
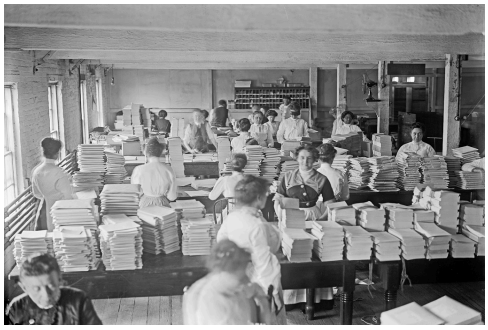


Environmental Scan of Government Information and Data Preservation Efforts and Challenges

Sarah K. Lippincott

December 2018

PRESERVATION OF ELECTRONIC GOVERNMENT INFORMATION
PEGI PROJECT

Publication Notes

Title: Environmental Scan of Government Information and Data Preservation Efforts and Challenges

Author: Sarah K. Lippincott

Editors: James R. Jacobs, Shari Laster, Katherine Skinner, Caitlin Perry

Publisher: Educopia Institute, 235 Peachtree Street, Suite 400, Atlanta, GA 30303

Cover Image Credits: (top left to right) Government Publishing Office (photos 1 and 2), Library of Congress (photo 3); and Markus Spiske, https://unsplash.com/photos/4T5MTKMrjZg (digital text screenshot).

Suggested Citation: Lippincott, Sarah K. *Environmental Scan of Government Information and Data Preservation Efforts and Challenges.* Atlanta, Georgia: Educopia Institute, 2018.

Table of Contents

Introduction

From technical reports and white papers, congressional committee reports and legislation, quantitative data and numeric datasets, posters and presentation slides, and informational websites, the federal government produces vast quantities of born digital information and data. Journalists, policymakers, educators, and the public rely on this data every day to power their own work and to hold the government accountable. Constituents increasingly expect timely, convenient access not only to the most current government information and data, but to historical and longitudinal data that allows for in-depth analysis. The web has facilitated instantaneous publishing in a variety of born digital media, reinforcing pushes for government transparency and data sharing, and leading to a proliferation of openly available government data. Widespread uptake of web publishing has also ushered in a profound shift in the definition of government documents. In the pre-digital era, most publicly disseminated government information came in the form of a static, stand-alone report or publication. In the born-digital era, these standardized formats have expanded and disaggregated and include multimedia collections that may include web pages, PDFs, data sets, or all of the above. In 2011, the Government Publishing Office (GPO)—which at that time was called the Government Printing Office—estimated that 97 percent of the federal government information and data was born-digital (GPO, 2011). A portion of this information (estimates vary on how much) is shared through myriad governmental websites. A 2017 review of the 95 agencies tracked by the General Services Administration (GSA) identified over 265,000 datasets hosted on government web domains (Brock, et al. 2018). There is no question that the public has access to greater quantities of government information than ever before. A 2013 executive order signed by former president Barack Obama "made open and machine-readable data the new default for government information," marking a new era in public access (Open Government Initiative, n.d.).

But distribution of vast quantities of data through the web has given rise to serious challenges and vulnerabilities regarding collection, long-term access, and preservation.

Project Scope and Methodology

In 2018, the PEGI Project team[1] received fiscal support from Arizona State University, Center for Research Libraries, Stanford University, University of Missouri, University of North Texas, University of North Carolina at Greensboro, and Yale University, to commission a scan to document current digital preservation activities and digital preservation gaps in the government information ecosystem. This scan was undertaken in 2018 by Sarah K. Lippincott, with guidance from Shari Laster, James R. Jacobs, and Scott Matheson. The scan activities included defining the information realm, surveying communities of interest, and preparing a report detailing the organizations, services, and infrastructures currently in place to preserve government information.

The environmental scan is focused specifically on born digital data and information produced by the federal government of the United States. For the purposes of this report, unless otherwise noted, the term “government information” refers to born-digital material produced by or through the direct participation of federal government agencies. As used in this report, the term “government

1 For more information on the PEGI Project see https://www.pegiproject.org.

information" covers a wide range of *information products*—including but not limited to datasets, reports and working papers, conference slides and posters, audiovisual materials, and web pages—and *media formats,* such as CSV, PDF, ASCII, and XML files. This scan does not generally include *digitized* government information products. The term "government data" is used throughout this report to refer to a specific subset of government information that comprises primarily quantitative or numeric datasets. In some cases, this report uses the term "government information and data" in certain contexts in order to emphasize the reference to numeric data as well as other types of government-produced content. This report strives for internal consistency in its use of terminology. However, federal government definitions of these terms may differ in meaningful ways from the ways they are used in this report. For example, a broader definition of "data" as "all structured information," was articulated in the 2013 Federal Open Data Policy, which mandated public access to newly created government data (Burwell, et al, 2013). Many more government information products could be considered "data" by this federal definition, especially in the increasingly common cases in which researchers conduct large-scale analyses on a corpus of information products. The term "archive" in this report is used interchangeably with "repository," as is common in the field of data management. Within the Federal government, however, the term "archive" is specific to the mission and activities of the National Archives and Records Administration (NARA).

This environmental scan was conducted between March 2018 and August 2018. The research employed web searches to identify original sources of government information and projects that aim to collect, replicate, preserve, and/or disseminate government information or data. This focus necessarily excludes the dozen or more organizations and initiatives who contribute to raising the visibility of government preservation challenges, who develop the infrastructure used by data preservation efforts, or who coordinate interest groups on this subject are not included in detail in this scan. Similarly, although this research included some specific stakeholder groups who rely on access to government data, including policy-makers, journalists, and scholars, these groups are not the primary focus of this scan and their needs are not considered in detail.

This report describes the landscape of initiatives within and outside of the federal government that aim to disseminate and preserve government information. It first describes government-led initiatives, from dissemination through official agency websites to publication on third-party platforms. Next, it considers the range of initiatives that have emerged in recent years outside of government to address perceived gaps and vulnerabilities in the federal government's curation initiatives and to add value to publicly available information and datasets. It briefly touches upon initiatives that focus on advocacy, awareness, or education, rather than on directly providing preservation and access. The report goes on to address the policies and infrastructures undergirding both government-led and non-government initiatives. It concludes with a brief summary of gaps and recommendations for collective action. Each section contains representative examples, but does not contain an exhaustive list of initiatives relevant to federal government information.

This report describes the landscape of initiatives within and outside of the federal government that aim to disseminate and preserve government information.

Government-Led Information and Data Initiatives

In the course of fulfilling their mandates, federal government agencies and programs produce vast quantities of original information and data, from the results of research studies and investigations to drafts of legislation, internal documents, conference slides, and posters authored by government employees, and informational websites on a wide range of topics. In many cases, these information products are designed for public consumption and are publicly disseminated by the agencies that create it through a vast network of government websites. This section primarily explores the landscape of government-created information and data designed for public use and briefly treats the parallel world of archival records. NARA explains that, according to the US Code (44 U.S.C. 3101), agencies are responsible for "*making* and *preserving* records that contain adequate and proper documentation of the organization, functions, policies, decisions, procedures, and essential transactions of the agency and designed to furnish the information necessary to protect the legal and financial rights of the Government and of persons directly affected by the agency's activities" (emphasis added) (NARA, n.d.). NARA does not explicitly direct agencies to provide access to agency records.

There is no common understanding of who qualifies as a "government information creator" in the federal agency context. As noted by David E. Lewis and Jennifer L. Selin in the *Sourcebook of United States Executive Agencies*, "there is no authoritative list of government agencies" (2012). NARA's inventory of records control schedules lists 15 departments (e.g., Department of Agriculture) and their subunits and 101 independent agencies as well as the executive, judicial, and legislative branch.[2] Indeed, definitions of what counts as an agency varies across legislation, with numbers of federal agencies ranging from 61 (Unified Agenda), to 220 (FOIA.gov), to 443 (USA.gov) (Crews, 2017). It is similarly difficult to pin down the total number of websites the federal government operates. James Jacobs (2015) estimates that there are 16,015 top-level .gov and .mil domains and "135,215 websites with government information." The exact quantity of government information may be difficult to scope, but the preservation challenge implied by these estimates needs no explanation. The rate of change in the quantity of government information being produced (taking into account the information being added as well as that being changed or removed) also defies easy calculation. However, if recent large-scale web harvests conducted at the end of the last two presidential terms are any indication, it is safe to assume that government information is rapidly proliferating. The number of URLs harvested in the 2016 iteration more than doubled relative to those harvested in 2012 (see *Figure 1*).

Further complicating this challenge, the U.S. federal government does not have a centralized information dissemination agency or statistics reporting body. A select group of government agencies assume statistical reporting as a core responsibility. As outlined in Title 44 of the U.S. Code, which articulates the federal government's responsibilities regarding the dissemination of government information, several different agencies perform some of the various functions related to government information stewardship (Public Printing and Documents, 2008). These agencies, which include the GPO

2 Also useful in this context is the NARA Records Management Statute, available at https://www.archives.gov/about/laws/fed-agencies.html

and the National Archives and Records Administration (NARA) are discussed in further detail in a subsequent section.

Statistics reporting is the domain of thirteen Federal Statistical Agencies (FSA), most of which are housed within larger government agencies. The US Census Bureau, for example, is located within the US Department of Commerce (DOC). Coordinated by the Office of Management and Budget (OMB), FSAs "provide essential statistical information for use by governments, businesses, researchers, and the public" (Federal Committee on Statistical Methodology, 2018). The thirteen FSAs produce some of the most heavily cited and impactful statistics, such as gross domestic product (GDP). The OMB provides oversight and coordination for these bodies to ensure quality standards and set policies, but does not bear responsibility for timely and effective dissemination or long-term preservation of the data produced by this loosely affiliated group of agencies. Changes to the structure of the FSAs may impact how critical information products are disseminated in the coming years. The Trump administration's federal agency reform plan, stemming from a 2017 executive order, proposes moving the Bureau of Labor Statistics (BLS) to the Department of Commerce (DOC), and merging the Census Bureau and the Bureau of Economic Analysis (BEA), both of which currently reside within the DOC.[3]

The result of the federal government's decentralized approach to public information dissemination is a system in which each agency essentially self-publishes its content to the web. Most agencies publish some form of information or data in the course of fulfilling their mission. However, they may not consider stewardship of that information a core responsibility and they lack sufficient incentives and resources to undertake the laborious and complicated processes of providing ongoing public access and long-term preservation of digital content. This self-published content (or metadata) often migrates into any number of aggregators, repositories, and other publishing platforms managed by the federal government and other entities, which also can create challenges relating to provenance.

