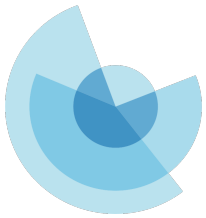
CODE

SAVING THE NATION'S DATA: A PROPOSED PROGRAM FOR ACTION

APRIL 2026



CODE

SAVING THE NATION'S DATA: A PROPOSED PROGRAM FOR ACTION

EXECUTIVE SUMMARY

The United States is at a critical juncture in the management and preservation of its national data assets. As the digital backbone of the nation's economy, research, and governance, these datasets are a public resource that requires both immediate protection and long-term strategic investment. We need a thorough, scalable, and repeatable process to ensure our national data ecosystem continues to serve America's needs.

In this white paper, the Center for Open Data Enterprise (CODE) proposes a strategic cycle of assessment, action, and institutional reform. Our proposal is based on the work of many outstanding organizations that are already taking action effectively. We recommend building on their efforts to create a coordinated, collaborative program that follows 10 steps like these:

1. **Track what we've lost.** We need a thorough, AI-enabled scan of the federal data ecosystem to see what's already been lost or changed, and set up automated monitoring to detect even subtle changes going forward.

2. **Build coalitions in key domains.** Public health experts know which datasets matter most to disease surveillance. Climate scientists know which environmental indicators are irreplaceable. Education researchers know which federal surveys track opportunity. These experts must work alongside data scientists, AI specialists, and philanthropic partners to map what truly counts.

3. **Prioritize core datasets.** Through interviews, surveys, and quantitative analysis—such as tracking citations in research or journalism—coalitions can identify a "core canon" of essential datasets in each field.

4. **Assess the risks.** Tools like the Data Checkup, developed by dataindex.us, can assess threats to federal datasets. This work can be automated and scaled with AI.

5. **Determine the federal role.** Some federal data—like satellite observations, national health surveillance, or economic indicators—cannot be replicated by states or private actors. Other data can be supplemented or replaced by state and local sources, private-sector datasets, crowdsourcing, or nontraditional data sources.

6. **Take action to save essential data.** When federal data is essential, coalitions can pursue advocacy, public comments, direct engagement with agencies, or litigation. When alternatives exist, they can be developed, benchmarked, and scaled.

7. **Put the data to work.** The best way to defend data is to use it. Publishing use cases, visualizations, tools, and plain-language insights helps the public see why this information matters. Generative AI can make federal and open data accessible to millions of non-technical users.

8. **Think globally.** The threats to data go beyond the U.S. We need to track the international impacts of U.S. data loss, study how international sources might replace U.S. data, and share lessons learned with other countries.

9. **Strengthen institutional protections.** In addition to managing today's immediate problems, we need to develop policies, laws, governance strategies, and guardrails for more stable, reliable data in the future.

10. **Sustain the cycle.** The threats will evolve. So must the response.

This white paper is an invitation to discussion, debate, and further improvement. We hope it will help encourage even greater collaboration and coordination among the many organizations working to protect and improve America's data. Working together, we can help ensure that our national data ecosystem becomes even more robust, resilient, and relevant.

INTRODUCTION

In March 2026, in a converted Christian Science church in San Francisco, more than a hundred experts from across the country gathered to discuss how to protect America's data and information. The nonprofit Internet Archive, whose mission is to provide "Universal Access to All Knowledge," convened this Information Stewardship Forum at its headquarters to find new ways to protect critical data from unprecedented risks.

The Forum gathered academics, librarians, policy experts, and representatives of both well-established and newly formed nonprofits – all part of a growing *data rescue movement*. Meeting in old church pews and a former Sunday school basement, the group brainstormed how to use information technology, collaboration, and voluntarism to save the information and common knowledge essential to a free society. This effort has roots in the broader open data movement, which began about two decades ago to promote freely available government data. The new data rescue movement began almost immediately after the 2024 election, spurred by the fear that the second Trump Administration would suppress or change essential national data for political reasons.

