Collecting and preserving born-digital government information and data

A holistic view from librarians leading grassroots efforts

Association of Public Data Users
Tuesday July 17, 2018, 2:15pm - 3:30pm
ROGUE SCIENTISTS RACE TO SAVE CLIMATE DATA FROM TRUMP
Who are we?

Deborah Caldwell // Graduate Research Assistant, PEGI Project, University of North Texas

James R. Jacobs // US Govt Information Librarian, Stanford University // @freegovinfo

Margaret Janz // Data Curation Librarian, University of Pennsylvania // @MargaretJanz

Shari Laster // Head, Open Stack Collections, Arizona State University // @rhonabwy
Agenda

- Background & context (Deborah)
- Independent library and consortial efforts (Deborah and James)
- Cooperative and grassroots projects (James and Margaret)
- Moving the conversation forward (Shari)
- Q&A
LIS Context

FDLP Libraries

The United States Federal Government is a critical source of information for researchers seeking data on the American population. From population counts to energy use to economic analysis, the United States Federal Government collects data and disseminates information that, often, no other private entities replicate; birth and death reports and Census questionnaires, for example, are legally mandated by the federal government, and thorough and accurate collection of these data by private institutions would be difficult to impossible.
LIS Context: FDLP Libraries
LIS Context: FDLP Libraries

Congress, Federal Agencies, etc

GPO
LIS Context: FDLP Libraries

Congress, Federal Agencies, etc

GPO

FDLP
Federal Depository Library Program
LIS Context: FDLP Libraries

Congress, Federal Agencies, etc

GPO

FDLP

Public
LIS Context: FDLP Libraries
LIS Context: FDLP Libraries
LIS Context: FDLP Libraries

Congress, Federal Agencies, etc
LIS Context: FDLP Libraries

Congress, Federal Agencies, etc

GPO

FDLP

Federal Depository Library Program
LIS Context: FDLP Libraries

Congress, Federal Agencies, etc.
LIS Context: FDLP Libraries
LIS Context: FDLP Libraries

Congress, Federal Agencies, etc

GPO

FDLP

Federal Depository Library Program

???

???
LIS Context

Since digital methods of information dissemination have risen to prominence, librarians, archivists, and government transparency advocates have been concerned with how this transition has affected existing discovery mechanisms and dissemination workflows.

In recent years, concern about the preservation and long-term accessibility of digital government information have grown, and 2016 and 2017 saw a concerted increase in conversations around these topics within the library and archives communities.
LIS Context

Various efforts have formed with the goal of ensuring long-term accessibility of these data for future researchers, the primary concern being that because a majority of these data are collected through mandate by the federal government, collection of these same data by private institutions would not be possible.
LIS Context

Various efforts have formed with the goal of ensuring long-term accessibility of these data for future researchers, the primary concern being that because a majority of these data are collected through mandate by the federal government, collection of these same data by private institutions would not be possible.
LIS Context

Various efforts have formed with the goal of ensuring long-term accessibility of these data for future researchers, the primary concern being that because a majority of these data are collected through mandate by the federal government, collection of these same data by private institutions would not be possible.
LIS Context

Various efforts have formed with the goal of ensuring long-term accessibility of these data for future researchers, the primary concern being that because a majority of these data are collected through mandate by the federal government, collection of these same data by private institutions would not be possible.

Library participates in effort to preserve government data
Sorry, the page you requested has either been moved or is not available on this server.

May we suggest these Gateway pages to census data:

- The 2010 Census
- American FactFinder
- Business Help Site
TAKE NOTICE: This Web site, AHRQ’s National Guideline Clearinghouse, will not be available after July 16, 2018.

Federal funding through AHRQ will no longer be available to support the NGC as of that date. For additional information, read our full announcement.

We will continue to post summaries of new and updated evidence-based clinical practice guidelines until July 2, 2018. For any questions, please contact Mary Nix, mary.nix@ahrq.hhs.gov.

https://www.guideline.gov
Libraries and digital data preservation

UNT digital library / CyberCemetery

Dataverse / UNT Data Archive / Data.stanford.edu

OpenGeoPortal / Earthworks.stanford.edu

ICPSR (DataLumos, OpenICPSR, US Census data repository, NACJD)
About This Collection

The CyberCemetery is an archive of government web sites that have ceased operation (usually web sites of defunct government agencies and commissions that have issued a final report). This collection features a variety of topics indicative of the broad nature of government information. In particular, this collection features web sites that cover topics supporting the university's curriculum and particular program strengths.