Most agencies publish some form of information or data in the course of fulfilling their mission. However, they may lack sufficient incentives and resources to undertake the laborious and complicated processes of providing ongoing public access and long-term preservation of digital content.

There are mixed signals about whether this may change in the coming years. H.R. 4631, the *Access to Congressionally Mandated Reports Act*, introduced in December 2017, is the latest incarnation of legislation that would require "the public release of any congressionally mandated report within 30 days of its submission to Congress, as long as the report is subject to the Freedom of Information Act," through a GPO-administered repository (Govtrack, 2017). According to Govtrack, this could mean providing rapid public access to around 4,000 reports annually. The legislation has received bipartisan support in the house, and the backing of a number of influential organizations (Govtrack, 2017). However, three previous versions of the bill have stalled. In some cases where the federal government

3 See https://www.performance.gov/GovReform/Reform-and-Reorg-Plan-Final.pdf.

has failed to provide convenient and timely public access, non-governmental organizations have stepped in to fill the gap. Projects like the Congressional Research Service (CRS) Report Archive from the University of North Texas (UNT) Libraries and the Environmental Impact Statement collection at Northwestern University Libraries, are discussed in further detail in the next section.

Among the government information that is made publicly available through government websites and databases, access and security protocols vary depending on the agency and the type of information product. Access restrictions may differ based on content (e.g., concerns about sensitivity or confidentiality) or format. Providing unrestricted downloads of raw datasets entails different considerations than formal publications, for example. In some cases, individuals can view and download raw datasets easily and without restriction. For example, the National Center for Environmental Information's Storm Events Database allows bulk downloads of storm data in CSV format via HTTPS and FTP.[4] Some agencies require users to register and agree to terms before accessing certain datasets. To access the National Postsecondary Student Aid Study data from the National Center for Education Statistics (NCES), for example, users are required to create an account and review a statement governing proper use of the data, which reads, "Under law, public use data collected and distributed by the National Center for Education Statistics (NCES) may be used only for statistical purposes. Any effort to determine the identity of any reported case by public-use data users is prohibited by law. Violations are subject to Class E felony charges of a fine up to $250,000 and/or a prison term up to 5 years." In other cases, the public may only be able to access summary statistics or visualizations, rather than the raw datasets. Finally, some datasets may only be accessed through trusted intermediaries, such as the Federal Statistical Research Data Centers (FSRDC) that provide mediated, on-site access to scoped subsets of sensitive administrative data that is otherwise not available for public consumption.

Occasionally, accessing electronic government information involves a fee, as in the case of PACER (Public Access to Court Electronic Records), a public access system for case and docket information from federal appellate, district, and bankruptcy courts.[5] PACER's fees of $0.10 per page have been the subject of controversy, though they waive these feels for litigants and their lawyers, users who request less than $15.00 per year in records, and some academic researchers. Critics say that the ostensibly minor costs actually pose a major hurdle to organizations and individuals who wish to perform big data analyses on court records, or build more usable public search interfaces (Browdie, 2016). The government has argued that the fees are appropriate as they are not charged for *access* to the documents, but rather for convenient electronic *delivery*. Critics have noted, however, that though the majority of fees are directed into the infrastructure of electronic filing and retrieval of court documents, about a quarter went towards unrelated expenses such as courtroom technology (Browdie, 2016). In April 2018, a judge ruled in favor of the plaintiffs in a class action lawsuit that alleged that PACER fees violated the E-Government Act (Scarcella, 2018).

Much of the government information that is made publicly available is distributed directly by the agencies that create it, via those agencies' top-level websites and subdomains. However, this is far from

4 See https://www.ncdc.noaa.gov/stormevents/ftp.jsp.

5 https://www.pacer.gov/.

the end of the story, as information and data may also be hosted or replicated through a variety of other government-administered platforms. These government-administered sources include entities such as:

- single-agency websites;
- inter-agency data portals;
- commercial platforms;
- websites of research organizations that partner with government agencies;
- information stewardship bodies, including the GPO and the Library of Congress (LC); and
- the National Archives and Records Administration (NARA), the federal records-keeping body.

This section examines how government agencies disseminate the information and data they create or manage through each of these channels.[6] Archival records, a distinct subset of government information, are also discussed briefly at the end of this section.[7]

Single-Agency Websites

As agencies increasingly adopt digital-only formats for their information products, their websites function as their primary public-facing content repositories and sources of information about their activities (Brock et al., 2018). These websites host a wide range of information and data products developed both *by* and *for* the agency, and for both internal and public use. Most of the information products available through government websites have been specifically designed for public consumption, or deemed to be of public interest, with the exception of internal or interagency materials shared online to facilitate access by government employees, which may incidentally be discoverable by the general public.

Government information designed for public consumption may aid scientific research, support local government activities, or perform a public service, among other purposes. It takes a range of forms, including everything from informational pamphlets to high-resolution images. Numerical data is one of the most commonly encountered information products. For example, the National Center for Education Statistics (NCES), the statistical arm of the Department of Education, provides access to a range of datasets and data analysis tools on its website. NCES hosts raw data from major government surveys and data collection efforts such as the Integrated Postsecondary Education Data System (IPEDS) alongside interactive tools such as DataLab, which allows web users to dynamically create tables and run analyses based on federal

Government information designed for public consumption may aid scientific research, support local government activities, or perform a public service.

6 A list of additional examples of government information and data initiatives, specifically efforts to aggregate or improve dissemination of government data, is included as part of Appendix I.

7 Government records are defined in Disposal of Records (44 U.S.C. Chapter 33) as "the "recorded information, regardless of form or characteristics, made or received by a Federal agency under Federal law or in connection with the transaction of public business and preserved or appropriate for preservation by that agency or its legitimate successor as evidence of the organization, functions, policies, decisions, procedures, operations, or other activities of the United States Government or because of the informational value of data in them."

datasets. To provide context for its own data, NCES also provides access to data from certain relevant surveys conducted by other organizations, such as the Trends in International Mathematics and Science Study conducted by the International Association for the Evaluation of Educational Achievement (IAEEA), an independent research body. A 2017 study conducted by the Federal Research Division of the Library of Congress on behalf of GPO found that, among the sample of government agencies it studied, third party content (including information produced by grantees and contractors) undergoes the same process of review before publication as the agency's own content (Brock et al, 2018). The example of the NCES illustrates one of the complexities of preserving government data when that data includes not only vast quantities of statistical information, but also web-based interactive tools, which are key to interpreting the data at different points in time.

Many government agencies host information not only on their top-level .gov website, but on subsidiary or discrete .gov sites that function specifically as publication repositories. For example, the Centers for Disease Control and Prevention (CDC) delivers its research reports and data through CDC Stacks, a subdomain of cdc.gov.[8] The portal was developed in response to the White House Office of Science and Technology Policy (OSTP) memo entitled "Increasing Access to the Results of Federally Funded Scientific Research," which directed major research funders to provide public access to and long-term preservation for research funded through their agency. A number of other agencies provide similar portals. For instance, PubDefense provides public access to journal articles that result from DoD and ODNI/IARPA-funded research. PubDefense content is in turn aggregated into the Department of Energy's (DOE) Office of Scientific and Technical Information (OSTI) portal, which makes available 70 years of research results from DOE and its predecessor agencies through a single search interface. The DOE also maintains the Public Access Gateway for Energy and Science, a repository of published research made possible through DOE funding.[9] Other examples include the USGS Publications Warehouse maintained by the United States Geological Survey (USGS) and Smithsonian Research Online.[10]

Several government agencies maintain dedicated portals for outdated content, often under the moniker of "archives." For instance, the Agency for Healthcare Research and Quality operates a discrete domain where it offers "Outdated information that may be useful for reference purposes" and "Materials and sites of historical or research interest."[11] The US Energy Information Administration (EIA) hosts an "archive" of historical energy outlook information and the US Geological Survey (USGS) provides access, through a dedicated web domain, to "scientific information websites formerly maintained by the U.S. Geological Survey."[12]

The internal policies that govern dissemination and retention of government information products vary by agency. A study conducted on behalf of the GPO in 2017 found that some agencies maintain formal policy documents that dictate how information products can be disseminated online, while others

8 See https://stacks.cdc.gov/.

9 See https://publicaccess.dtic.mil/padf_public/#/home, https://www.osti.gov/, and https://www.osti.gov/pages/.

10 See https://pubs.er.usgs.gov/ and https://research.si.edu/.

11 See https://archive.ahrq.gov/.

12 See https://www.eia.gov/outlooks/aeo/archive.php and https://archive.usgs.gov/.

review content on a case-by-case basis or via review through an internal chain of command (Brock et al, 2018). Preservation practices also vary, as detailed in this report. Some agencies systematically preserve master copies of their publications, through a central office or the agency's library, but the study found that most of the agencies it reviewed rely on their Office of the Chief Information Officer, or equivalent, to liaise with the National Archives and Records Administration (NARA) to facilitate the federally mandated transfer official records for preservation (Brock et al, 2018). The limitations of this records transfer system are discussed in a subsequent section.

Inter-Agency Portals

A range of government-run digital repositories aggregate metadata, information, and/or datasets from multiple government agencies for the purposes of preservation and/or dissemination. Information is ingested into these portals through both manual and automated processes.

Inter-agency portals may focus on a topical subset of information or on specific types of government entities. Some of the largest interagency portals include USA.gov, Data.gov, and govinfo. USA.gov launched in 2000 (as FirstGov) "as a search engine the public can use to navigate across government agencies to find information general users often seek" (Latham, 2018). Front-end users can access topical information portals on subjects like health, the environment, and consumer issues, while developers can access APIs and data feeds to build applications that further enable timely access to information.

A subset of government data can be identified through the federal government's main data portal, Data.gov, the U.S. government's closest approximation of a centralized statistical reporting service, Data.gov facilitates compliance with the 2013 Federal Open Data Policy requiring that "newly-generated government data . . . be made available in open, machine-readable formats, while continuing to ensure privacy and security" (Data.gov, 2018) by hosting metadata records for over 285 thousand individual datasets from a range of government agencies. This number is impressive, but likely represents only a fraction of publicly available datasets (Johnson and Kubas, 2018). Data.gov functions as a portal to government data, not as a data repository, in that it only hosts metadata records, which link to data hosted elsewhere. Ingest of metadata into Data.gov is largely automated. Federal agencies are required by the Federal Open Data Policy, OMB M-13-13 to provide a machine readable JSON file directly from their website (i.e., agency.gov/data.json) that Data.gov can crawl.

