Government and Online Data: Creation, Access, Preservation

Richard M. Mikulski
Portland State University, mm8@pdx.edu

Follow this and additional works at: https://pdxscholar.library.pdx.edu/onlinenorthwest
Let us know how access to this document benefits you.

https://pdxscholar.library.pdx.edu/onlinenorthwest/2019/schedule/10

This Presentation is brought to you for free and open access. It has been accepted for inclusion in Online Northwest by an authorized administrator of PDXScholar. Please contact us if we can make this document more accessible: pdxscholar@pdx.edu.
Government and Online Data: Creation, Access, Preservation

NOTICE: Due to a lapse in federal funding portions of this website are not being updated. Learn more.

Data

Rick Mikulski, Portland State University
Online Northwest 3/29/2019
rmm8@pdx.edu
Abstract

From the 1790 Census to the present, the US government has been a major user, producer, and distributor of data. Through its agencies and departments, it creates data; through funding, grants, and data sharing mandates, it makes research data accessible; through the FDLP and various agency platforms, it circulates and stores data online. This presentation discusses the implications of Government’s role in data creation, online access, and preservation. What are the potential strengths and weaknesses of this relationship, and how can librarians prepare for the future of data creation, preservation, and access? Special attention given to: Federal Data Mandates; Online Sources of Government Data; Challenges to Preservation and Access.
Overview: Government & Data

• History of relationship between government and Data creation/curation.

• What this relationship means for researchers and librarians.

• Challenges this relationship poses for preservation.

• The role of librarians in this relationship
Thinking about the relationship
There is no going back.

Ticket closed by Ed Hill.
On Mar 04, 2019 @ 01:41 pm, Ed Hill wrote:

Ticket closed: Set up a Linux VM, installed DOSBox. Mounted the USB disc drive in the Linux VM (Virtualbox).
Made disc images of the floppy discs with `dd bs=512 count=2880 if=/dev/fd0 out=/home/ed/disc1.img status=progress`.
Made empty directory and mounted disc 1 with `mount -o loop disc1.img /home/ed/install`.
Made empty directory for dosbox data.
In Dosbox mounted data directory to f: with `mount f /home/ed/dosboxdata`.
Mounted disc image to a: with `mount a /home/ed/install -floppy`.
When it asks, for disc 2, unmount disc 1 with `umount /install` and mount disc2 with same command as above.
Resume install.
Once install is complete, Excel should be able to open database files and Word should be able to open WordPerfect files.

For a full-on version of this, you would have something like:

- A (preferably Linux, to my mind) station with a floppy drive to make disk images of the floppy images. Each disk only takes a couple minutes to image. Ideally this would also involve checking the files in the image for bit-level accuracy. These images would be duplicated to an archive and a copy for processing. Ideally, if you were doing this for real, you would have a forensic hardware block in place so that you couldn’t accidentally overwrite the data on the disks, but that’s not strictly necessary.

- A (again, preferably Linux) station with Dosbox. The workflow for multi-disk installs would be:
  1. Create a directory in Linux for dosbox data.
  2. Create a directory in Linux to mount disk images. Mount image 1.
  3. In Dosbox, mount the data directory, mount the disk directory as -floppy
  4. Change to the disk directory and initiate install.
  5. For each requested disk image change, in Linux unmount the current disk and mount the next disk to the same Linux directory.
  6. Repeat 5 as needed until fully installed.
A [brief] history of Government & Data

1790 Census

1960s IRS “Brain Machines”
A brief history of Government & Data


Data.gov is launched in 2009 as an informal partnership with various agencies. Begins with 47 datasets. As of 3/28/19 there are 236,345.
February 22, 2013

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: John P. Holdren
Director

SUBJECT: Increasing Access to the Results of Federally Funded Scientific Research

I. Policy Principles

The Administration is committed to ensuring that, to the greatest extent and with the fewest constraints possible and consistent with law and the objectives set out below, the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community. Such results include peer-reviewed publications and digital data.

Scientific research supported by the Federal Government catalyzes innovative breakthroughs that drive our economy. The results of that research become the grist for new insights and are assets for progress in areas such as health, energy, the environment, agriculture, and national security.

Access to digital data sets resulting from federally funded research allows companies to focus resources and efforts on understanding and exploiting discoveries. For example, open weather data underpins the forecasting industry, and making genome sequences publicly available has
“The Office of Science and Technology Policy (OSTP) hereby directs each Federal agency with over $100 million in annual conduct of research and development expenditures to develop a plan to support increased public access to the results of research funded by the Federal Government.”

Department of Agriculture,
Department of Commerce,
Department of Defense,
Department of Education,
Department of Energy,
Department of Health & Human Services,
Department of Transportation,
Department of Veterans Affairs,
Environmental Protection Agency,
National Aeronautics & Space Administration,
National Science Foundation,
Office of the Director of National Intelligence (ODNI),
Smithsonian Institution,
US Agency for International Development,
US Geological Survey
Thorny Issues

Public Data vs. Open Data
- “Public” can be requested or obtained.
- “Open” is readily accessible and should be machine readable.

Federal Data Vs Federally-funded Data
- “Federal Data” is produced/collected by the government.
- “Federally-funded Data” is final research data produced in a funded study.

