Supporting the Identification, Monitoring and Preservation of Government Data Resources:

Findings from DataLumos Outreach Efforts

David A. Bleckley, Susan M. Jekielek, Ph.D., J. Trent Alexander, Ph.D., and Bianca Monzon

Inter-university Consortium for Political and Social Research
Institute for Social Research
University of Michigan

This research was funded by the Annie E. Casey Foundation. We thank them for their support but acknowledge that the findings and conclusions presented in this report are those of the authors alone, and do not necessarily reflect the opinions of the Foundation.

Table of Contents

Executive Summary

This report documents the findings of "Identification, Monitoring, and Preservation of Government Data Resources", an 18-month project involving outreach to government data producers, users, and intermediaries. Through this project, the Inter-university Consortium for Political and Social Research (ICPSR) sought to identify stakeholders' most-used government datasets that they perceive to be potentially less accessible in the future, among other goals. Interviews and less formal interactions with data advocates and intermediaries, government data producers, and a variety of data users provided insights into the use of government data and perceptions of these data's future accessibility.

The most important source of data to these stakeholders is the Census Bureau, and several of its products were identified as being critical to stakeholders' work. Data from other major statistical agencies, non-statistical federal agencies, and state and local data sources were also cited. The federal government data most used by stakeholders—and specifically the data of greatest importance to AECF-funded work—are perceived as accessible for future use. All of the federal datasets that stakeholders perceived to be potentially at risk were assessed and added to the DataLumos archive.

A noteworthy finding from these interactions is that data created or collected by KIDS COUNT grantees, National Neighborhood Indicators Partnership (NNIP) participants, and other data intermediaries may not have a long-term data archiving or sharing plan. The analysts at these organizations spend significant effort gathering, aggregating, and analyzing data for their products, but they generally have no mechanism to archive or share these data. Given the investment in this work and the potential value of these data to community organizations, researchers, and even local and regional government agencies, there is a real opportunity for data intermediaries to store and share these data in a secure manner for the long term.

Recommendations based on the project's findings can be grouped into two major categories: advocacy and data sharing. Data users, intermediaries, and funders should continue to advocate that the Census Bureau and other principal statistical agencies provide access to the data products needed to successfully complete their work. Advocacy is also needed at the state and local levels, with the goals of targeting the creation of transparency laws and sunshine clauses, budget line items for data sharing, and infrastructural investments like open data portals and data application programming interfaces (APIs). Beyond traditional advocacy work, sustained and increased collaboration between government data producers and data users, intermediaries, and advocates is needed. As for data sharing, we recommend that data creators and intermediaries like KIDS COUNT grantees and NNIP partners work with data repositories like ICPSR to make their data available to others now and in the future. The archiving of these data would require both the infrastructure of a secure data repository as well as specialized curation and technical assistance related to sharing these types of data. The creation of an archive for data intermediaries' data would extend the value of intermediaries' important work, creating new resources for community members, institutions, and researchers.

Introduction

From September 2017 through February 2019, the Inter-university Consortium for Political and Social Research (ICPSR) reached out to a wide variety of government data producers, users, and intermediaries. The primary goal of these interactions was to identify stakeholders' most-used government datasets that they perceive to be potentially less accessible in the future. Secondary goals were to:

- better understand a diversity of stakeholders' interactions with government data,
- clarify workflows of government data users and producers,
- identify respondents' interest areas, and
- identify and develop connections between the respondents' work and DataLumos/ICPSR.

While the main emphasis of these efforts has been to identify data with perceived decreased future accessibility—efforts which have been documented in the grant's narrative reports—the conversations surrounding this work have been interesting, rich, and informative. The following whitepaper seeks to document these findings.

Background

ICPSR has a long commitment to safekeeping and disseminating U.S. government and other social science data. On February 16, 2017, ICPSR launched DataLumos (www.datalumos.org), an open-access archive where the public can archive valuable government data resources, ensuring their long-term availability. Later that year, ICPSR worked with the Annie E. Casey Foundation (AECF) toward a project called "Identification, Monitoring, and Preservation of Government Data Resources." This project sought to expand ICPSR's capacity to identify, monitor, and archive government data resources that are of key concern to AECF—specifically resources whose continued accessibility or discoverability are of concern. Our core objective in this project was to identify important data sources that are required to maintain and enhance the products and services that Foundation partners provide to the research community. The project supported outreach and engagement between ICPSR project staff and individuals across a number of data communities who possess insights and knowledge about the future accessibility of these data. Project outreach focused on connecting with federal data producers, liaising with civil servants, and developing relationships with data advocates and data intermediaries who use federal data to inform policy and practice.

