (2025). *Disappearing Data: Trump Administration Removing Climate Information*. <https://nsarchive.gwu.edu/briefing-book/climate-change-transparency-project-foia/2025-02-06/disappearing-data-trump>
4. Auerbach, J., Bowen, C. M. K., Citro, C., Pierson, S., Potok, N., & Seeskin, Z. (2024). *The Nation's Data at Risk: Meeting America's Information Needs for the 21st Century*. American Statistical Association. <https://www.amstat.org/docs/default-source/amstat-documents/the-nation's-data-at-risk-report.pdf>
5. Beitsch, R. (2025). USAID order to delete classified records sparks flurry of litigation. *The Hill*. <https://thehill.com/homenews/administration/5191064-usaids-document-destruction/>
6. Bouton, L. J. A., & Redfield, E. (2026). *Removal of Sexual Orientation and Gender Identity from Federal Data Collections: January 2025 to January 2026*. Williams Institute. <https://williamsinstitute.law.ucla.edu/publications/sogi-data-collection-removal/>
7. Bowen, C. M. K., Citro, C., Crosby, M., Pierson, S., Potok, N., & Seeskin, Z. (2025). *The Nation's Data at Risk: 2025 Report*. The American Statistical Association. <https://www.amstat.org/policy-and-advocacy/the-nations-data-at-risk-2025-report>
8. Brady, J. (2025). More environmental data is deleted in Trump's second term. *NPR*. <https://www.npr.org/2025/08/08/nx-s1-5495338/climate-change-environment-websites-trump>
9. CEJ. (2025). *Climate and Economic Justice Screening Tool (CEJST)*. Harvard Dataverse. [10.7910/DVN/B6ULET](https://dataverse.harvard.edu/citation?persistentId=doi:10.7910/DVN/B6ULET)
<https://dataverse.harvard.edu/citation?persistentId=doi:10.7910/DVN/B6ULET>
10. Cheney, K., & Gerstein, J. (2025). Appeals court rules Trump clamp-down on spending data defies Congress' authority. *Politico*. <https://www.politico.com/news/2025/08/09/appeals-court-rules-trump-clamp-down-on-spending-data-defies-congress-authority-00501348>
11. Clark, L. (2026). Judge dismisses lawsuit over feds' climate data erasure. *E&E News*. <https://subscriber.politicopro.com/article/eenews/2026/03/13/judge-dismisses-lawsuit-over-feds-climate-data-erasure-cw-00824893>
12. Dayak, S., & Kramer, A. (2026). Federal Data Is Disappearing. *NOTUS*. <https://www.notus.org/trump-white-house/federal-data-is-disappearing>
13. Elkind, P. (2025). The Latest Trump and DOGE Casualty: Energy Data. *ProPublica*. <https://www.propublica.org/article/the-latest-trump-and-doge-casualty-energy-data>
14. Emma, C. (2025). Appeals court rules Trump clamp-down on spending data defies Congress' authority. *Politico*. <https://www.politico.com/news/2025/08/09/appeals-court-rules-trump-clamp-down-on-spending-data-defies-congress-authority-00501348>
15. Fiorentino, D. A., & Riccard, T. N. (2025). *Office of Management and Budget (OMB) Reporting on Apportionments* (CRS Insight No. IN12538; Issue IN12538). Congressional Research Service. <https://crsreports.congress.gov/product/pdf/IN/IN12538>
16. Fischer, W. (2025). *Letter to Mr. Christopher Colbrow, Agency Records Officer, US Agency for International Development*. <https://www.archives.gov/files/records-mgmt/resources/ud-2025-0047-usaid-open.pdf>
17. Fowler, S., Joffe-Block, J., & Bond, S. (2026). The government is investigating new claims that DOGE misused Social Security data. *NPR*. <https://www.npr.org/2026/03/11/nx-s1-5745153/doge-social-security-data-whistleblower-investigation>
18. Frazin, R. (2025). Energy Secretary Chris Wright says admin 'reviewing' past climate reports. *The Hill*. <https://thehill.com/policy/energy-environment/5441347-energy-department-chris-wright-national-climate-assessment-review/>
19. Freilich, J., & Kesselheim, A. S. (2025). *Data manipulation within the US Federal Government*. [https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(25\)01249-8/fulltext](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(25)01249-8/fulltext)
20. Garza, F. (2026). After a lawsuit, USDA agrees to share climate risk data with farmers. *Government Executive*. <https://www.govexec.com/management/2026/03/after-lawsuit-usda-agrees-share-climate-risk-data-farmers/411848/>
21. Gehrke, G. (2026). The Trump administration is disappearing climate change data. *The Hill*. <https://thehill.com/opinion/energy-environment/5756951-federal-data-threatened-trump-era/>
22. Hamad, R., Warren, M., Ross, D., Maury, M., & Jarosz, B. (2026). *Rapid Response Data Briefing: Pregnancy Risk Assessment Monitoring System (PRAMS)*. Webinar hosted by dataindex.us, Association of Public Data Users, Population Reference Bureau, March of Dimes, and Population Association of America. <https://dataindex.us/events/Rapid-Response-Data-Briefing-Pregnancy-Risk-Assessment-Monitoring-System-PRAMS>
23. Hartman, M. (2025). Federal data has been disappearing under Trump. *Marketplace*. <https://www.marketplace.org/story/2025/07/28/federal-data-has-been-disappearing-under-trump>
24. Heckman, J. A. (2025). 'Bedrock' federal data sets are disappearing, as statistical agencies face upheaval. *Federal News Network*. <https://federalnewsnetwork.com/big-data/2025/12/bedrock-federal-data-sets-are-disappearing-as-statistical-agencies-face-upheaval>
25. Heilweil, R. (2025). GAO says OMB takedown of apportionments website violates federal statutes. *FedScoop*. <https://fedscoop.com/gao-omb-takedown-apportionments-website-federal-statutes/>
26. Hill, C. (2025). GAO says OMB takedown of apportionments website violates federal statutes. *Fedscoop*. <https://fedscoop.com/gao-omb-takedown-apportionments-website-federal-statutes/>
27. Hirji, Z. (2025). Six Environmental Mapping Tools the White House Doesn't Want You to See. *Bloomberg*. <https://www.bloomberg.com/news/articles/2025-05-07/six-environmental-mapping-tools-the-white-house-doesn-t-want-you-to-see>
28. Jones, L. A. (2025). A Shortlist of Federal Data the Trump Administration Has Tampered With or Destroyed. *Talking Points Memo*. <https://talkingpointsmemo.com/news/a-shortlist-of-federal-data-the-trump-administration-has-tampered-with-or-destroyed>
29. Katz, E. (2026). Court orders OMB to publish more info about how federal funding is distributed. *Government Executive*. <https://www.govexec.com/management/2026/01/court-orders-omb-publish-more-info-about-how-federal-funding-distributed/411092/>
30. Katz, E. (2025). Spending transparency data posted by Trump budget office after court order. *Government Executive*. <https://www.govexec.com/management/2025/08/spending-transparency-data-posted-trump-budget-office-after-court-order/407548/>
31. Kiersz, A. (2026). America's economy is 'driving through the fog.' *Business Insider*. <https://www.businessinsider.com/disappearing-economic-data-bad-economy-recession-unemployment-bls-jobs-report-2026-2>
32. Kim, A. (2025). Federal Data Are Disappearing. *Washington Monthly*. <https://washingtonmonthly.com/2025/11/26/federal-data-are-disappearing/>
33. Klein, M., & Medina, C. (2026). One Year In: The Cost of Rolling Back Federal LGBTQ Data. *America's Data Index*. <https://dataindex.us/newsletter/article/f69108a5-7346-4257-9fbc-e2ef784bd96b>
34. Koebler, J. (2025). Archivists Work to Identify and Save the Thousands of Datasets Disappearing from Data.gov. *404 Media*.