The Center for Open Data Enterprise (CODE) – whose mission is to harness the power of open and shared data for the public good – has been tracking disruptions to public data since January 2025. In reports on America's data future and the state of public health data in particular, we have analyzed changes this administration has made to data collections, analytic tools, and the operations of data-collecting agencies. Many appear politically motivated, such as reducing access to environmental and climate data, data on LGBTQ+ individuals, and health data in many areas.

A number of new and existing organizations, tracked comprehensively by the Data Rescue Project, have responded by downloading, archiving, and preserving essential data and tools. These heroic, highly effective efforts have involved thousands of volunteers in essential ongoing work. But they are limited by a basic fact: You can only rescue data that already exists. *The emerging challenge is how to ensure that America's national data ecosystem remains robust and continues to serve the country into the future.*

This is more than a technical challenge: It is a core challenge for America's democracy. Americans across the political spectrum have largely lost trust in our national government and national institutions. We can only restore that trust with a shared, widely accepted understanding of facts and information about the issues

that most impact the country. And we can only reach that understanding with the support of a healthy, durable system of unbiased national data.

UNDERSTANDING THE CHALLENGE: NATIONAL DATA AT RISK

The threats to American data go beyond the federal data and tools that have already been discontinued and others that are at immediate risk. The Trump Administration's actions have shown that our national data systems are much more fragile than anyone realized. In a little over a year, we have seen that:

- Federal agencies have **discontinued** or **dramatically changed** data collections simply by getting the approval of the White House Office of Management and Budget (OMB). While OMB is required to seek public comment in most cases, the office is not required to follow public opinion.
- Agencies **selectively changed** the ways they collect data, for example, by dropping data on LGBTQ+ individuals.
- Agencies **fired the teams** responsible for analyzing data collections, such as some key public health data programs, making it difficult or impossible to use the data even if it is still collected.
- Agencies **removed data and tools from their websites**, making that information effectively unavailable to the public.
- Agencies summarily **disbanded federal advisory committees** responsible for reviewing data programs.
- The President **fired the leadership** of the Bureau of Labor Statistics at will, and could presumably do the same for other national statistical offices, although replacements must be approved by Congress.
- National scientific agencies, notably the National Science Foundation (NSF) and National Institutes of Health (NIH), **dramatically revised** their criteria for funding and gutted scientific projects that had been running for years. These scientific projects have become an increasingly important source of public data for the nation.
- The administration **eliminated major data collections** in targeted areas such as environmental research and proposed draconian funding cuts to data-collecting agencies. Although research budgets are ultimately set by Congress, even the attempt to slash funding creates a climate of uncertainty.

A wide range of stakeholders have opposed these changes with some success. Early on, physicians' groups sued successfully to have CDC data and information restored, and a lawsuit from organic farmers led the U.S. Department of Agriculture (USDA) to republish climate data that is essential to their livelihood. Advocacy campaigns by nonprofit groups, notably the new organization dataindex.us, have generated public comments supporting important national data collections. Congress rejected proposed cuts to research in the 2026 President's Budget proposal (although the 2027 President's Budget proposes similar cuts).

Despite these efforts, however, the administration's actions have fundamentally revealed the weaknesses in our data infrastructure. Since the mid-2000s, members of the open data movement have assumed that government data, and federal data in particular, was a massive source of insight just ready to be tapped. The challenge, it seemed, was to make the data even better, ensure that it is easy to access and use, help data providers and data users collaborate more effectively, and create new tools for analyzing the data and

publishing the results. Now, it's suddenly clear that many of those valuable data sources could disappear overnight.

CODE has been studying this problem for more than a year, together with the many outstanding organizations that have been working to protect government data in a wide variety of ways. In 2025, CODE partnered with the nonprofit USAFacts to hold two Roundtables on "Building the National Data Ecosystem America Deserves" with more than 150 expert participants. We also worked with the Robert Wood Johnson Foundation (RWJF) and the National Conference on Citizenship to hold a roundtable and analyze threats to public health data - a project we are building on in 2026.