Methods

We used qualitative approaches to learn how various groups of stakeholders interact with government data. These approaches included two main types of interactions: (1) semi-structured interviews and (2) less formal discussions. The interviews were used when ICPSR could schedule time with the respondent ahead of time. Informal discussions occurred when opportunities arose more organically, for example, during a conference or other meeting.

The interviews were guided by a set of core questions (see Appendix). The core questions were modified and customized according to the field and role of each respondent. Prior to conducting the interviews, we collected background information about the interviewees and their work from institution/project websites and reports. This background preparation helped to both inform interviews and facilitate the building of rapport with our participants.

We conducted these interviews mainly through conference calls and occasionally in-person. When possible, two ICPSR staff members would conduct the interviews. We also encouraged interviewees to include two or more team members in the discussion, bringing a greater diversity of perspectives to the conversation and allowing interviewees to build on each other's responses. The less formal discussions were based on similar goals and questions as the interviews but without the benefit of the pre-interaction research and preparation due to the generally spontaneous nature of these discussions. After each interview or discussion, ICPSR staff typed and collated notes. Once all interactions were complete, we synthesized all notes and analyzed for common emerging themes. Those themes are presented below, in the Findings section.

Overview of Participants/Respondents

Throughout this project, ICPSR spoke with 81 individuals from widely varied backgrounds, including 30 government data producers, 24 data intermediaries/users, and 27 data advocates. Two of the main groups of respondents were identified specifically in the grant application—KIDS COUNT Grantees and National Neighborhood Indicators Partnership (NNIP) Partners—and we worked closely with AECF and NNIP to make initial contact with these individuals. There was a great deal of diversity even within these two groups with data users, intermediaries, and producers from academic, government, and nonprofit institutions. We also spoke with data producers employed by multiple agencies within the federal government, from major statistical agencies as well as non-statistical agencies. Finally, we spoke with data advocates and intermediaries, including librarians, archivists, and others working to ensure government data are accessible well into the future.

Findings

The overarching goals of this project were to identify what government data these stakeholders use and to determine whether they are concerned about the future accessibility of these data. We found that the most-used data are products of the Census Bureau. The vast majority of data users and data intermediaries mentioned using Census products, including the decennial census and the Current Population Survey, but the greatest emphasis was on the American Community Survey and its public-use microdata sample. These data are important to stakeholders interested in understanding community-level issues because they capture a wide variety of social and economic aspects of American life, are released annually, and utilize relatively small geographic regions. Beyond data from the Census Bureau, stakeholders discussed the importance of other major statistical agencies' data, especially the Bureau of Labor Statistics and the Bureau of Economic Analysis.

Several individuals also cited the importance of data from non-statistical agencies in their work. The most commonly referenced departments, agencies, and institutions were:

- Department of Education,
- Department of Health and Human Services,
- Department of Housing and Urban Development,
- Federal Financial Institutions Examination Council (Home Mortgage Disclosure Act), and
- Internal Revenue Service.

While many different data sources and datasets were identified through this process, respondents named very few datasets that they perceived to be less accessible in the future. This was a key finding and bears repeating: **the federal government data most used by stakeholders—and specifically the data of greatest importance to AECF-funded work—are considered to be accessible for future use.** Data producers and users alike perceive those data to be accessible in the future. All of the federal datasets stakeholders perceived to be potentially at risk were assessed and added to the DataLumos archive (either by the stakeholders themselves or by ICPSR). These data include:

- Centers for Disease Control and Prevention - Social Vulnerability Index,
- Department of Education - Civil Rights Data Collections,
- Department of Housing and Urban Development - Affirmatively Furthering Fair Housing, and
- Department of Labor - Trade Adjustment Assistance Cases.