35. Kundra, V. (2009). *Data.gov Launch Announcement*. U.S. General Services Administration, Technology Transformation Services. <https://data.gov/timeline/>
36. Kutz, A. (2025). How much federal data has Trump really purged? *NewsNation*. <https://www.newsnationnow.com/politics/trump-federal-datasets-websites/>
37. LaRose, A. (2025). *Internal Memorandum* [Email sent to all staff]. U.S. Energy Information Administration, Office of Energy Analysis. <https://www.documentcloud.org/documents/25928821-larose-memo/>
38. Levenstein, M., & Kubale, J. (2025). Data that taxpayers have paid for and rely on is disappearing – here's how it's happening and what you can do about it. *The Conversation*. <https://theconversation.com/data-that-taxpayers-have-paid-for-and-rely-on-is-disappearing-heres-how-its-happening-and-what-you-can-do-about-it-251787>
39. Malesky, E. (2025). *2025 Letter from the Director - Duke Center for International Development*. <https://dcid.sanford.duke.edu/2025-letter-director/>
40. Mandel, K. (2026). The Women Saving America's Climate Data. *TIME*. <https://time.com/7344773/women-saving-federal-climate-data/>
41. Marcum, C. S. (2026). The World Factbook was a Valuable Data Resource. In *Open Evidence*. <https://www.doi.org/10.59350/npagv-r3682> <https://www.chrismarcum.com/marcum-blog/2026/02/06/The-World-Factbook-Was-A-Valuable-Data-Resource.html>
42. Mauran, C. (2025). Thousands of datasets from Data.gov have disappeared since Trump's inauguration. What's going on? *Mashable*. <https://mashable.com/article/government-datasets-disappear-since-trump-inauguration>
43. Maury, M., & Marcum, C. (2025). How You Can (and should) Shape Federal Data Collections. *America's Data Index*. <https://dataindex.us/newsletter/article/6cfecae3-3c89-487e-b8f7-cfcd85ded6c7>
44. Maury, M., & Ross, D. (2026). *Take Action: How to Write a Public Comment on Federal Data*. America's Data Index. <https://dataindex.us/events/Take-Action-How-to-Write-a-Public-Comment-on-Federal-Data>
45. Mys, A. (2025). HHS Rescinds Richardson Waiver, Reducing Public Input in Rulemaking. *American Bar Association*. https://www.americanbar.org/groups/health_law/news/2025/3/hhs-rescinds-richardson-waiver-reducing-public-input-in-rulemaking/
46. Noor, D. (2025). Green groups sue Trump administration over climate webpage removals. *The Guardian*. <https://www.theguardian.com/us-news/2025/apr/15/trump-climate-webpage-removal-lawsuit>
47. O'Leary, M. (2025). Data Rescue Project Thwarts Government Censorship Wave. *Information Today*. <https://www.infotoday.com/IT/nov25/OLeary-Data-Rescue-Project-Thwarts-Government-Censorship-Wave.shtml>
48. Obama, B. (2009). *Memorandum on Transparency and Open Government*. Presidential Memorandum. <https://obamawhitehouse.archives.gov/the-press-office/2009/07/06/transparency-and-open-government>
49. Palmer, K. (2025). Preserving the Federal Data Trump Is Trying to Purge. *Inside Higher Ed*. <https://www.insidehighered.com/news/government/science-research-policy/2025/06/10/preserving-federal-data-trump-trying-purge>
50. Popova, Y. (2026). CKAN Turns 20: Two Decades of Open Data Infrastructure. *CKAN Blog*. <https://ckan.org/blog/ckan-turns-20-two-decades-of-open-data-infrastructure>
51. Rabin, R., & Mandavilli, A. (2025). CDC Web Pages and Data Vanish Following Trump's DEI and Gender Orders. *The New York Times*. <https://www.nytimes.com/2025/01/31/health/trump-cdc-dei-gender.html>
52. Reiss, D. (2025). Administrative Changes That Decrease Transparency at HHS. *The Regulatory Review*. <https://www.theregview.org/2025/03/24/reiss-administrative-changes-that-decrease-transparency-at-hhs/>
53. Riley, W. T., & Blizinsky, K. D. (2017). Implications of the 21st century cures act for the behavioral and social sciences at the national institutes of health. *Health Education & Behavior*, 44(3), 356–359. <https://doi.org/10.1177/1090198117707964>