The GPO-administered repository govinfo, a successor to the Federal Digital System (FDsys) database, "provides free public access to official publications from all three branches of the Federal Government," (govinfo, 2018). Govinfo touts its three grounding principles of facilitating public access through its high-quality search system, ensuring content integrity through robust content management practices, and providing long-term access through the implementation of standards-compliant digital preservation practices. Although govinfo hosts millions of documents, this represents only a small subset of the content within the scope of GPO's legislative mandate.

A range of other inter-agency information portals focus on narrower subsets of government information. They may facilitate compliance with public access policies or legislative requirements, as in

the case of PubMed Central, the designated repository for papers submitted in accordance with the NIH's public access policy.[13] PubMed Central aggregates content produced by or with the support of a range of government agencies (ACL, AHRQ, ASPR, CDC, DHS, EPA, FDA, NASA, NIH, NIST, VA), and is primarily concerned with providing public access to federally funded research. Some of these repositories, including PubMed Central, are managed by one of the four national libraries, National Agricultural Library, the National Library of Education, the National Library of Medicine, and the National Transportation Library.

Some of these narrower interagency initiatives also have a specific topical focus. These portals bring together data, publications, and other content relevant to specific disciplines or research communities. For instance, the Federal Interagency Forum on Aging Related Statistics provides access to data on aging aggregated from a variety of government agencies, including the Bureau of Labor Statistics, the Census Bureau, and the Department of Veterans Affairs among others. The Repository and Open Science Access Portal maintained by the National Transportation Library (NTL) serves as a portal for open access transportation-related research produced by the US Department of Transportation (DOT) as well as state departments of transportation and other sources.[14] AGRICOLA, the catalog of the National Agriculture Library not only facilitates USDA compliance with the OTSP memo, but also aggregates over five million records for publications and resources encompassing all aspects of agriculture and allied disciplines in a variety of media and from global sources.[15] Science.gov boasts 200 million pages of "authoritative federal science information" aggregated from 60 databases and 2,000 government websites through its federated search interface.[16] PACER, a service operated by the Administrative Office of the U.S. Courts, provides fee-based public access to "case and docket information online from federal appellate, district, and bankruptcy courts."[17]

Commercial Platforms

Social media and transient information formats have become increasingly common channels for government agencies to disseminate information. Many of these forms of communication rely on commercial media platforms. A 2017 study conducted by the GPO found that the agencies it reviewed disseminated content on at least one (and usually multiple) digital-only channels, including blogs, email lists, mobile apps, podcasts, and social media platforms, including Facebook, Flickr, Google+, Instagram, LinkedIn, Pinterest, Slideshare, Storify, Tumblr, Twitter, and YouTube (Brock et al, 2018). In many cases, agencies use social media to point back to content hosted on the agency's main website. However, agencies are also creating original content specifically tailored for social media platforms. For example, the federal government maintains an official YouTube channel, where it posts educational and instructional videos.[18] The Environmental Protection Agency (EPA) hosts a selection of presentation slides and posters authored by employees on Figshare.[19] The State Department maintains an account on

13 See https://www.eia.gov/outlooks/aeo/archive.php and https://archive.usgs.gov/.

14 See https://rosap.ntl.bts.gov/.

15 See https://agricola.nal.usda.gov/.

16 See https://www.science.gov/.

17 See https://www.pacer.gov/.

18 See https://www.youtube.com/usagov1.

19 See https://epa.figshare.com/.

the microblogging platform Medium, where it shares original content as well as re-posts from its official website.[20] NARA hosts open source code its staff has developed on GitHub.[21] Users can download code for "a module to pull images from the National Archives Catalog to edit in Drupal 8" or "a captcha module that helps identify text vs. handwritten documents." Many other government agencies also use GitHub as a repository for source code and open data as well as for the public drafting of policy.[22] The White House Office of Management and Budget (OMB), for example, used GitHub to create a site to collect public feedback on proposed revisions to Circular A-130, "Managing Information as a Strategic Resource."[23] A GitHub "community" that accepts participation from anyone with a government email address allows for knowledge sharing around best practices.[24]

Agencies must comply with government policies and regulations around the use of third party services. The General Services Administration (GSA) is responsible for negotiating terms of service agreements compatible with "federal laws and regulations on security, privacy, accessibility, records retention, ethical use, and other specific agency policies," and maintains a list of free tools that have federal-compatible terms of service agreements (DigitalGov, 2018).[25] Per a 2011 Office of Management and Budget (OMB) memorandum, commercial cloud services that hold federal data must be authorized by the Federal Risk and Authorization Management Program (FedRAMP), which evaluates cloud services based on their information security and other criteria. The FedRAMP website lists 147 agencies that currently use, or plan to use, commercial cloud products to host data.[26] The federal data hosted by these cloud services may not be intended for public use.

Government partnerships with academic institutions, non-profits, and other entities result in a wealth of information and data that may or may not be simultaneously hosted in other repositories.

Government agencies are also collaborating with commercial entities to integrate government data into their platforms. In 2015, for example, Data.gov announced a program to enable users to open selected datasets directly in the cloud-based visualization tools Plotly and CartoDB (AshLCk and Williams, 2015). Content hosted on commercial third-party platforms is some of the most difficult to identify. Agencies do not necessarily provide links to all their social media and web publishing presences on their primary websites, and the third-party platforms rarely offer metadata and curation that would allow for easy identification of all federal government content on their sites.

20 See the State Department's Medium account at https://medium.com/@StateDept. A list of all government agencies using Medium as of 2015 is available at https://www.govloop.com/community/blog/medium-perfect-fit-government/.

21 See https://github.com/usnationalarchives.

22 See https://government.github.com/.

23 See https://a130.cio.gov/.

24 See https://github.com/government/welcome#readme.

25 The full list of free tools that have federal-compatible terms of service agreements is available at https://digital.gov/resources/negotiated-terms-of-service-agreements/.

26 See https://tinyurl.com/y8xbmnya.

Research Partnerships

Government partnerships with academic institutions, non-profits, and other entities result in a wealth of information and data that may or may not be simultaneously hosted in other repositories. This category does not include data created by third parties through government *funding*, but instead focuses on instances where formal research partnerships between the government and an academic, non-profit, or commercial organization support ongoing information and data collection, archiving, and analysis.

One example of this type of partnership is the National Archive of Criminal Justice Data, which "archives and disseminates data on crime and justice for secondary analysis."[27] The database is hosted by the Inter-University Consortium for Political and Social Research (ICPSR) and receives ongoing support from the Department of Justice and the Bureau of Justice Statistics. It "contains data from over 2,700 curated studies or statistical data series," including significant datasets from the FBI and other government agencies.[28] The corollary for agricultural data is the USDA Economics, Statistics and Market Information System (ESMIS), "a joint effort between several key agencies within the USDA and Mann Library at Cornell University," which archives dozens of reports and datasets created by the USDA.[29] Additional examples include the Cooperative Institute for Research in Environmental Sciences (CIRES), a partnership between NOAA and the University of Colorado Boulder, and the Department of State Foreign Affairs Network (DOSFAN) electronic research collection, a partnership between the United States Department of State and the Federal Depository Library at the Richard J. Daley Library, University of Illinois at Chicago (UIC).[30]

Some research partnerships use information or data generated by federal agencies to create powerful tools for visualization and analysis. For example, The Opportunity Atlas, which allows users to visualize social mobility data, is the result of a collaboration between researchers at the Census Bureau, Harvard University, and Brown University.[31]

Research partnerships between federal government agencies and corresponding state agencies have emerged to improve local access to federal government information. The Census Bureau's State Data Center Program liaises with US states and territories to facilitate data access, partner on special projects, and provide other resources. This program enables work like that of the Missouri Census Data Center, which has built its own enhanced search interfaces and analysis tools for datasets obtained from the Census Bureau and other federal and state-level agencies.[32]

The federal government also engages in research partnerships with commercial entities. In some cases, the exact nature of such partnerships can be difficult to determine. For example, a business management consulting firm, ICF, manages a repository of information and data created by USAID's Demographic and Health Surveys Program. Though a disclaimer cautions that the site's content is not

27 See https://www.icpsr.umich.edu/icpsrweb/content/NACJD/index.html.

28 Ibid.

29 See http://usda.mannlib.cornell.edu/MannUsda/homepage.do.

30 See https://cires.colorado.edu and http://dosfan.lib.uic.edu/ERC/.

31 See https://www.opportunityatlas.org/. UIC recently ended this 21-year collaboration, giving this content to GPO. See https://www.census.gov/about/partners/sdc.html.

32 See https://www.census.gov/about/partners/sdc.html.

official government information, the documents in the repository appear to be U.S. government publications that fall within the public domain.[33]

Federal Information Stewardship Bodies

The Government Publishing Office (GPO) and the Library of Congress (LC) each have a unique role in the collection and dissemination of government information. The GPO has a dual responsibility to "produce and distribute information products and services for all three branches of the Federal Government" in print and digital formats and to "provide for permanent public access to Federal Government information at no charge through the Federal Depository Library Program (FDLP), the Federal Digital System (FDsys) and govinfo" (GPO, 2018).[34] The GPO has historically provided printing services for federal agencies, meaning that many government publications were automatically funneled through GPO (Brock et al, 2018). This allowed the GPO to keep track of a large proportion of the federal government's published content. Even so, researchers have estimated that the GPO may have contact with only about half of the printed government publications produced each year, not to mention born-digital content (Brock et al, 2018). Though so-called fugitive documents existed in the print era, they have proliferated with the emergence of easy digital publishing (Jacobs, 2018c).

A 2017 study noted an acute lack of awareness among federal agencies about how and when to report the creation of digital content to GPO.

The ease of web publishing means not only that government-created information has proliferated over the past few decades, but that it no longer requires GPO as intermediary before reaching the public. This has exponentially complicated GPO's mission. A 2017 study noted an acute lack of awareness among federal agencies about how and when to report the creation of digital content to GPO (Brock et al, 2018). Most indicated that they had no regular schedule for notifying GPO of new digital publications, and several expressed confusion about what types of publications they should report.