As a service to the public data community, we have created an online hub with information about the organizations that participated in CODE's roundtables: https://bit.ly/NationalDataResources. We hope this will be a useful resource for collaboration, and welcome suggestions for additional organizations doing this work or other changes to the hub (see the "About" section for contact information). At the same time, we are continuing to track progress in the field and exploring CODE's own role in maintaining America's national data ecosystem moving forward.

Our convenings and related research have included a range of organizations addressing the threats to data in real time while enabling new and improved uses of public data. We believe it's now possible to develop an integrated, collaborative, and scalable program to meet these goals by bringing different lines of work together and building on them. This program could combine expert collaboration, a decision tree, and rapid action; utilize both human expertise and AI-enabled analysis; and be driven by alliances and communities of practice between subject-matter specialists, data advocates, technical experts, and philanthropic organizations. The following is a high-level draft proposal for such a program, intended for discussion, development, and collaboration.

A PROGRAM TO SAVE ESSENTIAL DATA AND FIND NEW ALTERNATIVES

No organization, program, or plan could be designed to save every current federal data collection. The centralized website data.gov now provides access to over 400,000 datasets from all over the federal government. While it's possible to archive these datasets in their current form, as Harvard's Library Innovation Lab has recently done, there is no way to monitor, assess, evaluate, and ensure resources to protect all of them. And even if that were possible, it might not be the best strategy, since these data sources vary widely in their quality, utility, and importance.

A different approach is to develop a methodology to identify the most valuable federal datasets, determine which ones are at risk, and take action accordingly. This methodology can act as a funnel to focus action on the most important data sources in any given domain. CODE is already testing aspects of this approach in our work on public health and exploring further domain-based partnerships with education and environmental organizations. We believe a complete methodology could include steps like those shown in the infographic below and spelled out in detail beginning on the next page.

1. **Track what we have lost.** Several programs are now monitoring changes to federal data in different ways, but some subtle or hidden changes may still go unnoticed. The field would benefit from an AI-enabled approach to scan federal websites systematically and compare them to archived data as a benchmark. Harvard's replication of data.gov could be one valuable benchmark for this effort.

2. **Build coalitions in key domains.** Implementing this program will require close, ongoing collaboration between people and organizations with diverse perspectives and expertise. Some will have deep policy and practical experience in specific areas of public concern, while others will have broad expertise in data policy, monitoring, and assessment. CODE, for example, is now assembling working groups with a wide range of health data experts, policy advocates, and science communicators to advance public health data. Stakeholder groups can form the basis of coalitions and communities of practice including the following.

    a. **Domain experts.** With some exceptions like Census and economic data, most federal data collections have high value to experts in specific domains rather than the public at large.

Coalitions can tap subject-matter experts to identify the most important data sources for their work, using surveys, interviews, and convenings. Initial domains could include:

- Public health
- Climate and the environment
- Education and workforce opportunity
- Economic development
- Public safety and criminal justice
- Scientific research

b. **Data, technical, and AI experts.** These individuals and their organizations have knowledge that can be applied across all domains, and that is critical to preserving and improving federal data. They include experts in data science and statistics, research methodology, data policy, data privacy and security, and related fields. Experts in AI can also help automate and scale methodologies to assess hundreds or thousands of datasets rather than doing that work manually.

c. **Philanthropic institutions.** A number of foundations, and some new funder coalitions, have started to fund projects to protect national data resources. Many others have shown interest but have been challenged by the scope of America's data landscape. A clear program with a scalable methodology could give them a path for highly leveraged, high-impact investment.