54. Robbins, R. (2025). *STAT is backing up and monitoring CDC data in real time: See what's changing*. STAT News. <https://www.statnews.com/2025/02/14/tracking-cdc-data-changes-trump-executive-order-targets-gender/>
55. Satter, R. (2025). Harvard Law Library acts to preserve government data amid sweeping purges. *Reuters*. <https://www.reuters.com/world/us/harvard-law-library-acts-preserve-government-data-amid-sweeping-purges-2025-02-06/>
56. Schilling, E., & Slowey, E. (2026). IRS Improperly Shares Immigrants' Data with ICE: Explained. *Bloomberg Tax*. <https://news.bloombergtax.com/daily-tax-report/irs-overshares-thousands-of-immigrants-data-with-ice-explained>
57. Smith, J.-M. (2025). USDA cancels survey tracking how many Americans struggle to get enough food. *NPR*. <https://www.npr.org/2025/09/22/nx-s1-5549115/usda-food-insecurity-survey-hunger>
58. Smith, M. (2026). America's Statistical System Is Breaking Down. *Bloomberg*. <https://www.bloomberg.com/news/articles/2026-01-09/why-the-trump-administration-is-choosing-not-to-collect-some-us-data>
59. Stone, R. (2025). *Researchers from China and five other 'countries of concern' barred from NIH databases*. Science. <https://www.science.org/content/article/researchers-china-and-five-other-countries-concern-barred-nih-databases>
60. Stuessy, M. M., & Knoedl, T. R. (2026). *Availability of Federal Data: Policy Considerations for Disclosure, Preservation, and Governance*. Congressional Research Service, Library of Congress. <https://www.congress.gov/crs-product/R48889>
61. Wang, H. L. (2026). The Trump administration is adding a citizenship question to the 2030 census. *NPR*. <https://www.npr.org/2026/03/09/nx-s1-5613878/us-census-citizenship-question-redistricting>
62. Willson, M. (2025). Groups archive environmental justice data scrapped by Trump. *E&E News*. <https://www.eenews.net/articles/groups-archive-environmental-justice-data-scrapped-by-trump/>
63. Zimmermann, A. (2026). *FY 2026 R&D Appropriations: Final R&D Report*. AAAS. https://www.aaas.org/sites/default/files/2026-02/Final%20Report%202026_0.pdf
64. American Statistical Association. (2025). *Bureau of Labor Statistics*. <https://www.amstat.org/docs/default-source/amstat-documents/the-nations-data-at-risk-2025/bureau-of-labor-statistics.pdf>
65. Citizens for Responsibility and Ethics in Washington. (2025). *OMB's latest effort to conceal spending data*. CREW Investigations. <https://www.citizensforethics.org/reports-investigations/crew-investigations/ombs-latest-effort-to-conceal-spending-data/>
66. Congressional Research Service. (2022). *The OPEN Government Data Act: A Primer* (No. IF12299; Issue IF12299). Congressional Research Service. <https://www.congress.gov/crs-product/IF12299>
67. Data Foundation. (2025). *Taking Stock of Federal Open Data in 2025*. Data Foundation Blog. <https://datafoundation.org/news/blogs/707/707-Taking-Stock-of-Federal-Open-Data-in->
68. FitzSimons, C. USDA's Decision to End 30-Year Food Security Report Will Hide the Struggle of Millions of Families to Put Food on the Table. In *Food Research and Action Center*. Retrieved March 4, 2026, from <https://frac.org/news/foodsecuritysurveyterminationsept25>
69. KFF. (2025). *A Look at Federal Health Data Taken Offline*. KFF Policy Watch. <https://www.kff.org/policy-watch/a-look-at-federal-health-data-taken-offline/>
70. National Science and Technology Council. (2023). *A Framework for Federal Scientific Integrity Policy and Practice*. White House Office of Science and Technology Policy. <https://bidenwhitehouse.archives.gov/wp-content/uploads/2023/01/01-2023-Framework-for-Federal-Scientific-Integrity-Policy-and-Practice.pdf>
71. National Security Archive. (2025). *Disappearing Data: Trump Administration Removing Climate Information from Government Websites*. National Security Archive, George Washington University. <https://nsarchive.gwu.edu/briefing-book/climate-change->