GPO has undertaken a range of efforts to adapt to this new digital publishing environment. GPO administers its own digital repository, govinfo, a successor to the Federal Digital System (FDsys) database, which "provides free public access to official publications from all three branches of the Federal Government" (govinfo, 2018). GPO also maintains a federated search engine called MetaLib that simultaneously searches the databases of multiple government agencies and administers the Catalog of Government Publications (CGP), a searchable index that aims to cover all government publications within the scope of the FDLP.[35] The breadth and depth of the content indexed are substantively greater in the CGP than on govinfo; however, CGP is a catalog and not a repository and therefore does not commit to provide long-term stewardship for the content it indexes. GPO also administers The Federal

33 See https://dhsprogram.com/publications/.

34 FDSys will be fully replaced by govinfo in December 2018.

35 See https://catalog.gpo.gov/F?RN=311536437 and https://metalib.gpo.gov/V?RN=579295977.

Depository Library Program (FDLP) Web Archive, which harvests selected federal government websites using the Archive-it digital archiving service.[36]

GPO's collection development plan defines its priorities regarding digitization and ingest into govinfo (Office of the Superintendent of Documents, 2018). The 2018 plan emphasizes digitization of content that completes document series also represented in govinfo, and the inclusion of more executive branch content. The collection development plan also highlights GPO's preservation mandate, stating that "content ingested into GPO's system of online access will remain permanently publicly accessible, except under very rare and specific circumstances" (Office of the Superintendent of Documents, 2018). It also recognizes the preservation statutory authority of the National Libraries (agriculture, medicine, transportation, education), LC, and NARA.

In addition to its efforts to collect and index government information, GPO has been involved in coordination efforts that aim to bolster the infrastructure for the dissemination and preservation of born digital government information. For example, GPO is partnering with around three dozen academic libraries to support the LOCKSS-USDOCS network, a privately-run network also known as the Digital Federal Depository Library Program, which "replicates key aspects of the United States Federal Depository System" by hosting born digital data on a geographically distributed network of servers (LOCKSS, 2018). GPO also manages the Federal Information Preservation Network (FIPNet), which works with a range of partner agencies on the stewardship of digitized and born-digital content.

Unlike GPO, LC's core functions do not include the collection, preservation, or dissemination of government information. Rather, LC aims to provide a comprehensive research library in support of congress, the federal government, and the American public (Library of Congress, 2015). Nevertheless, LC has, for well over a decade, contributed leadership and support through programs such as the National Digital Information Infrastructure and Preservation Program (NDIIPP)—predecessor to National Digital Stewardship Alliance (NDSA), now hosted by the Digital Library Federation (DLF)—which provided funding and leadership on digital preservation of cultural heritage, including state government information. LC's long-running *The Signal* blog has become a well-used resource in the field of digital cultural heritage stewardship.[37] LC has also taken an active role in digital preservation efforts related to government information and to other content with significant cultural or historical value. LC is also a partner on the End of Term web harvesting project alongside GPO, the California Digital Library (CDL), George Washington University (GWU), the Internet Archive (IA), Stanford University, and the University of North Texas (UNT). LC makes available thousands of archived government webpages organized as thematic collections. For example, the United States Congressional Web Archive contains over one thousand archived websites of congressional representatives and committees.[38] LC is also an important creator of government information in its role as administrator of the Congressional Research Service (CRS). As of September 2018, LC began providing access to CRS reports through a new web portal.[39] Finally, although LC does not typically curate *government* information, it has assembled distinctive born-

36 See https://www.fdlp.gov/project-list/web-archiving.

37 See https://blogs.LC.gov/thesignal/.

38 See https://www.loc.gov/collections/united-states-congressional-web-archive/.

39 See https://crsreports.congress.gov/.

digital collections that hold important cultural and historic value. For example, the LC Web Archives provides datasets for analysis, such as a collection of 10,972 unique GIFs harvested from the image-sharing website Giphy.[40]

Official Government Records: National Archives and Records Administration

NARA and federal agency staff are required to manage government information in the form of official government records. NARA bears responsibility for the stewardship of archival government records, and manages the process for scheduling and appraisal of all government records in order to fulfill a four-fold mission of identifying, protecting, preserving, and making publicly available the "historically valuable records of all three branches of the Federal government" (NARA, 2018a). Government records are a broad category that includes published information, sensitive and confidential information, and internal documentation that results from the day-to-day activities of the federal government (e.g., emails, meeting minutes, documentation of legislatively required duties).

Federal records laws and regulations require agencies to identify all types of records created, and propose a records schedule outlining which records should be retained for what lengths of time, to support business use and provide protection for public rights and government accountability. Agencies also propose which records might be of permanent historical value. Only the Archivist of the United States has the authority to approve these schedules, and the Archivist has final authority to determine what is a record. No agency can legally destroy records without authorization from the National Archives. Given that government information of public and historical interest is now produced in a range of forms and formats, there are policy questions that require consideration. Less than five percent of records are scheduled for permanent retention and must be transferred to NARA.[41] Because many online publications and websites are scheduled as temporary records, "few federal agencies have any legal obligation to preserve web content that they produce long-term," leading to a lack of clear protocols for deposit and long-term stewardship (Lazorchak, 2015).

NARA's role, as defined in its legislative mandates, prioritizes long-term preservation. However, public access to the permanently valuable records of the federal government is also at the forefront of the agency's mission statement, and NARA's most recent strategic plan puts forward the ambitious goal of making 82 percent of NARA holdings available for public discovery and access by 2021 by accelerating processing and description activities.[42] NARA's holdings include 750 million unique born-digital files, some of which are discoverable through its Access to Archival Databases (AAD) tool.[43] NARA also provides a searchable catalog of its collection, including digitized and born-digital content that can be accessed online.[44] The catalog contains records of 85 percent of NARA's actual holdings.

NARA's public access goals reflect the ongoing challenge of discovery of electronic records (and other types of government information). Even where searchable catalogs and federated search portals exist,

40 See https://labs.LC.gov/experiments/webarchive-datasets/.

41 See: https://www.archives.gov/files/records-mgmt/email-management/final-capstone-white-paper.pdf

42 See https://www.archives.gov/about/plans-reports/strategic-plan/strategic-plan-2018-2022.

43 See https://aad.archives.gov/aad/.

44 See https://catalog.archives.gov/.

poor metadata quality, proliferating information silos, a lack of deposit requirements, and poor awareness of preservation mandates and guidelines remain stubborn challenges. Since 2014, a Federal Web Archiving Working Group, comprising representatives from GPO, NARA, and LC, alongside several other agencies including the National Library of Medicine and the Smithsonian Institute continue to meet monthly to share best practices and develop collective solutions.

In addition to accepting regular deposits of print and electronic materials from the federal government, NARA conducts periodic web harvests of congressional and agency websites. NARA has conducted biennial web harvests of congressional websites since 2006, with the most recent harvest occurring in 2017; however its last large-scale harvest of executive agency websites took place in 2004.[45] NARA's announcement that it would not conduct a comprehensive web harvest at the end of the Bush administration in 2008 spurred the creation of the End of Term .gov/.mil Crawl project, a collaborative initiative of the California Digital Library (CDL), Internet Archive (IA), LC, University of North Texas (UNT), Stanford University, George Washington University, and GPO (Nye, 2017).[46]

45 See https://www.webharvest.gov/.

46 Stanford University and George Washington University officially joined the project for the 2016 harvest.

Non-Government Initiatives

Providing widespread access to government information is not a new concept. Libraries have long facilitated public access to Census data, legislation, and other government information as part of their core services to researchers and the public. Many serve as official stewards of government information through the Federal Depository Library Program (FDLP). The U.S. Government Publishing Office (GPO) distributes select government information to these libraries and requires them to make it freely and publicly available.

However, the proliferation of born-digital government information requires new approaches to collection and preservation. The web has made *access* to government information faster and digital government publications more readily available for many users. However, the sheer volume of information being produced, and its decentralized distribution through hundreds of government domains, has disrupted traditional models of preservation. Jacobs (2014) illustrated the scope of the preservation challenge in a chart depicting the 10,200 tangible items distributed by GPO to FDLP libraries in one year compared to the 160 million URLs harvested in the 2008 End of Term Crawl, a large-scale web harvesting project. An updated version of this chart is presented as Figure 1, comparing the approximately 2,511,610 titles currently managed and made available by GPO through govinfo, other GPO servers, and official agreements with agency partners, to the 160 million and 350 million URLs harvested in 2012 and 2016 End of Term Crawls, respectively (Sherman, 2018).

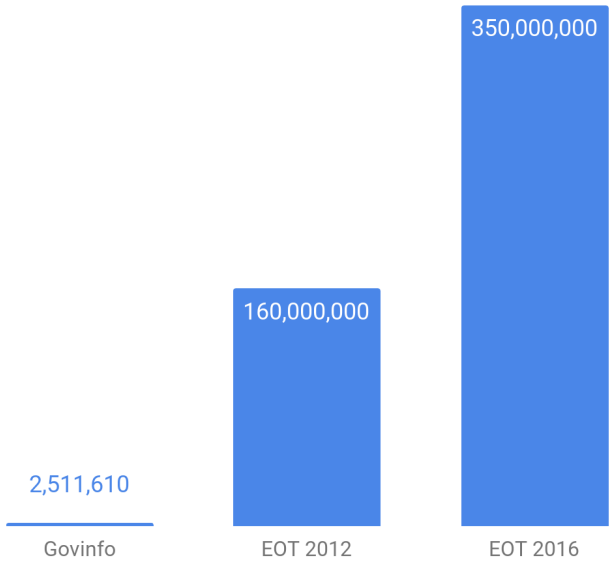

Figure 1. Titles made available by GPO as of October 2018 through its repository govinfo, other GPO servers, and official agency partnerships; compared with URLs harvested in the 2012 and 2016 End of Term Crawls.

To make sense of the deluge, dozens of non-governmental projects have emerged to curate and preserve born digital government information. These projects have developed in response to a variety of needs and concerns and have therefore adopted heterogeneous goals and approaches. In general, they function independently of the government agencies whose data and information they archive. This section describes several of these projects in detail and outlines several broad categories of projects.

It is important to note that, though information-collection and preservation efforts cover a wide range of government agencies, none is comprehensive. At their most ambitious, they aspire to duplicate and provide access to the most at-risk information, as identified by qualified experts in the field. Other

projects have more modest or specific goals, such as efforts to preserve and disseminate Census Bureau data from a particular year.

The breadth of agency content represented in non-government-administered repositories is impressive. This scan identified non-government repositories that collect and store information and data from at least 58 different government agencies and sub-agencies and the executive branch, including all 13 Federal Statistical System (FSS) agencies. Table 1 shows how many non-government repositories identified in this scan duplicate information and data from each of the listed federal agencies.