3. **Prioritize core datasets.** By taking a domain-by-domain approach, subject matter experts can work with data experts to identify the most valuable federal data resources. CODE has piloted this approach in our work with RWJF. Through interviews, surveys, and an expert roundtable, we identified about 70 major data sources that public health experts widely agree are core to their work. A similar process could be a starting point for analyzing data in other domains. After an assessment by subject matter experts, data analysts could use quantitative measures to rate federal datasets for their relative importance – for example, by analyzing the use of datasets in scientific publications, in news reports, or in other ways. Combining human expertise and technical analysis with the help of AI can develop a "core canon" of essential data for each domain and refine it over time.

4. **Assess the risks.** The next step is triage: Assessing the risk to each dataset in the core canon for any given sector. The organization dataindex.us has developed the Data Checkup, which combines several different factors to assess the threats to any federal dataset. While this kind of work now requires human research and analysis, it should be possible to automate and apply at a large scale. The Impact Project and other organizations are also tracking cuts to public funding that may endanger federal data indirectly.

5. **Determine the federal role.** Up to this point, this methodology could produce a selective list of highly important, at-risk data sources for each domain. The next step is to determine whether these data sources can only be provided by the federal government or whether there are realistic alternatives to explore. This can include considering the expense of data collection, the methods of collecting data (satellites are harder to deploy than surveys), geographical scope and granularity, and

other factors. Determining the need for federal leadership – through a consistent, replicable methodology that will need to be developed – can then guide the choices for action.

6. **Take action to save essential data.** At a high level, there are three paths at this point.

    a. **Holding the government accountable: Persuasion, advocacy, and litigation.** When the federal government is the only realistic source of essential data, the stakeholders who rely on that data can push the government to play this critical role. Stakeholders can pursue a wide range of options, including meetings and discussions with key decisionmakers, broad-based advocacy through open letters or the public comment process, or litigation, which requires an analysis of administrative law, authorizing language for data collections, and the possible harms of data loss.

    Insisting on federal data collection does not have to mean continuing to collect data in exactly the same way. For years, for example, responses to the U.S. Census and other surveys have been declining, and statisticians have been exploring the use of administrative data as a possible alternative. But the kind of administrative data that would be needed – such as tax records, birth and death records, and other personal information – would still need to come from the government.

    b. **Developing bridging solutions.** In some cases, philanthropies or other organizations may step in to support data programs that the federal government has abandoned. For example, the Robert Wood Johnson Foundation has a program to fund researchers who have lost their NIH grants, and the Gates Foundation is funding the international Demographic and Health Survey program that was formerly under USAID. But these are only temporary measures, and depend on the hope that a future administration will restore federal support.

    c. **Finding alternative data: Tapping other levels of government and sources.** In many cases, new kinds of data sources may be able to replace or even improve on the federal data that has been used in the past. These can include state and local data analyzed in new ways, crowdsourced data, social media data, data from the private sector, or "nontraditional" data sources of other kinds. A future national data ecosystem should include a diversity of data types, and ultimately become less reliant on federal data than it has been in the past.

7. **Put the data to work.** The same coalitions that work to preserve data can also promote its public use. This can include publishing use cases as America's Essential Data is doing; publishing articles and visualizations that turn data into insights, as USAFacts does; convening working groups to coordinate on specific projects as well as broader efforts; and developing new tools for data access and application. These efforts are important in their own right, and also support data preservation by giving more people first-hand experience with government data's value.

    Generative AI can enable non-technical users to access, analyze, and publish findings from federal and other open data, although it is still far from perfect at this. CODE is now experimenting with AI-enabled tools to make data easy to use through plain-language prompts.

8. **Think globally.** The U.S. data ecosystem operates in a global context. While monitoring federal actions is the primary focus, we should also develop a global vision from several perspectives. This includes tracking the impacts of U.S. data loss on global knowledge, as discussed here; looking for international data sources that may be able to replace U.S. data, for example in earth observation or global health; and sharing and refining lessons learned with other countries that have worked on data governance or face similar challenges to data resilience.