- [transparency-project-foia/2025-02-06/disappearing-data-trump](https://www.whitehouse.gov/wp-content/uploads/2025/01/M-25-05-Phase-2-Implementation-of-the-Foundations-for-Evidence-Based-Policymaking-Act-of-2018-Open-Government-Data-Access-and-Management-Guidance.pdf)
72. Office of Management and Budget. (2025). *Phase 2 Implementation of the Foundations for Evidence-Based Policymaking Act of 2018: Open Government Data Access and Management Guidance*. <https://www.whitehouse.gov/wp-content/uploads/2025/01/M-25-05-Phase-2-Implementation-of-the-Foundations-for-Evidence-Based-Policymaking-Act-of-2018-Open-Government-Data-Access-and-Management-Guidance.pdf>
 73. Office of Management and Budget. (2025). *Guidance on Implementing Section 3(e) of Executive Order 14168 in Accordance with the Paperwork Reduction Act and the Privacy Act*. Office of Information and Regulatory Affairs.
 74. Office of Management and Budget. (2016). *Circular No. A-130: Managing Information as a Strategic Resource*. Executive Office of the President. https://www.whitehouse.gov/wp-content/uploads/legacy_drupal_files/omb/circulars/A130/a130revised.pdf
 75. Office of Management and Budget. (2009). *OMB Memorandum M-10-06 - Open Government Directive*. <https://obamawhitehouse.archives.gov/open/documents/open-government-directive>
 76. Rep. Don Beyer. (2025). *79 U.S. Representatives Demand the Restoration of Public Access to Federal Data Sets Purged by the Trump Administration*. Official Website of U.S. Representative Don Beyer. <https://beyer.house.gov/news/documentsingle.aspx?DocumentID=6384>
 77. Reuters. (2025). *NOAA "fully staffed" with forecasters, scientists, US commerce secretary says*. <https://www.reuters.com/world/us/noaa-fully-staffed-with-forecasters-scientists-us-commerce-secretary-says-2025-06-04/>
 78. Reuters. (2025). *White House aims to eliminate NOAA climate research in budget plan*. <https://www.reuters.com/sustainability/climate-energy/white-house-proposes-eliminate-noaa-climate-research-budget-proposal-2025-04-11/>
 79. U.S. Bureau of Labor Statistics. (2025). *Notice of CPI collection reductions*. <https://www.bls.gov/cpi/notices/2025/collection-reduction.htm>
 80. U.S. Bureau of Labor Statistics. (2025). *BLS to Discontinue Selected PPIs*. <https://www.bls.gov/ppi/notices/2025/bls-to-discontinue-selected-ppis.htm>
 81. U.S. Department of Education. (2025). *Agency Information Collection Activities; Submission to the Office of Management and Budget for Review and Approval; Comment Request; National Assessment of Educational Progress (NAEP) 2026*. In *Federal Register*. <https://www.federalregister.gov/documents/2025/05/15/2025-08602/agency-information-collection-activities-submission-to-the-office-of-management-and-budget-for>
 82. U.S. Department of Health and Human Services. (2025). *Policy on Adhering to the Text of the Administrative Procedure Act*. Federal Register. <https://www.federalregister.gov/documents/2025/03/03/2025-03300/policy-on-adhering-to-the-text-of-the-administrative-procedure-act>
 83. U.S. General Services Administration. (2024). *Data.gov Program Timeline*. Data.gov. <https://data.gov/timeline/>
 84. U.S. General Services Administration. (2014). *Data.gov CKAN Catalog Launch*. Data.gov Blog. <https://data.gov/announcements/datagov-ckan-catalog/>
 85. U.S. General Services Administration. (2024). *Data.gov User Guide*. Data.gov. <https://data.gov/user-guide/>
 86. United States Congress. (2019). *Foundations for Evidence-Based Policymaking Act of 2018* (No. P.L. 115-435; Issue P.L. 115-435). United States Congress.
-