Federal Agency Name	Non-Government Data Collection Efforts Identified
Consumer Financial Protection Bureau (CFPB)	**1**
Department of Commerce (DOC) including: • *Bureau of Economic Analysis (BEA)* • *Census Bureau (CB)* • National Oceanic and Atmospheric Administration (NOAA)	**18**
Department of Defense (DOD)	**3**
Department of Education (ED) including: • *National Center for Education Statistics (NCES)*	**4**
Department of Energy (DOE) including: • *Energy Information Administration (EIA)* • National Renewable Energy Laboratory (NREL)	**7**
Department of Health and Human Services (HHS) including: • Centers for Disease Control (CDC) • Centers for Medicare and Medicaid Services (CMS) • Food and Drug Administration (FDA) • *National Center for Health Statistics (NCHS)*	**5**
Department of Homeland Security (DHS)	**3**
Department of Housing and Urban Development (HUD)	**3**
Department of the Interior (DOI) including: • United States Geological Survey (USGS) • United States Forest Service (USFS) • National Water Quality Monitoring Council (NWQMC)	**6**
Department of Justice (DOJ) (5) including: • Federal Bureau of Investigation (FBI) • *Bureau of Justice Statistics (BJS)* • Drug Enforcement Agency (DEA)	**5**
Department of Labor (DOL) including: • *Bureau of Labor Statistics (BLS)* • Occupational Safety and Health Administration (OSHA)	**7**
Department of Transportation (DOT) • *Bureau of Transportation Statistics (BTS)*	**4**
Department of the Treasury (TRE) • *Statistics of Income*	**2**

Institute of Museum and Library Services (IMLS)	**1**
Environmental Protection Agency (EPA)	**8**
Equal Employment Opportunity Commission (EEOC)	**1**
Executive Branch (e.g., executive orders)	**3**
Federal Deposit Insurance Corporation (FDIC)	**1**
Federal Reserve System (FRB)	**2**
Federal Trade Commission (FTC)	**2**
Freddie Mac	**1**
National Aeronautics and Space Administration (NASA)	**6**
National Endowment for the Humanities (NEH)	**1**
National Interagency Fire Center (NIFC)	**1**
National Institutes of Health (NIH) including: • National Cancer Institute (NCI)	**1**
National Labor Relations Board (NLRB)	**1**
National Science Foundation (NSF) • *National Center for Science and Engineering Statistics*	**5**
Nuclear Regulatory Commission (NRC)	**1**
Securities and Exchange Commission (SEC)	**2**
Small Business Administration (SBA)	**1**
United States Agency for International Development (USAID)	**1**
United States Consumer Product Safety Commission (CPSC)	**1**
United States Department of Agriculture (USDA) including: • Agricultural Marketing Service (AMS) • *Economic Research Service (ERS)* • FAS (Foreign Agricultural Service) • *National Agricultural Statistics Service (NASS)* • World Agricultural Outlook Board (WAOB)	**8**
United States Department of State (DOS)	**4**
United States Department of Veterans Affairs (VA)	**1**
United States Social Security Administration (SSA) • *Office of Research, Evaluation, and Statistics*	**3**

Table 1. *Coverage of federal government agency information in non-government administered information repositories. These numbers are estimates and represent the minimum number of initiatives ongoing at the time of publication. Federal Statistical Agencies (which assume statistical reporting as a core responsibility) are indicated in italics.*

As Table 1 indicates, coverage is uneven and reflects the perceived vulnerability and significance of the data produced by various federal agencies. DOC data is collected and stored by fifteen repositories, for example, largely because climate data produced by NOAA has been identified as being particularly at-risk.

Some efforts also focus on mirroring metadata from interagency portals including Data.gov, or capturing documents from other government entities, including the executive branch and the Congressional Research Service. A more comprehensive list of non-government information and data curation efforts can be found in the living appendix to this report (see Appendix I). Government information preservation initiatives[47] led by non-governmental organizations take a number of forms. Their approaches to collection and curation are informed by a variety of goals and desired outcomes, from aggregating information relevant to specific research communities to fostering greater government transparency. This section describes several categories of government information preservation initiatives and provides examples of each.

Topical or Format-Based Collections

Given that government information is produced by various agencies, who publish it through their own websites, it is not necessarily organized in a way that is intuitive to researchers and policy-makers interested in specific topics or disciplines. Multiple agencies produce information and data relevant to individuals interested in aging, for example. Discipline-based research repositories aggregate information from multiple government agencies, and in some cases non-governmental sources, into a central clearinghouse relevant to a specific research community or discipline.

One of the largest and well-known aggregators of government data is the Inter-university Consortium for Political and Social Research (ICPSR), and international consortium of research institutions that provides leadership and services to the social science research community.[48] According to their most recent strategic plan, ICPSR's data curation efforts focus on "data that has the greatest likelihood to have a transformational scientific impact and make a difference in future research" (ICPSR, 2017). ICPSR aggregates and provides access to data from a wide range of government and non-government sources, which it curates into thematic collections such as the National Archive of Criminal Justice Data (previously discussed in the section of this report on government research partnerships) and the ICPSR Census repository.[49]

A range of other initiatives have developed to serve research communities interested in climate change, criminal justice, and other topics. For example, the Georgetown Climate Center Adaptation Clearinghouse aggregates data relevant to policymakers and other individuals interested in helping communities adapt to climate change. It assembles and contextualizes relevant resources from governmental agencies like NOAA and the EPA as well as research produced by scholars and policymakers working in the areas of climate change resilience. The GCCAC, operated by the Georgetown Climate Center, receives financial support from several foundations, and partners with non-profit organizations and government agencies (such as the EPA) to collect and curate content.[50] Other repositories focused on climate data include Climate Mirror and the Azimuth Climate Data Backup Project, both volunteer-run organizations that aim to create redundant copies of publicly available

47 The term "preservation" here is used by many of these groups to mean many different things, from collecting and storing the content to actively curating the content, to full lifecycle management.

48 See https://www.icpsr.umich.edu/icpsrweb/.

49 See https://www.icpsr.umich.edu/icpsrweb/content/NACJD/index.html and https://census.icpsr.umich.edu/census/.

50 See the GCCAC collection of federal government data at http://www.adaptationclearinghouse.org/networks/federal-adaptation-resources/.

climate data.[51] The University at Albany School of Criminal Justice *Sourcebook of Criminal Justice Statistics*, which was discontinued around 2013, was another example of a curated collection that brought together relevant reports and data sets from the state and federal government and other sources for the benefit of researchers, policymakers, and the public.[52] To be included in the database, the research must only be deemed to have national relevance and be methodologically sound. The American Presidency Project, a non-profit organization hosted at the University of California, Santa Barbara, compiles data sets and documents, including speeches, executive orders, and public papers, among other content.[53]

Some collections focus more on format than topic. Many projects target raw datasets (e.g., CSV files of government statistics, geospatial data). DataLumos, a project of the social science data repository ICPSR, archives government data on a range of subjects.[54] DataLumos describes itself as a crowd-sourced repository; it encourages the public to submit relevant datasets and improve metadata. Several other projects focus on audiovisual materials. The Internet Archive's FedFlix collection aggregates government-produced films on a variety of subjects.[55] The C-SPAN video library makes available hundreds of thousands of hours of footage of live congressional coverage.[56] Although the public benefits from free and open access to much government information, some data is deemed too sensitive to share openly. Solutions have emerged to provide mediated access to raw government datasets that would not otherwise be publicly available, or to reduce costs or effort involved in accessing such data. The US Census Bureau makes data available through a network of Federal Statistical Research Data Centers, "partnerships between federal statistical agencies and leading research institutions. They are secure facilities providing authorized access to restricted-use microdata for statistical purposes only" (US Census Bureau, 2018).

Finally, many commercial content databases, often developed for the academic library market, draw from government information sources, repackaging them for use by students and scholars. For example, EBSCO provides subscription access to Agricola, the catalog of the National Agricultural Library (NAL) and ERIC, a database of indexed and full-text education literature and resources maintained by the Department of Education.[57] The vendor HeinOnline specializes in digitized and born digital government

Many commercial content databases, often developed for the academic library market, draw from government information sources, repackaging them for use by students and scholars.

51 See https://climatemirror.org/ and http://math.ucr.edu/home/baez/azimuth_backup_project/.

52 See https://www.albany.edu/sourcebook/about.html.

53 See http://presidency.proxied.lsit.ucsb.edu/ws/.

54 See https://www.datalumos.org/datalumos/.

55 See https://archive.org/details/FedFlix.

56 See https://www.c-span.org/about/videoLibrary/.

57 See https://www.ebsco.com/products/research-databases/agricola and https://www.ebsco.com/products/research-databases/eric.

documents, including Congressional Record, Federal Register, and Code of Federal Regulations, and legislation going back over a century.[58] ProQuest Legislative Insight compiles full-text PDFs of documents created in the course of making and passing laws while ProQuest Congressional compiles content documenting the entire history of the US Congress.[59]

WestLaw and LexisNexis, well-known databases of resources for lawyers and legal scholars, include some government-created content, such as documents that guide federal agencies on the interpretation of the law (J. Selin, personal correspondence, December 18, 2017). FindLaw, a Thomson Reuters subsidiary, provides free online access to Supreme Court documents going back to 1760, alongside other content for legal professionals[60]. Google's Patent Search function facilitates access to digitized and born-digital US patents.[61]

Accountability- and Transparency-Focused Collections

Some not-for-profit organizations make government information freely available and searchable through their own web interfaces, in the interests of transparency and increased government accountability. For example, Guidestar, a non-profit organization that aggregates information about the finances of US-based charitable organizations, provides public access to millions of 990 tax forms filed with the IRS.[62] The non-profit GovTrack aggregates the text of bills and other legislative information in a searchable interface based on structured data from government and other sources.[63] The Internet Archive runs one of the most ambitious web harvesting programs with its End of Term Web Archive (EOT), which aims to capture government web pages when they are most vulnerable to destruction or alteration: when presidential administrations change. The EOT has captured millions of government web pages over the course of three crawls in 2008, 2012 and 2016.

Organizations such as OpenSecrets (for campaign finance data), VoteSmart (for elected officials' voting records), and Muckrock (for providing archival information obtained via FOIA and Sunshine requests) aim to increase government accountability by providing convenient access to collections of government information that is not readily available from other sources. The CRS Report Archive, hosted by the University of North Texas, also addresses the challenge of providing permanent public access to a class of government information that is otherwise only available in an ad hoc manner. As the Archive's website explains, "The Congressional Research Service (CRS) does not provide direct public access to its reports, requiring citizens to request them from their Member of Congress." These reports end up hosted on the websites of individual congressional representatives or non-profit organizations, leaving them vulnerable to loss over time, especially as elected officials leave office. At least two other non-governmental organizations also aim to provide aggregated collections of CRS reports.[64] Sites like these have the effect of putting pressure on government agencies to be more transparent and open with their

58 See https://home.heinonline.org/.

59 See https://www.proquest.com/products-services/legislativeinsight.html and https://www.proquest.com/libraries/academic/databases/proquest-congressional.html.