Beyond accessibility of existing data, respondents cited concerns about future data. Among the prospective perceived problems were transparency, accessibility, and data quality. Some stakeholders expressed concern that some data collection efforts may not be continued in the future due to shifts toward deregulation and decreased government oversight. Several stakeholders also worried that government priorities may be shifting away from releasing data

and making these releases easily discoverable. Somewhat related to that, a few respondents noted that some data previously released as tabular data files are now released as reports with embedded summary statistics or tables. This method of data release makes these data less useful and accessible, whether intentionally or unintentionally. It also surfaces another long-term preservation issue, as not all government documents are disseminated in the same manner or archived permanently; issues like these are being explored by initiatives like the Preservation of Electronic Government Information (PEGI) Project. Finally, a major concern cited by many stakeholders is the future quality of data, including the lack of funding for data creation and dissemination. Specifically, issues with the funding and leadership of the Census Bureau as well as the inclusion of a citizenship question in the 2020 Census lead many respondents to fear a decrease in this decennial census's quality. There is a related concern that local funding, community programming and research will be negatively impacted by poor quality data for the coming decade.

The federal government is a key source of data for the stakeholders involved in this project, but many also use data from state government agencies. While respondents had few concerns about specific datasets' being less accessible in the future, several different issues were raised related to state government data access, including funding, staffing, and informality. Stakeholders reported that budget cuts over the past decade have limited the number of datasets agencies across the country are releasing as a matter of practice, as governments focus on activities deemed internally as high priorities than open data dissemination. Staffing is also partially a financial issue, because budget cuts generally result in staffing cuts, resulting in fewer employees available to respond to with data requests. The staffing issues identified by stakeholders, however, extend beyond layoffs. A few respondents described interactions with government employees who lacked sufficient knowledge about data sharing (and data-related laws such as FERPA) to provide adequate data to data users or intermediaries. In one case, the government staff member stated that no education data could be shared because of FERPA. This leads to the last state government data issue: a lack of formal, standardized means of distributing government data. Access to data is often based on relationships between government workers and data requesters. While these personal connections may facilitate the initiation of public access to data, they are limiting if they are not codified into policies and procedures (who has access to data? when? through what means?). A key employee's retirement may mean the end of public data access in these situations. Stakeholders also provided mixed comments about trends in state data availability. On one hand, the lower costs of data storage and data portal creation, along with citizen demands for data access, have led to the widespread emergence of state government data sharing. Conversely, a few stakeholders noted that some states have seen governors rise to power and create new barriers to data sharing.

Another set of issues noted by data producers—both government employees as well a data intermediaries who compile and aggregate secondary data—is related to paradata. Paradata are data about the process of data production. Examples may be related to execution of a survey collection, how geographies are defined, or other administrative data. Data producers noted that even when a dataset is publicly available, the paradata may not be released or even documented.

One respondent noted finding a database of administrative data but did not know if there was a plan for archiving them or if they were included in a records disposition schedule.

This leads to our final finding—one that, while only tangentially related to the government data project, we found noteworthy. Data created or collected by KIDS COUNT grantees, NNIP Partners, and other data intermediaries may not have a long-term data archiving or sharing plan. The analysts at these organizations spend significant effort gathering, aggregating, and analyzing data for their products, but they generally have no mechanism to archive or share these data. One stakeholder team told ICPSR that they are running out of server space, and one solution they are considering to remedy this would be to delete data older than ten years. Given the investment in this work and the potential value of these data to community organizations, researchers, and even local and regional government agencies, there is a real opportunity for data intermediaries to store and share their data in a secure manner for the long term.

Recommendations

Based on the findings discussed above, ICPSR would like to provide several recommendations. They can all be grouped into two major categories: advocacy and data sharing.

Advocacy

One of our main takeaways from these interactions was that respondents perceived data from federal statistic agencies to be secure in their future accessibility. What stakeholders questioned was the funding, transparency, and quality of future data creation and dissemination. We recommend that data users, intermediaries, and funders continue to advocate that the Census Bureau and other principal statistical agencies provide access to the data products needed to successfully complete their work. This will require advocacy for government investment in the implementation and infrastructure required for data collection, processing, analysis, and dissemination.

Another key finding was that some important data are only available in government reports as tables or summary statistics. Stakeholders reported that this can create issues for discoverability and access. Respondents also noted that the U.S. Government Publishing Office's transition from print dissemination to largely electronic dissemination has created some issues of its own. Some stakeholders were unclear of the ongoing role of Federal Depository Libraries, including what gets stored, how it gets stored, and how to maximize electronic documents' discoverability, searchability, and utility. Other respondents pointed out there can be gray areas about what documents need to be released in the first place or how widely they need to be disseminated. All of this points to a need to advocate for:

- public release of stand-alone data files (in addition to reports) wherever possible;

- clarity on what documents get released; and
- procedures and infrastructure to support the discoverability and long-term access to those documents.