9. **Strengthen institutional protections.** We have learned that the U.S. national data ecosystem is much more fragile than anyone may have realized. In addition to managing the disruptions we're facing today, we need to develop laws, governance strategies, and guardrails that will give us more stable, reliable data in the future. This is a long-term project, but now is the time to start.

10. **Sustain the cycle.** The state of federal data will be fluid for years to come. Any program for progress needs to be iterative, with coalitions working continually to analyze, build, and improve the state of the national data ecosystem. We will need to build new tools, communication channels, and information hubs to track progress and support innovation.

THE LONG-TERM GOAL: NEW RULES AND NEW SAFEGUARDS

We believe this ten-point program, or a version of it developed through collaboration, can protect and improve America's data in the face of immediate challenges. But in the long run we need to do more. We need new rules and new safeguards for a national data ecosystem that balances continuity and stability with the need for constant evaluation and thoughtful change.

Ultimately, we need new legislation and new laws to govern America's data infrastructure. There is a long bipartisan tradition of Congressional support for robust federal data. Most notably, the Foundations for Evidence-Based Policymaking Act (the Evidence Act), passed almost unanimously by Congress and signed into law by President Trump in 2019, established rules for open data publication, statistical agencies' independence, and other elements of federal data infrastructure.

While the Evidence Act was landmark legislation, we need stronger legislation to protect America's data from political manipulation, defunding, or other interference. But at the same time, we need a data system that can evolve in positive ways. America's national data was far from perfect before the changes that began in 2025, and simply returning to the status quo, even if that was possible, would not be the best option. To develop a structure for data that balances continuity and change, for example, Congress could:

- Amend the Evidence Act to strengthen the protections it provides
- Write more precise authorizations for the most important, highest-priority datasets
- Set up multiyear funding for core data collections to remove year-to-year uncertainty
- Update national archiving requirements for the digital age
- Create clear guidelines for making government data AI-ready
- Set clear, enforceable criteria to ensure accurate scientific data is made public
- Define measurable criteria for sunsetting data collections that are no longer needed
- Mandate phase-out processes for changing or replacing longstanding data collections, including benchmarking against current data collections to ensure comparability and continuity

There's a good chance that Congress could reach a bipartisan consensus to strengthen our data infrastructure in these or other ways. While some policies could be enacted quickly, broader legislation will take years to develop and pass and even longer to have an impact. (Seven years after becoming law, the Evidence Act is still not fully implemented.) But it's worth beginning to work on policy and legislation now for the long-term solutions the country needs.

CODE's proposals for both immediate and long-term action are just a start - a first-draft, high-level proposal for the work ahead. We hope these proposals help give everyone concerned about America's data a path to collaborate, focus our efforts, and use a shared, scalable methodology to be efficient and effective. We look forward to discussing and implementing these and other ideas as we envision the future together.

This white paper is published by the Center for Open Data Enterprise (CODE), a nonprofit organization whose mission is to harness the power of open and shared data for the public good. It was written by Joel Gurin, CODE's president and founder, as part of his Fellows for Democracy and Public Service Initiative, supported by the Bridge Alliance and the National Academy of Public Administration. Throughout 2026, his initiative will address the topic, "Trust in Information Leads to Trust in Government," and explore ways to restore that trust.

Readers are invited to share this paper at https://bit.ly/NationsDataProgram and to contact Joel Gurin about this work at joel@odenterprise.org. More information about CODE can be found at www.opendataenterprise.org.

CODE

The Center for Open Data Enterprise (CODE), a 501(c)3 nonprofit organization based in Washington, DC, was founded in January 2015. CODE's mission is to harness the power of open and shared data for the public good. We achieve our mission by working with government agencies, businesses, nonprofits, and researchers who are both data providers and data users.

We support the application of fully open data – free, publicly available data that anyone can access and use, without limitation – as well as strategies for sharing and exchanging data that requires privacy or security restrictions. CODE welcomes ideas and opportunities for collaboration at contact@odenterprise.org.