60 See https://caselaw.findlaw.com/court/us-supreme-court.

61 See https://www.google.com/advanced_patent_search.

62 See https://www.guidestar.org/Home.aspx.

63 See https://www.govtrack.us/.

64 See https://crsreports.com/ and https://www.everycrsreport.com/.

data. For example, the presence of Guidestar heavily influenced the IRS' decision to make similar data available in bulk, and public access to CRS reports was recently included in the government's omnibus spending bill (Jacobs, 2018). The EDGAR Database of forms filed by public companies, now hosted by the Securities and Exchange Commission (SEC), was created in the early 1990s by technologist and activist Carl Malamud because the SEC did not provide convenient public access to such information at the time. The Free Law Project, which advocates for the elimination of fees for delivery of court information through PACER, also maintains an archive of the content it has paid to access (or has been donated by others), around 100,000 files each day.[65] They merge files received from PACER with their collection of oral argument recordings (totaling 1.4 million minutes). The Free Law Project also maintains a browser plug-in that automatically adds PACER files a user purchases to the organization's public archive.

Data Rescue Initiatives

The 2016 presidential election ushered in considerable concern about public access to government data and fueled interest in more aggressive preservation of federal government information. Data rescue initiatives aim to "preserve"[66] data that they consider vulnerable to loss, alteration, or destruction. Vulnerability is determined by a variety of criteria, though many organizations specifically cite threats to data integrity by the current presidential administration. The Sunlight Foundation recently reported that, "despite widespread concern, we do not have evidence that *data* has been removed from federal websites after a year into the Trump presidency" (emphasis added). However, they have documented "substantial removals and overhauls of web pages, documents, research, and websites, reducing public access to public information across the federal webspace" (Sunlight Foundation, 2018). The Environmental Data and Governance Initiative (EDGI) also runs a website monitoring program and released a report documenting the ways in which the Trump administration has removed or obscured information on government websites, particularly information about climate change.[67] Like many of the data collections cited in this report, data rescue repositories may not equate to preservation repositories.

A range of efforts materialized specifically in response to threats to the preservation of climate change data under the Trump administration (Janz, 2018). At least fifteen distinct projects reported that they have collected or are collecting Department of Commerce (DOC) information and data, which includes data generated by NOAA, making it the most frequently targeted agency. Longitudinal climate data is also considered especially vulnerable because it cannot be replicated: should the data disappear, scientists cannot replicate the conditions of the day on which it was collected (Shorish, 2017). The Data Refuge project, one of the most high-profile organizations working in this area, reported hosting 50 events in cities and academic institutions across the country in 2017 to collect and curate endangered government datasets (Janz 2018). It now preserves nearly 400 such datasets or around four terabytes of data in its repository (Sunlight Foundation, 2017).[68] The project's leaders estimate that other data

65 See https://free.law/.

66 This term is defined in different ways by different data rescue initiatives, often meaning "to collect and store" rather than formally adhering to a specific archival or library processes and practices.

67 See http://100days.envirodatagov.org/changing-digital-climate/.

68 See the Data Refuge repository at https://www.datarefuge.org/dataset.

rescue events have captured petabytes more in local repositories and have seeded more than 30,000 URLs into the Internet Archive's WayBack Machine. Crowd-sourced data rescue projects have given attention not only to the technical and logistical challenges of collecting government information, but on ensuring the quality, integrity, and authenticity of the content they ingest (Janz, 2018). Data rescue events received coverage in a range of mainstream media outlets including *NY Times*, *Business Insider*, *Wired*, and CNN. The events have largely tapered off since 2017, though the data archives remain live.

Given resource constraints, selecting the right data to capture and potentially preserve is one of the key responsibilities of data rescue projects. Two key factors include the perceived vulnerability of the data and its perceived significance to researchers or the public at large. Vulnerability is determined by a variety of metrics, including whether data is already duplicated elsewhere (in governmental or non-governmental repositories), whether the data is considered controversial or at risk of being suppressed for political reasons, and whether the data could be replicated if lost or destroyed. Significance may be evaluated based on the information's relevance to specific disciplinary communities, widespread usage of the data by decision-makers, or the impact of conclusions that can be drawn from the data, among other factors. Data rescue efforts therefore require not only preservation professionals, but disciplinary experts, who can evaluate a dataset's significance and advise on its vulnerability. This has led to some creative partnerships, such as a data rescue event focusing on agricultural data organized by JSTOR and the Qualitative Data Repository (QDR) in conjunction with the New York Botanical Garden. [69]

Visualization and Analysis Engines

A comprehensive scan of the consumers of government information is outside the scope of this report. However, some of the individuals and organizations that rely on government information to power their work have become de facto repositories for that information. Journalists, activist organizations, think tanks, scholars, and academic libraries have created a range of visualizations, and analytical tools, and other digital projects that rely largely or exclusively on the availability of government information. University of Pennsylvania Penn Wharton Budget Model (PWBM) maintains a large archive of data from 25 top-level government agencies that powers its USAFacts platform.[70] USAFacts provides a series of dynamic visualizations generated from the underlying data, links to the data sources on government agency webpages, and allows users to download snippets of data as CSV files. However, users cannot retrieve raw data files directly from the repository. The for-profit company Social Explorer gives users even more power to build interactive charts and maps using its store of data from the FBI, the Census Bureau, and other sources (including non-governmental sources).[71] Users can build custom visualizations in order to show, for example, the crime rate in major American cities based on FBI data.[72] The Measures for Justice project, a citizen-driven group, uses public data to assess county-level information about the US criminal justice system.[73]

---

69 See https://qdr.syr.edu/qdr-blog/jstor-and-qdr-partner-organizing-data-rescue-event.

70 See https://usafacts.org/.

71 See http://company.socialexplorer.com/.

72 See https://viz.socialexplorer.com/mapspice/01/00/fbi_crime_rate/.

73 See https://measuresforjustice.org/.

Other projects led by journalists and academics have already done much of the interpretive work, using government data to create digital projects that tell a story or encourage public action. The journalism production company Invisible Institute developed a data-driven project to increase public awareness of policing and encourage accountability.[74] The Torn Apart/Separados project used government data to visualize the reality of family separations at the U.S.-Mexico border.[75]

Library-Based Collections

Although collections of government information curated by libraries could fall into any of the categories described above, libraries' long history of engagement in this work merits its own discussion. Many library projects related to government information are focused primarily on access and are therefore designed to facilitate more convenient, complete, and contextualized engagement with the content produced by government agencies. This mission manifests in a range of library-based projects, from web-based research guides that connect users with sources of government information (including government agencies and non-government initiatives) to involvement in advocacy and awareness projects.[76]

Some academic libraries, particularly Federal Depository Libraries, maintain specialized collections focusing on specific types of government information. For example, Northwestern University holds one of the most comprehensive collections of Environmental Impact Statements, including born digital and digitized documents.[77] The University of North Texas (UNT) maintains the CyberCemetery archive of defunct government websites and also hosts an End of Term Publications digital collection in its institutional repository, consisting of reports, presentations, and documents collected during End of Term web harvesting efforts.[78] Wichita State University Libraries developed and has maintained its Document Data Miner (DDM) system for nearly 20 years, building a unique collection of metadata about the documents available from GPO to Federal Depository Libraries.[79] The Mann Library at Cornell University hosts the Economics, Statistics and Market Information System (ESMIS), a collaborative effort with the USDA. The Mann Library's collaboration with a government agency is not unique. This scan identified several additional examples of such partnerships, both current and defunct. As previously mentioned, the Daley Library at UIC maintained its DOSFAN electronic research collection in partnership with the State Department and, at one time, the Kelvin Smith Library at Case Western Reserve University hosted a mirror of ASCII data from the 2000 Census; the site is currently only available through the Internet Archive's Wayback Machine.[80] The Virtual CDROM/Floppy Disk Library at Indiana University "provides public access and preservation services for the nearly 5,000 CD-ROMs, DVDs, and floppy disks distributed by the GPO under the Federal Depository Library Program (FDLP)."[81]

74 See https://invisible.institute/.

75 See https://xpmethod.plaintext.in/torn-apart/.

76 See, for example, http://libraryguides.missouri.edu/govdocs.

77 See https://libguides.northwestern.edu/environmentalimpactassessment.

78 See https://digital.library.unt.edu/explore/collections/GDCC/ and https://digital.library.unt.edu/explore/collections/EOT/.

79 See http://govdoc.wichita.edu/ddm2/gdocframes.asp.

80 See https://wayback.archive-it.org/8653/*/http://library.case.edu/ksl/census/.

81 See https://webapp1.dlib.indiana.edu/virtual_disk_library/index.cgi/.

Consistent with their mission of enhancing access and use of information, libraries are also developing new ways for users to interact with government information. The now defunct Historical Census Browser, developed by the University of Virginia Libraries, allowed users to analyze and visualize historical census data.[82] More recently, the University of California Berkeley Libraries have launched a protocol that enables researchers to access XML files of the Congressional Record for text mining purposes.[83]

Academic libraries have been at the forefront of initiatives grappling with the future of capturing and providing access to government information in the digital environment, including the Digital Federal Depository Library (LOCKSS-USDOCS) program and the End of Term web harvests.[84]

82 See https://web.archive.org/web/20161230021404/https:/mapserver.lib.virginia.edu/.

83 See https://guides.lib.berkeley.edu/c.php?g=491766&p=4444575.

84 See https://lockss-usdocs.stanford.edu and http://eotarchive.cdlib.org.

Challenges Presented by Government Information

In a summary of the proceedings of a meeting of the Council of State Archivists, Maryland State Archivist and CoSA President Tim Baker cautioned, "We may be looking at a 'gap' of 95% of all we would wish to preserve" (Baker, 2018). Baker cited a range of significant barriers to creating a comprehensive archive of born-digital state government information and records. The same could be said of federal government information, which encompasses an even greater volume and diversity of content in a hodgepodge of format types, disseminated through a larger and more distributed network. The challenges of digital preservation are not unique to government information. Stewards of born-digital materials must grapple with decisions about what gets preserved, given technology and resource constraints; all archives contend with the difficulty of discovering information in distributed repositories; and all archives face the limitations of current web archiving tools, which cannot easily collect files hosted in dynamically queried databases, unless the databases include site maps or structured URLs. This section examines these broad challenges in the context of government information preservation.