Glossary

The glossary provides select list of terms and definitions used in this project. When feasible, term definitions were drawn from U.S. Federal Government sources as federal definitions can sometimes deviate from consensus-based definitions in standard sources. Additional general resources for terms related to this project include the [Open Data Handbook Glossary](#) and the [Turing Way Glossary](#). The code for this glossary was adapted from [jekyll-glossary](#) by the author.

[A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#) [N](#) [O](#) [P](#) [R](#) [S](#) [U](#) [V](#) [W](#) [X](#)

A

Access Control

In physical security and information security, access control (AC) is the action of deciding whether a subject should be granted or denied access to an object (for example, a place or a resource). Source: [Wikipedia](#)

Administrative Procedures Act

A federal act that governs the procedures of administrative law, including notice and comment related to formal and informal rulemaking. Source: [Cornell Legal Information Institute](#)

Anonymization

The process of permanently and irreversibly transforming data so that it cannot be linked to any specific individual. Source: [NIST.gov](#)

Application Programming Interface (API)

A predefined protocol for reading and/or writing data using a filesystem, a database, or across a network. Source: [Data.gov](#)

B

Bulk Data

Data that is available for download in its entirety, allowing users to efficiently retrieve the complete dataset. Source: [Open Data Handbook](#)

Bureau of Economic Analysis (BEA)

An agency that produces economic accounts statistics, including the nation's Gross Domestic Product (GDP). Source: [BEA.gov](#)

Bureau of Justice Statistics (BJS)

The principal statistical agency of the Department of Justice, providing information on crime and the justice system. Source: [BJS.gov](#)

Bureau of Labor Statistics (BLS)

The principal federal agency responsible for measuring labor market activity, working conditions, and price changes. Source: [BLS.gov](#)

Bureau of Transportation Statistics (BTS)

The principal statistical agency of the Department of Transportation providing transportation statistics for the nation and its various regions and sectors. Source: [BTS.gov](#)

C

CDO Council

A cross-agency council that coordinates data management policy and establishes government-wide best practices for data use. Source: [CDO.gov](#)

CKAN

An open-source data portal platform used to store, manage, and distribute data assets, powering Data.gov. Source: [CKAN.org](#)

Census Bureau

The principal statistical agency responsible for the decennial census and producing data about the American people and economy. Source: [Census.gov](#)

Chief Data Officer (CDO)

An agency official responsible for data management, governance, and implementing the Evidence Act at a federal agency. Source: [44 U.S.C. § 3520](#)

Comma-Separated Values (CSV)

A common file format for storing tabular data in plain text, where each row is a record and columns are separated by commas. Source: [Open Data Handbook](#)

Comprehensive Data Inventory (CDI)

An inventory of all data assets held by a federal agency, typically stored in a JSON format. An agency comprehensive inventory is usually the source of data used by Data.gov harvesters to populate the Federal Data Catalog. Source: [M-25-05](#) and [44 U.S.C. § 3502](#)