Terminology
- “Final Research Data,” “timely,” “reasonable cost” not defined.
Data Management Plans

b) Ensure that all extramural researchers receiving Federal grants and contracts for scientific research and intramural researchers develop data management plans, as appropriate, describing how they will provide for long-term preservation of, and access to, scientific data in digital formats resulting from federally funded research, or explaining why long-term preservation and access cannot be justified;

c) Allow the inclusion of appropriate costs for data management and access in proposals for Federal funding for scientific research;

d) Ensure appropriate evaluation of the merits of submitted data management plans;

e) Include mechanisms to ensure that intramural and extramural researchers comply with data management plans and policies;

f) Promote the deposit of data in publicly accessible databases, where appropriate and available;

g) Encourage cooperation with the private sector to improve data access and compatibility, including through the formation of public-private partnerships with foundations and other research funding organizations;

h) Develop approaches for identifying and providing appropriate attribution to scientific data sets that are made available under the plan;
H.R.4174 - Foundations for Evidence-Based Policymaking Act of 2017 (signed 1/14/19)

Summary: H.R.4174 — 115th Congress (2017-2018)

This bill requires departments and agencies identified in the Chief Financial Officers Act to submit annually to the Office of Management and Budget (OMB) and Congress a plan for identifying and addressing policy questions relevant to the programs, policies, and regulations of such departments and agencies. The plan must include: (1) a list of policy-relevant questions for developing evidence to support policymaking, and (2) a list of data for facilitating the use of evidence in policymaking.

The OMB shall consolidate such plans into a unified evidence building plan.

The bill establishes an Intergency Council on Evaluation Policy to assist the OMB in supporting government-wide evaluation activities and policies. The bill defines “evaluation” to mean an assessment using systematic data collection and analysis of one or more programs, policies, and organizations intended to assess their effectiveness and efficiency.

Each department or agency shall designate a Chief Evaluation Officer to coordinate evidence-building activities and an official with statistical expertise to advise on statistical policy, techniques, and procedures.

The OMB shall establish an Advisory Committee on Data for Evidence Building to advise on expanding access to and use of federal data for evidence building.

Open, Public, Electronic, and Necessary Government Data Act or the OPEN Government Data Act

This bill requires open government data assets to be published as machine-readable data.

Each agency shall: (1) develop and maintain a comprehensive data inventory for all data assets created by or collected by the agency, and (2) designate a Chief Data Officer who shall be responsible for lifecycle data management and other specified functions.

The bill establishes in the OMB a Chief Data Officer Council for establishing government-wide best practices for the use, protection, dissemination, and generation of data and for promoting data sharing agreements among agencies.
Overall Historical Trend

- Government has produced and collected data for centuries.
- Government has Computerized their data since the mid 20th century.
- Electronic government data circulated physically after 1980s via FDLP.
- Government data widely accessible online by late 90s
- Government began setting data standards in 2013 through funding.
- Government requiring agency data to be machine readable in 2019.
Part II
What does this relationship mean for librarians?

- Government agencies produce incredible amounts of data, which our users want/need to access.
- This centralization has resulted in an irreversible dependence.
- Government funding requirements is slowly setting standards.
- These standards are increasingly making librarians key to grant funding, as we assist with DMPs and archiving.

It is within this environment that Librarians must...
- 1) Provide access to data.
- 2) Work to preserve government data.
- 3) Assist researchers in preserving their own data.
Postings on ALA Joblist with “data” in the title
March 2018 – March 2019
(Collected by rmm for future project)
Providing Access to Government Data

(Social Explorer)
Threats to Access and Preservation

Privatization

Obsolete Technology (or poor planning)

Reliance on Federal Agencies
Efforts to Preserve Federal Data
Preserving Research Data

**Data Management**

**What is Data Management?**

Data management covers all steps in the collection, processing, storing and sharing of research data, including:

- Processing and analyzing your data
- Verifying and documenting your data
- Data formats, naming, and organization
- Data archiving and sharing
- Ethics and attribution of data

The Library can help you with data management planning and archiving of research data; Research Computing can help with storage and processing needs while you are working on your project.

**Data Management Services**
Data Management Planning

https://dmptool.org/
Government policy directly shapes data curation, preservation, and access

- Through agencies, it produces data, which is increasingly accessible (only!) online.
- Through funding it is requiring federally-funded data to be freely accessible online.
- By implementing Data policies and setting the criteria, it is setting the standards.
- For better or worse, librarians and researchers are largely dependent upon the online infrastructure the government has set in place for continued access.
If we keep active, it’ll be fine.
Questions?

Rick Mikulski, rmm8@pdx.edu
Acknowledgements
Special thanks to Ed Hill and Sho Ikeda for their work and patience regarding my floppy disk antics. Colleagues on the Govdoc-L list (Susanne Caro, Jim Jacobs, Brooke Johnson, Julia Cater) similarly assisted in gathering material on archiving outdated FDLP data.

Selected References
https://www.datarefuge.org/
https://webapp1.dlib.indiana.edu/virtual_disk_library/
https://shsulibraryguides.org/c.php?g=622180&p=5782944
https://library.pdx.edu/services/data-management/
http://guides.library.pdx.edu/data
https://freegovinfo.info/
https://dmptool.org/
https://openaccess.unt.edu/denton-declaration
https://sparcopen.org/open-data/

Government and Online Data: Creation, Access, Preservation by Richard M. Mikulski is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.