The Volume Challenge

The sheer volume of government information presents the most obvious challenge for both access and preservation efforts. Web harvesting and data curation projects have accomplished impressive feats: between Fall 2016 and Spring 2017, End of Term partners archived over 350 terabytes of government websites and data for the EOT archive. In addition, a related effort at the Internet Archive collected an estimated 159 TB of public data from federal FTP file servers (End of Term Archive, n.d.). Brock et al. (2018) reported that this crawl found that "federal agencies host approximately 6,000 websites containing 32 million webpages." Websites represent only a fraction of the corpus of government information. The volume of federal government information in the form of data and publications is also staggering. For example, Data.gov, the most comprehensive database of publicly available government data, includes over 300,000 records, which some researchers believe is still a vast underestimate of extant data sets. Deposit into Data.gov is ad hoc; there is no systematic effort to collect metadata records across agencies or over time. Additionally, Data.gov also only includes records of content that meets a narrow definition of data as "structured information," meaning that its inventory database does not cover vast quantities of important government information, from reports and white papers, to conference presentations, that may warrant long-term preservation and access. Another example of the volume of information confronting preservationists is govinfo, the GPO's repository of government publications, which contains millions of items, representing only a small sliver of government publications: legislative and regulatory information published by GPO, primarily for Congress and the judicial and executive branches. The fastest growing, and most widely used collection held in govinfo is the U.S. Courts Opinion collection, with more than 2.5 million opinions.[85]

Regardless of the exact quantity of government information, the process of preservation remains too onerous to enable a comprehensive archiving effort. For example, the Data Refuge Project's 400 datasets, which represent only a fraction of available environmental data, occupy 4 terabytes of space and took months to collect, describe, and preserve (Wiggin, 2017). The volume challenge is further

85 Information provided by Cynthia Etkin, Senior Program Planning Specialist, US Government Publishing Office (GPO).

exacerbated by the potential for information to change daily or even hourly. One-time captures may quickly become outdated and may fail to make clear when information has been altered or removed, a significant concern for many researchers. Further, the overwhelming amount of information presents challenges for preservation and curation, that is, "the active and ongoing management through its lifecycle of interest and usefulness" (Hudson-Vitale, et al, 2017). Checking the accuracy and comprehensiveness of a collection, for example, may be challenging or impossible. NARA provides the following caveats regarding its periodic web harvests, which might be easily adapted to apply to all data preservation initiatives, warning that "the accuracy of each harvest was affected by these factors: the completeness of URL source lists, whether URLs resolved successfully, and the capabilities of crawler tools used ... and the server environment being crawled" (NARA, 2018b).

The Discovery Challenge

Historically, public government information products followed a relatively predictable route from the authoring agency to the GPO for printing, and into the hands of Federal Depository Libraries and other consumers. Now, researchers and libraries must wade through a tangle of individual agency websites, inter-agency data portals, and other repositories to locate data. The publishing supply chain has radically changed given the ease with which agencies can publish their own content directly to the web. Efforts to create centralized discovery portals, including Data.gov and the Catalog of Government Publications (CGP), have undoubtedly made it easier to locate government information, but they are far from comprehensive. The repositories maintained by individual agencies or groups of agencies do not typically expose their metadata in standards-compliant formats, making it difficult or impossible to inventory their content. Nor do most agencies systematically apply unique identifiers such as digital object identifiers (DOIs) to their publications or datasets. Informally published content (e.g., social media content, pamphlets, blog posts, videos) may not have any unique identifier or permanent link. Official publications may receive one or more unique identifiers. For example, documents distributed by GPO to libraries in the Federal Depository Library Program (FDLP) receive Superintendent of Documents Classification (SuDoc) numbers. Government agencies may also have internal cataloging conventions and may assign standardized identifiers that include an agency abbreviation, year, and publication number (e.g., GAO-18-###). However, there is no mechanism for external consumers to access these identifiers as a raw dataset.

The publishing supply chain has radically changed given the ease with which agencies can publish their own content directly to the web.

The challenges of discovering government information are only magnified when it comes to locating data that has been preserved or aggregated by entities other than the original creator. In particular, non-government information initiatives replicate some of the same weaknesses of the current decentralized system of data dissemination. These initiatives generally focus on small subsets of information products (e.g., all data from a single agency or selected datasets from multiple agencies) rather than wholesale and systematic duplication. They tend to favor numeric data, rather than documents and other media. And they base their collection efforts on the perceived vulnerability of the information or on its relevance to the organization's audience. This is an approach often determined by

necessity, as the sheer volume of government information (as discussed above) makes responsible data mirroring of an entire agency's output daunting if not impossible. This approach ensures that certain datasets persist into the future and retain their integrity. However, future researchers may be left to reconstruct their origin and context.

There is no equivalent to Data.gov for locating government data that lives in the distributed network of institutional and disciplinary repositories, vendor databases, and other websites. This data therefore remains difficult to locate for those not in-the-know. Data Refuge estimates that data rescue events have captured petabytes of government data in institutional or disciplinary repositories (Wiggin, 2017). However, this scan failed to organically identify any federal government datasets in institutional repositories through internet searches. One of the major goals of data rescue initiatives is to ensure that researchers retain access to government data that is lost, altered, or destroyed. Yet, researchers are likely to begin their search for government information through government channels. Unless they already know what they are looking for, how will these researchers identify information or datasets that are missing or altered? Researchers who visit government websites may not realize they are seeing incomplete catalogs; they may not take the time to compare versions downloaded several months apart.

The Technology Challenge

While digital preservation of tabular datasets and static PDFs has become relatively commonplace, preservationists remain ill-equipped to deal with the increasing prevalence of multimedia and dynamic content produced by government agencies. As illustrated in Figure 1, the 350 million URLs harvested during the 2016 End of Term web archiving project, which may contain multimedia and dynamic information, embedded social media feeds, and hyperlinks in addition to text-based content, vastly overshadow the 2.5million titles GPO manages, including those hosted in govinfo, hosted on GPO servers, or made available through partnerships and cooperative agreements with other agencies. Many agencies have adopted at least one (and often multiple) social media channels and produce more informal content than ever before. This content not only poses challenges due to its formats, but lacks the kinds of built-in metadata of more formal publications and these commercial services can severely limit access to content hosted on their platforms. Formal publications typically list a title, author, publication date, and version. Blog posts, social media posts, videos, and visualizations may not have a cited creator and may be easily altered or destroyed without notice. Descriptive metadata that allows for curation and access must be created manually. Website content may change daily or even hourly, and web technologies allow for the personalization of content displayed to a given user based on their Location, device, or other characteristics. If government websites were to implement personalization technologies, a version of record for a webpage could be increasingly difficult to determine. The transition from a text-based web to a dynamic one disrupts traditional digital preservation methods, which rely on static packages of information.

Verifying authenticity and integrity becomes increasingly important and challenging when dealing with electronic publications. GPO has recognized this problem for several decades, writing in 1996, "Due to the ease in which it currently is possible to manipulate electronic source files, the obligation to provide long range assurances of authenticity will become increasingly important as more Government

information moves to electronic formats" (GPO, 1996). GPO authenticates all content held in govinfo (and its predecessor FedSys) and continues to explore next generation practices, processes, and tools. GPO has taken steps toward ISO 16363 certification as a Trusted Digital Repository (TDR). If its certification is approved, it would become the first government-run TDR. While certification as a TDR is an important part of building the systems needed for long-term preservation, it is an expensive and time-consuming process that involves developing technical and policy infrastructure. In July 2018, the Big Data Interagency Working Group (IWG) recommended that agencies expand the use of trustworthy digital repository certification (Big Data Interagency Working Group, 2018).

Conclusion

The web has caused a radical shift in how government information is created, disseminated, and archived. It has upended traditional deposit protocols in the United States that allowed central agencies to keep an official inventory of government content, and the distribution mechanisms that deliver that content into the hands of consumers.

Preserving government information is a long-term responsibility that requires ongoing coordination between technologists, librarians, archivists, and disciplinary experts, both within and outside the government. Collaboration between stakeholders is essential to ensure that information is preserved responsibly, safeguarded against technological or other failures, and broadly discoverable and accessible. Given the immense volume of digital government information, developing preservation priorities requires consultation between digital archivists and disciplinary experts, who can identify the most important information sources for their fields. Collaboration with technologists ensures that preservationists can adopt the most robust methods for curating new media content. And a holistic community effort is necessary to ensure that the failure or sunsetting of individual projects does not lead to catastrophic losses. The breakout success and vibrance of data rescue initiatives following the 2016 elections demonstrates the enthusiasm and resourcefulness of this community, but the quick demise or stagnation of some of these efforts also demonstrates the sustainability risks the resulting collections may face. Even highly popular, established services, such as the University of Virginia Libraries' Historical Census Browser, may not survive budget cuts, staff turnover, or changing administrative priorities, leaving them vulnerable to loss or degradation over time.

There is a clear need for the implementation of open standards and interoperable metadata that force accountability by empowering external organizations to keep track of and analyze the full range of government information.

Despite the many accomplishments of non-government initiatives, the greater issue of preserving and providing long-term access to government information cannot be sufficiently addressed through a series of uncoordinated efforts. Going forward, several pressing needs stand out. There is a clear need for an organizing body to leverage the existing distributed infrastructure and expertise into a coordinated effort by setting priorities, staying aware of imminent threats, parceling out responsibility, and advocating for greater government information transparency. There is a clear need to reevaluate government policies, regulations, and legislation regarding the deposit of digital information products with federal publishing and record-keeping agencies. The disruption of historical information supply chains (in which documents reliably migrated from the creator through the GPO to FDLP libraries to the general public or from the creator to NARA to the general public) presents grave threats to long-term preservation and access. Each day that this system remains broken represents the potential loss of massive quantities of historically significant information. There is a clear need for the implementation of open standards and interoperable metadata that force accountability by empowering external organizations to keep track of and analyze the full range of government information. Government

agencies should be held responsible for producing and disseminating their information products in standards-compliant formats and repositories (Jacobs, 2018b).

The challenges are great, but the risks of failing to address them are profound. The American public stands to lose access to decades worth of historically and scientifically significant information. The actions that the federal government, and the non-governmental organizations addressing these issues, take will have consequences for future generations of scholars, scientists, and the public at large.