Latest Additions
http://www.library.unt.edu/scholarly-works/data-repository
https://data.stanford.edu
https://dataverse.org
Geoportals

http://data.opengeoportal.org

https://earthworks.stanford.edu
Cooperative and grassroots projects

- LOCKSS-USDOCS
- End of Term Archive
- DataRefuge
Digital Federal Depository Library Program

The Digital Federal Depository Library Program, or the "USDocs" private LOCKSS network replicates key aspects of the United States Federal Depository System. The content is held in geographically distributed sites and replicated many times. Citizens have oversight and responsibility for the long-term care and maintenance of the content. All these characteristics mean the content will be preserved so that any alteration of the content (either deliberate or accidental) will be detected and repaired. For all documents, preservation in a tamper-evident environment is important, but for government documents, this is essential.

Our current focus is harvesting and preserving collections hosted on the The U.S. Government Publishing Office's (GPO) Federal Digital System (FDsys). See also, Digital Federal Depository FAQ

https://lockss-usdocs.stanford.edu
https://www.govinfo.gov
LOCKSS-USDOCS

- 36 libraries in a collaborative preservation network (map)
- Collecting and preserving copies of all collections on https://Govinfo.gov
- Added assurance to information published by GPO
- Primarily Congressional information
- Preserved w FDLP principles and open source software
- http://lockss-usdocs.stanford.edu
End of Term Web Archive
US Federal Web Domain at Presidential Transitions

Project Background

Project Partners

Browse by Timeframe:
- 2008-2009
- 2012-2013
- 2016-2017
- Browse All

Browse Web Archive

EOT at a Glance
Crawl dates:
- September 2008 to May 2009
- September 2012 to March 2013
- September 2016 to March 2017
Number of websites captured:
- 2008: 3,305 websites
- 2012: 3,273 websites
- 2016: 53,324 websites

Archive Scope
The End of Term Web Archive contains federal government websites (.gov, .mil, etc) in the Legislative, Executive, or Judicial branches of the government. Websites that were at risk of changing (i.e., whitehouse.gov) or disappearing altogether during government transitions were captured. Local government websites, or any other site not part of the federal government domain were out of scope.

http://eotarchive.cdlib.org
Volunteer seed contributions

End of Term Presidential Harvest 2016

2008: 457 from 26 nominators
2012: 1476 from 31 nominators
2016: 15,000+ from 400+ nominators (via UNT form)
Plus!: Over 100,000 from DataRescue/EDGI events/tools
EOT 2016 results

~300 TB data total!

~110 TB web crawls + ~130 TB of gov ftp site archiving
310,000,000 web URLs + 12,000,000 ftp files
9,000+ social media accounts (scrape of gov SM registry API) 44% FB, 37% TW, 10% YT

[http://eotarchive.cdlib.org](http://eotarchive.cdlib.org)
ABOUT NEW YORK

Harvesting Government History, One Web Page at a Time

In conversation with many partners, such as you, we can build refuge for federal climate action. The fact of ongoing climate change. We are c
Preservation of Electronic Government Information (PEGI) is a two-year initiative to address national concerns regarding the preservation of electronic government information by cultural memory organizations for long term use by the citizens of the United States.

The focus is at-risk government digital information of long term historical significance which is not being adequately harvested from the Web or by other automated means.
Partner Institutions

- University of North Texas
- University of North Carolina at Greensboro
- Arizona State University
- University of Missouri
- Stanford University
- Yale University
- Center for Research Libraries
- Educopia

Year two of the project (2018) is made possible in part by the Institute of Museum and Library Services grant #LG-88-17-0129-17.
Goals

1. Conduct a multimodal scan of at-risk federal digital content.

2. Produce guidance for agencies and other government information producers, taking the form of recommendations for best practices that can improve the preservability and accessibility of electronic information.

3. Conduct an educational awareness and advocacy outreach program.

4. Develop recommendations for a collaborative agenda for future work to continue improving preservation of electronic government information.
Thank you.

Questions? Contact us:

- www.pegiproject.org
- info@pegiproject.org
- @PEGIProject