Confidential Information Protection and Statistical Efficiency Act (CIPSEA)

Title III of the Evidence Act providing legal protections for confidential information collected by federal agencies for statistical purposes. Source: [Congress.gov](#)

Controlled Unclassified Information (CUI)

Information that requires safeguarding or dissemination controls pursuant to and consistent with law, regulations, and government-wide policies that is not classified. Source: [Archives.gov](#)

D

DCAT-US

The metadata schema used for the Federal Data Catalog, based on the W3C Data Catalog Vocabulary (DCAT). Source: [Resources.data.gov](#)

Data Asset

A collection of data elements or data sets that may be grouped together. Source: [44 U.S.C. § 3502](#)

Data Catalog Vocabulary (DCAT)

An vocabulary designed to facilitate interoperability between data catalogs published on the Web. Source: [Data.gov](#)

Data Deletion

The process whereby data is removed from active files and storage structures and rendered inaccessible except through specialized data recovery tools. Source: [Society of American Archivists Glossary](#)

Data Dictionary

A data dictionary is a document that outlines the structure, content, and variable definitions for a dataset or collection of data. A data dictionary is a critical tool for reproducibility because it allows others to understand your data. Source: [Harvard](#)

Data Governance

Data governance is the set of principles, policies, and processes that guide the effective and responsible use of data within an organization. Source: [Wikipedia](#)

Data Integrity

The maintenance of, and the assurance of, data accuracy and consistency over its entire life-cycle. Source: [Wikipedia](#)

Data Inventory

A list of data assets and their metadata maintained by an agency to track the information it collects and produces. Source: [44 U.S.C. § 3511](#)

Data Management and Sharing Plan (DMSP)

A document describing how data will be managed, stored, protected, and shared throughout the lifecycle of a research project. Source: [NIH](#)

Data Quality

The fitness for use of data, often measured by its accuracy, completeness, consistency, timeliness, and validity. Source: [FCSM](#)

Data Standard

A technical specification that describes how data should be stored or exchanged for consistent collection and interoperability. Source: [Data.gov](#)

Data.gov

The federal government's federal data catalog indexing public data assets from across all agencies. Source: [Data.gov](#)

De-identification

The process of removing or masking identifying information from a dataset so that the individuals cannot be readily identified. Source: [NIST.gov](https://www.nist.gov)

Differential Privacy

A mathematical framework for sharing information about a dataset while providing strong, quantifiable privacy guarantees for individuals. Source: [Census.gov](https://www.census.gov)

Discontinuation

The termination of an Information Collection Request (ICR), ending the legal authority for an agency to gather specific data from the public under the Paperwork Reduction Act. Source: [Digital.gov](https://www.digital.gov)

E

Economic Research Service (ERS)

One of two principal statistical agencies of the Department of Agriculture that provides economic and social statistical data and analysis for agriculture, food, and the environment. Source: [USDA.gov](https://www.usda.gov)

Energy Information Administration (EIA)

The principal statistical agency of the Department of Energy responsible for collecting and analyzing independent energy information to promote sound policymaking. Source: [EIA.gov](https://www.eia.gov)

Evaluation Officer

An official designated to coordinate evidence-building activities and provide leadership over an agency's evaluation functions. Source: [5 U.S.C. § 313](https://www.5U.S.C.313)

Evidence

Information produced as a result of statistical activities conducted for a statistical purpose, used to inform policymaking. Source: [DOL.gov](https://www.dol.gov)

F

FAIR Principles

Findable, Accessible, Interoperable, and Reusable (FAIR) principles aim to improve data use and reuse. Source: [Go-FAIR.org](https://go-fair.org)

Federal Committee on Statistical Methodology (FCSM)

An interagency committee dedicated to improving the quality of Federal statistics. Source: [StatsPolicy.gov](https://statspolicy.gov)

Federal Data Catalog

The metadata schema used for the Federal Data Catalog, based on the W3C Data Catalog Vocabulary (DCAT). Source: [M-25-05](https://www.m-25-05) and [44 U.S.C. § 3502](https://www.44U.S.C.3502)

Federal Data Strategy

A framework for a consistent approach to federal data stewardship, use, and dissemination across the Executive Branch. Source: [CIO.gov](https://www.cio.gov)

Federal Information Security Modernization Act (FISMA)

A law that requires federal agencies to implement information security programs to protect their data and systems. Source: [CISA.gov](https://www.cisa.gov)

Federal Statistical Research Data Center Program (FSRDC)

A network of 34 data centers across the United States working as a partnership between federal statistical agencies and research institutions that provides secure environments for authorized researchers to access confidential statistical data. Source: [U.S. Census Bureau](https://www.uscensusbureau)