Acknowledgements

Thanks to James R. Jacobs, Shari Laster, Marie Concannon, and Scott Matheson for their invaluable guidance throughout the process of researching and writing this report. Thanks to Katherine Skinner of the Educopia Institute for connecting me with Alex Chassanoff and Sam Meister, also of the Educopia Institute. Alex's firsthand experience with data rescue initiatives and Sam's pointers about digital projects powered by government data improved the scope and depth of this report.

Appendix I. Inventory of Efforts

This living appendix inventories each of the organizations and initiatives identified in this environmental scan. The first tab of the spreadsheet concerns collected government information managed by non-governmental organizations. This list is not comprehensive, but aims to be representative of the diversity of projects and organizations working in this space. The second tab provides details on government-run repositories, information portals, and catalogs. Again, this inventory is far from complete, but gives a sense of the complexity of the federal government information ecosystem and the many channels consumers must navigate to access federal information.

https://docs.google.com/spreadsheets/d/1P6oBRNFcDSYThzOqmeehBxI8hVJGxaMc6u26pHiDIag/edit#gid=358697030

Works Cited

Ashlock, Philip and Williams, Rebecca. (2015, Mar. 18) “Open with Apps.” *Meta - The Data.gov Blog.* Retrieved from https://www.data.gov/meta/open-apps/.

Baker, Tim. (2018). “Update: Preserving Electronic Government Information (PEGI).” *Council of State Archivists.* Retrieved from https://www.statearchivists.org/connect/blog/2018/02/update-preserving-electronic-government-information-pegi/.

Big Data Interagency Working Group. (2018). “Measuring the Impact of Digital Repositories.” Washington, D.C.: National Science and Technology Council. Retrieved from https://www.nitrd.gov/pubs/BD-IWG-Digital-Repository-Recommendations-2018.pdf.

Brock, Marieke Lewis, David, Katarina, and Miller, Patrick M. (2018). “Disseminating and Preserving Digital Public Information Products Created by the U.S. Federal Government: A Case Study Report.” Washington, D.C.: Federal Research Division, Library of Congress. Retrieved from https://www.fdlp.gov/file-repository/about-the-fdlp/superintendent-of-documents-policy-statements/3532-final-gpo-frd-digital-case-studies-082218.

Browdie, Brian. (2016, Oct. 14). “The cost of electronic access to US court filings is facing a major legal test of its own.” *Quartz.* Retrieved from https://qz.com/800076/the-cost-of-electronic-access-to-us-court-filings-is-facing-a-major-legal-test-of-its-own/.

Burwell, Sylvia, VanRoekel, Steven, Park, Todd, and Mancini, Dominic J. (2013). “Open Data Policy-Managing Information as an Asset.” Washington, D.C. Retrieved from https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2013/m-13-13.pdf.

Crews, Clyde W. (2017). “How Many Federal Agencies Exist? We Can't Drain The Swamp Until We Know.” Retrieved from https://www.forbes.com/sites/waynecrews/2017/07/05/how-many-federal-agencies-exist-we-cant-drain-the-swamp-until-we-know/#350d85601aa2.

Das, Tara. (2015). Measuring scholarly use of government information: An altmetrics analysis of federal statistics. Government Information Quarterly, 32(3), 246–252. doi.org/10.1016/j.giq.2015.05.002 Retrieved from http://www.sciencedirect.com/science/article/pii/S0740624X1500060X.

DigitalGov. (2018, Jul. 10). “Federal-Compatible Terms of Service Agreements.” Retrieved from https://digital.gov/resources/federal-compatible-terms-of-service-agreements/.

End of Term Web Archive. (n.d.) “Project Background.” Retrieved from http://eotarchive.cdlib.org/background.html.

Federal Committee on Statistical Methodology. (2018). “Federal Statistical Agencies.” Retrieved from https://nces.ed.gov/FCSM/agencies.asp.

Government Printing Office, Superintendent of Documents. (1996). “Study to identify measures necessary for a successful transition to a more electronic Federal Depository Library Program as required by Legislative Branch Appropriations Act, 1996, Public Law 104-53.” Retrieved from https://www.gpo.gov/fdsys/pkg/GOVPUB-GP3-83702f16b5d4a3823308c2c477545669/content-detail.html.

Government Printing Office. (2011). *GPO’s Strategic Plan FY 2011 – 2015, Customer Centric and Employee Driven.* Retrieved from https://www.gpo.gov/docs/default-source/mission-vision-and-goals-pdfs/gpo-strategic-plan-fy-2011-2015.pdf.

Government Publishing Office. (2018). “Mission, Vision, and Goals.” Retrieved from https://www.gpo.gov/who-we-are/our-agency/mission-vision-and-goals.

Govtrack. (2017). "H.R. 4631: Access to Congressionally Mandated Reports Act." Retrieved from https://www.govtrack.us/congress/bills/115/hr4631/summary.

Hudson-Vitale, Cynthia; Imker, Heidi; Johnston, Lisa R.; Carlson, Jake; Kozlowski, Wendy; Olendorf, Robert; Stewart, Claire. (2017). *SPEC Kit 354: Data Curation*. Washington, D.C.: Association of Research Libraries. https://doi.org/10.29242/spec.354.

Jacobs, James A. (2014). Born Digital U.S. Federal Government Information: Preservation and Access. Center for Research Libraries Global Collections Forum. Retrieved from http://www.crl.edu/sites/default/files/d6/attachments/pages/Leviathan%20Jacobs%20Report%20CRL%20%C6%92%20(3).pdf.

Jacobs, James R. (2018, Aug. 24). "GODORT pens thank you letter re CRS reports. LC needs to do this right." *Free Government Information*. Retrieved from https://freegovinfo.info/node/13055.

Jacobs, James R. (2018b, Apr. 26). "FGI at GovInfo Day 2018: how to build a sustainable govt information ecosystem." Retrieved from https://freegovinfo.info/node/12841.

Jacobs, James R. (2018c). "'Issued for Gratuitous Distribution:' The History of Fugitive Documents and the FDLP." *Against the Grain*, *29*(6). Retrieved from https://purl.stanford.edu/yc376vd9668.

Janz, Margaret. (2018). Maintaining Access to Public Data: Lessons from Data Refuge. *Against the Grain*. Retrieved from https://doi.org/10.31229/osf.io/yavzh.

Johnson, Eric and Kubas, Alicia. (2018). "Spotlight On Digital Government Information Preservation: Examining The Context, Outcomes, Limitations, And Successes Of The Datarefuge Movement." *In the Library with the Lead Pipe.* Retrieved from http://www.inthelibrarywiththeleadpipe.org/2018/information-preservation/.

Latham, Bethany. (2018). *Finding and Using U.S. Government Information: A Practical Guide for Librarians.* Rowman & Littlefield Publishers.

Lazorchak, Butch. (2015). "Introducing the Federal Web Archiving Working Group." *The Signal* (blog). Retrieved from https://blogs.LC.gov/thesignal/2015/02/introducing-the-federal-web-archiving-working-group/.

Lewis, David E. and Selin, Jennifer L. (2012). *Sourcebook of United States Executive Agencies.* Vanderbilt University. Retrieved from https://www.acus.gov/sites/default/files/documents/Sourcebook%202012%20FINAL_May%202013.pdf.

Library of Congress. (2015). "Strategic Plan 2016-2020." Retrieved from https://www.LC.gov/portals/static/about/documents/library_congress_stratplan_2016-2020.pdf.

LOCKSS. (2018). "Digital Federal Depository Library Program." Retrieved from https://www.lockss.org/community/networks/digital-federal-depository-library-program/.

NARA. (n.d.) "Frequently Asked Questions about Records Management in General." Retrieved from https://www.archives.gov/records-mgmt/faqs/general.html.

NARA. (2018a). "Strategic Plan 2018-2022." Retrieved from https://www.archives.gov/about/plans-reports/strategic-plan/strategic-plan-2018-2022.

NARA. (2018b). "Congressional and Federal Government Web Harvests." Retrieved from https://www.webharvest.gov/.

Nye, Valerie. (2017). "Preserving Government Websites with 'End of Term President Harvest'." *Intellectual Freedom Blog*. Retrieved from https://www.oif.ala.org/oif/?p=9005.

Office of the Superintendent of Documents. (2018). "GPO's System of Online Access Collection Development Plan." Retrieved from https://www.fdlp.gov/file-repository/about-the-fdlp/gpo-projects/trustworthy-digital-reports/3620-final-systemcolldevplan-09282018.

Open Government Initiative. (n.d.) Retrieved from: https://obamawhitehouse.archives.gov/open.

Public Printing and Documents, 44, U.S.C. §§ 301–318. 2008. Retrieved from https://www.gpo.gov/fdsys/pkg/USCODE-2008-title44/html/USCODE-2008-title44.htm.

Scarcella, Mike. (2018, Mar. 31). "Federal Judiciary Misused PACER Fees, Judge Says in Class Action Ruling." The National Law Journal. Retrieved from https://www.law.com/nationallawjournal/2018/03/31/federal-judiciary-misused-pacer-fees-judge-says-in-class-action-ruling/?cmp=share_twitter&slreturn=20181030101740.

Selin, Jen. (2018, Dec. 18). Interview with Marie Concannon.

Sherman, Andrew M. (2018). "Prepared Statement before the Subcommittee on Legislative Branch Appropriations Committee on Appropriations U.S. House of Representatives On GPO's Appropriations Request For FY 2019." Washington, D.C.: Government Publishing Office. Retrieved from https://www.gpo.gov/docs/default-source/congressional-relations-pdf-files/testimonies/sherman_house_prepared_statement_4_2018.pdf.

Shorish, Yasmeen. (2017, Feb. 1). "Data Refuge and the Role of Libraries." *ACRL TechConnect*. Retrieved from https://acrl.ala.org/techconnect/post/data-refuge-and-the-role-of-libraries/.

The Sunlight Foundation. (2018). "Tracking US Government Data Removed from the Internet During the Trump Administration." Retrieved from https://sunlightfoundation.com/tracking-u-s-government-data-removed-from-the-internet-during-the-trump-administration/.

The Sunlight Foundation. (2017). "How Data Refuge Works and How You Can Help Save Federal Open Data." Retrieved from https://sunlightfoundation.com/2017/02/06/how-data-refuge-works-and-how-you-can-help-save-federal-open-data/.

US Census Bureau. (2018). "Federal Statistical Research Data Centers." Retrieved from https://www.census.gov/fsrdc.

Wiggin, Bethany. (2017, Feb. 26). "How Data Refuge Works, and How YOU Can Help Save Federal Open Data." *Sunlight Foundation*. Retrieved from https://sunlightfoundation.com/2017/02/06/how-data-refuge-works-and-how-you-can-help-save-federal-open-data/.