The findings also point to the need for advocacy at the state and local levels. Data users, intermediaries, and funders must continue to inform government officials (both lawmakers and agency staff) of the importance of government data access for community programming, planning, and research. One potential goal for this type of advocacy would be the creation of transparency laws and sunshine clauses at multiple levels of government, partnered with budget line items for data sharing. Based on stakeholder feedback, governments should move toward infrastructural investments like open data portals and data application programming interfaces (APIs).

Beyond traditional advocacy work, our observations from the past 18 months of outreach and interactions also suggest the need for sustained and increased collaboration between government data producers and data users, intermediaries, and advocates. The data producers with whom ICPSR spoke as well as many of the producers described in stakeholders' anecdotes are deeply dedicated to the creation, dissemination, and analysis of government data. Some of these producers were offended by the data rescue movement because of the lack of outreach to data producers and the underlying assumption that data producers would ignore the politicized destruction of data or data dissemination channels. Some of the stakeholders and organizations with which ICPSR worked during this project—notably the PEGI Project—are doing exemplary work in collaborating with government workers to ensure continued or improved accessibility.

Data Sharing

All these interactions with stakeholders—especially KIDS COUNT grantees and NNIP partners—have made it clear that data intermediaries and data users who analyze and aggregate data for special regions and populations are creating really valuable datasets and resources. These data, however, are often created for a report or project, and that may be the extent to which the data are used or disseminated. ICPSR sees great potential in the curation, archiving, and sharing of these types of data, allowing communities, organizations, and researchers access to all the value these intermediaries are adding to the data. We, therefore, recommend that data creators and intermediaries like KIDS COUNT grantees and NNIP partners work with data repositories like ICPSR to make their data available to others now and in the future.

Based on discussions with data intermediaries, we would recommend an archive infrastructure with a variety of data security measures. The system would provide secure upload and download capabilities with varied permissions, allowing a variety of users access to data ranging from public-use to more sensitive restricted-use. Another potential function of a secure repository project would be the provision of an interface for data producers to supply confidential data to data users and intermediaries with a data use agreement. This secure system would ensure that the data remain at the repository while users access the data remotely for analysis. When

analyses are complete, the aggregate data and statistical output could be released to the user as well as archived with the repository, while the sensitive raw data files are deleted.

The archiving of these data would require not only the infrastructure of a secure data repository but also specialized curation and technical assistance related to sharing these types of data. The most critical support would be training data intermediaries on best practices in data sharing. An important facet of that would be methods, examples, and assistance in incorporating data sharing into data use agreements with local and state government. Data use agreements often include stringent confidentiality and destruction clauses which preclude external data sharing. With some revision, these agreements can also include specifications for redistribution and requirements for aggregation (minimum population size or geographic area). One way these data could be curated would be to harmonize similar datasets into a collection, allowing for interregional or interstate comparison. Some data intermediaries noted that they do not necessarily want to be forced into aligning definitions of all indicators with their peers, but most noted the value of some manner of data harmonization and sharing. Creating a centralized repository for data intermediaries would also provide an opportunity for these stakeholder groups to learn from each other to promote best practices in data management, curation, sharing, and archiving.

The findings of this project suggest that the creation of an archive for data intermediaries' data and archive-related technical assistance would extend the value of intermediaries' important work, creating new resources for community members, institutions, and researchers.

Appendix: Core Questions for Interview Protocol

1. Could you please tell me a little bit about your data collection process? Do you pull data from government websites? Receive data files directly from government programs or agencies through relationships with staff there?

2. [Drawing on what I've seen on your reports/website/etc. it looks like you use…] Are there other federal datasets you use to inform your work and reporting? Which ones?

3. Are there any federal datasets you are concerned may be less accessible in the future? Why?

4. [I see that you also use… (insert state/local dataset names).] Are there any state or local datasets which come from federally-funded programs or other programs with federally-mandated reporting?

5. Are you concerned that any of these datasets may be less accessible in the future? Why?

6. What data do you wish you could access to better answer your questions? What barriers do you currently face to accessing those data? What would help you access those data? Tools? Relationships?

7. What types of data tools would improve your ability to do your job?

8. Are there any partner organizations or individuals you think might be able to give us additional insight into some of the issues we discussed today? Particularly data that may be at risk?