Federal Statistical System (FSS)

The decentralized network of federal agencies that produce official statistics to inform the public and policy makers. Source: [StatsPolicy.gov](https://statspolicy.gov)

Foundations for Evidence-Based Policymaking Act of 2018 (Evidence Act)

A 2018 law (Pub. L. 115–435) that requires federal agencies to modernize data management, increase data availability, and develop evidence to support policymaking. Source: [Congress.gov](https://www.congress.gov)

Freedom of Information Act (FOIA)

A law that provides the public the right to request access to records from any federal agency, subject to certain exemptions. Source: [FOIA.gov](https://www.foia.gov/)

H

Harmonized Tariff Schedule of the United States (HTUS)

A federal data product by the USITC that provides the applicable tariff rates and statistical categories for all merchandise imported into the United States; it is based on the international Harmonized System, the global system of nomenclature that is used to describe most world trade in goods. Source: [usitc.gov](https://www.usitc.gov/)

I

Inter-university Consortium for Political and Social Research (ICPSR)

An American political science and social science research consortium, based at the University of Michigan, ICPSR maintains and provides access to a vast archive of social science data for research and instruction (over 16,000 discrete studies/surveys with more than 70,000 datasets). Source: [Wikipedia](https://www.icpsr.umich.edu/)

Interagency Council on Statistical Policy (ICSP)

A council chaired by the U.S. Chief Statistician that coordinates the federal statistical system and sets government-wide best practices for data. Source: [StatsPolicy.gov](https://www.statspolicy.gov/)

J

JavaScript Object Notation (JSON)

A lightweight, text-based, language-independent data interchange format that is easy for humans to read and machines to parse. Source: [Wikipedia](https://en.wikipedia.org/wiki/JSON)

L

Learning Agenda

A multi-year plan (Agency Evidence-Building Plan) that identifies priority policy questions and the data/methods needed to answer them. Source: [Evaluation.gov](https://www.evaluation.gov/)

Link Rot

The phenomenon of hyperlinks tending over time to cease to point to their originally targeted file, web page, or server due to that resource being relocated to a new address or becoming permanently unavailable. Source: [Wikipedia](https://en.wikipedia.org/wiki/Link_rot)

M

Machine-Readable Format

A format that can be easily processed by a computer without human intervention while ensuring no semantic meaning is lost. Source: [44 U.S.C. § 3502](https://www.govinfo.gov/constitution/44%20U.S.C.%20%26%203502)

Metadata

Data that defines and describes the characteristics of other data. Source: [Wikipedia](https://en.wikipedia.org/wiki/Metadata)

N

National Agricultural Statistics Service (NASS)

One of two principal statistical agencies of the Department of Agriculture responsible for providing timely, accurate, and useful statistics in service to U.S. agriculture. Source: [USDA.gov](https://www.usda.gov/)

National Center for Education Statistics (NCES)

The principal statistical agency of the Department of Education, housed within the Institute of Education Sciences, collecting and analyzing statistical data related to education in the United States. Source: [NCES.ed.gov](https://nces.ed.gov/)

National Center for Health Statistics (NCHS)

The nation's principal health statistics agency, providing data to guide actions and policies to improve American health. Source: [CDC.gov](https://www.cdc.gov)

National Center for Science and Engineering Statistics (NCSES)

The principal statistical agency of the National Science Foundation, providing statistics regarding the U.S. science and engineering, and research and development enterprise. Source: [NSF.gov](https://www.nsf.gov)

O

OMB Circular A-130

OMB policy titled 'Managing Information as a Strategic Resource' which establishes general policy for information governance, data sharing, and privacy. Source: [CIO.gov](https://www.eo.gov)

OPEN Government Data Act (OGDA)

Title II of the Evidence Act which requires federal agencies to publish information online as open data using standardized, machine-readable formats. Source: [Congress.gov](https://www.congress.gov)

Office of Information and Regulatory Affairs (OIRA)

A statutory component of the Office of Management and Budget (OMB) that reviews federal regulations and oversees the implementation of the Paperwork Reduction Act. Source: [National Archives](https://www.archives.gov)

Office of Research, Evaluation, and Statistics (ORES)

The principal statistical agency of the Social Security Administration responsible for statistical data on social security programs and the beneficiaries they serve. Source: [SSA.gov](https://www.ssa.gov)

Open Data Plan

A mandatory annual plan describing an agency's progress in making its public data assets available as open data. Source: [GSA.gov](https://www.gsa.gov) and [M-25-05](https://www.fda.gov)

Open Format

A file format for storing digital data, defined by an openly published specification usually maintained by a standards organization, and which can be used and implemented by anyone. Source: [Wikipedia](https://en.wikipedia.org)

Open Government Data Asset

A public data asset that is machine-readable, available in an open format, based on an open standard, and not encumbered by restrictions that impede use. Source: [44 U.S.C. § 3502](https://www.govinfo.gov)

Open License

A legal guarantee that a data asset is made available at no cost and with no restrictions on copying, publishing, distributing, transmitting, citing, or adapting such asset. Source: [44 U.S.C. § 3502](https://www.govinfo.gov)

Open Source Software

Open-source software (OSS) is computer software that is released under a license in which the copyright holder grants users the rights to use, study, change, and distribute the software and its source code to anyone and for any purpose. Source: [Wikipedia](https://en.wikipedia.org)

P

Paperwork Reduction Act (PRA)

A law governing how federal agencies collect information from the public. Source: [Digital.gov](https://www.digitall.gov)

Personally Identifiable Information (PII)

Information that can be used to distinguish or trace an individual's identity, either alone or when combined with other information. Source: [NIH Privacy Glossary](https://www.nih.gov)

Pregnancy Risk Assessment Monitoring System (PRAMS)

A site-specific population-based surveillance system designed to identify groups of women and infants at high risk for health problems, to monitor changes in health status, and to measure progress towards goals in improving the health of mothers and infants. Source: [cdc.gov](https://www.cdc.gov)

Privacy Act of 1974 (Privacy Act)

A law that establishes a code of fair information practices governing the collection, use, and dissemination of information about individuals. Source: [Justice Department Office of Privacy and Civil Liberties](https://www.justice.gov)

Privacy Impact Assessment (PIA)

An analysis of how information is handled to ensure compliance with privacy requirements and evaluate risks to PII. Source: [OMB Circular A-130](#)

Public Access Removal

A disruption in federal data availability where proactive disclosure or dissemination is halted, often resulting in datasets or tools being withdrawn from public-facing portals like Data.gov. Source: [Congressional Research Service \(CRS\) Report R48889](#)

Public Data Asset

A data asset maintained by the Federal Government that has been, or may be, released to the public. Source: [44 U.S.C. § 3502](#) and [M-25-05](#)

R

Reproducible Research

Reproducible research is work that can be independently recreated from the same data and the same code that the original team used. Source: [The Turing Way](#)

S

Schema

A data model or database structure that defines the relationships between different pieces of information. Source: [Resources.data.gov](#)

Standard Application Process (SAP)

The centralized portal and process required by the Evidence Act for researchers to apply for access to confidential statistical data from federal statistical agencies, often through an FSRDC. Source: [ResearchDataGov](#)

Statistical Official

A designated agency official with expertise in statistics who advises on statistical policy, techniques, and procedures. Source: [5 U.S.C. § 314](#)

Statistical Purpose

The description, estimation, or analysis of the characteristics of groups without identifying the individuals or organizations in those groups. Source: [44 U.S.C. § 3561](#)

Statistics of Income Division (SOI)

The principal statistical agency of the Internal Revenue Service that compiles and publishes statistical data on the operation of the U.S. tax system. Source: [IRS.gov](#)

Synthetic Data

Information that is generated by a computer model that mimics the statistical properties of a real-world dataset but contains no real records. Source: [Census.gov](#)

System of Records Notice (SORN)

A public notice required by the Privacy Act that informs the public of the existence and character of a system of records. Source: [GSA](#)

U

United States Agency for International Development (USAID)

The USAID is a *de jure* agency of the executive branch of the United States federal government that, until its effective shuttering by the Trump Administration in 2025, served as the world's largest funder of direct foreign assistance. Source: [Wikipedia](#)

United States Chief Statistician

A position within OMB responsible for coordinating the federal statistical system, US official international statistical activities, and setting government-wide statistical standards. Source: [StatsPolicy.gov](#)

United States International Trade Commission (USITC)

The USITC is an independent, nonpartisan, quasi-judicial federal agency that fulfills a range of trade-related mandates. The agency provides high-quality, leading-edge analysis of international trade issues to the President and the Congress. The USITC produces the HTUS dataset. Source: [usitc.gov](#)

V

Vocabulary

A set of standardized terms with consistent semantic definitions, typically constrained to a particular namespace or domain. Source: [Resources.data.gov](https://resources.data.gov)

W

Wayback Machine (WBM)

A digital archive of the World Wide Web provided by the Internet Archive that captures and preserves snapshots of websites to prevent the loss of information when pages are changed or removed. Source: [Internet Archive](https://archive.org)

X

XML (eXtensible Markup Language)

A markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable. Source: [W3C](https://www.w3.org)

[↑ Back to Top](#)
