Preservation of Electronic Government Information Project (PEGI)

Overview

Librarians, technologists, and other information professionals from the Center for Research Libraries, the Government Publishing Office (GPO), the University of North Texas, the University of California at Santa Barbara, the University of Missouri, and Stanford University are undertaking a two year project to address national concerns regarding the preservation of electronic government information (PEGI) by cultural memory organizations for long term use by the citizens of the United States. The PEGI project has been informed by a series of meetings between university librarians, information professionals, and representatives of federal agencies, including the Government Publishing Office and the National Archives and Records Administration.

There is now a growing awareness nationally of the serious ongoing loss of government information that is electronic in nature. This issue has loomed larger in recent years, reaching a point of criticality that drove intense discussions in recent meetings of information management thought leaders from across the United States. Summits held in April and December of 2016 sought to explore ways of undertaking urgently needed cross-sector activities to preserve and provide access to electronic government information. The first event was entitled The Digital Preservation of Federal Information Summit (DPFIS) and brought together stakeholders from more than two dozen public, private, and federal organizations, including archivists, librarians, technologists, program officers, executive directors, and other interested parties.[1] The meeting sought to engage these national leaders in a structured, facilitated dialogue on at-risk digital government records and information, with an aim to explore the development of a national agenda to address the preservation and access of priority content in this area. The second event was held in December 2016 in Washington D.C., and continued the discussions between many of the participants from the first summit, leading to the conviction that action should be undertaken through a project now by means of this project. These two meetings built on extensive prior discussions of the issues conducted at *CRL's 2014 Global Resources Collections Forum*, which ultimately resulted in the *Leviathan Report*[2] summarizing serious threats to the long-term integrity and accessibility of electronic government information.

The core problem discussed in all these events can be summarized as follows. In the pre-digital production era, a clear workflow accounted for the preservation of most government records and information. Federal agencies created the content and when that content was ready to be disseminated or archived, appropriate

print material was sent to NARA, GPO, and depository libraries. These workflows were effective in the print era. However, today, most government information is produced and disseminated digitally. Digital workflows are neither as predictable nor systematic as print workflows; further, the number of publications has exploded.[3] Individual federal agencies now have the ability to quickly and easily publish their work themselves, without involving the GPO. So long as the work is not categorized officially as a "record" or a "report" or a "publication," regulatory authority does not require agencies to maintain content themselves, nor to provide it to the GPO or NARA for ongoing care. Information produced by agencies in digital form should be scheduled as a record or publication, however it often falls through the cracks. The Federal Records Act covers information created or received by agencies in the course of conducting government business, regardless of format. Records that are appraised as permanent should be transferred to NARA, as per publicly published schedules set by each agency with widely varying timeframes for transfer depending upon the record types.[4] Whether that is consistently happening is another issue, one which NARA has regularly questioned in its agency assessment efforts.[5] The resulting electronic circulation of information that falls into the gap between official "records" (NARA) and official "reports and publications" (GPO) arguably serves the short-term interests of U.S. citizens; however, it breaks the chain of custody that has long ensured that government records and information are assessed, selected, and preserved. In other words, this workflow shift has undercut the GPO/NARA-based central pathway to selection and long-term retention and preservation for much of our government's output as necessary for the "National Bibliography", the official information of the nation.

The organizations partnering on this project include several university libraries, the Government Publishing Office, and the Center for Research Libraries, all of which are deeply concerned about and committed to ensuring the long term preservation and access to electronic government information of a critically important and lasting historical value to the citizens of the United States. For years both the University of North Texas and Stanford University have sought to preserve at-risk electronic government information, and regularly participates in the national End-of-Term (EOT) web archiving effort.[6] Further, Stanford coordinates the LOCKSS USDOCS network[7], a key repository for government documents.

At-risk Government Digital Information

The focus of the PEGI project is at-risk government digital information of long term historical significance. Historians and public scholars are now concerned that "the age-old conflict between the interests of historians and those of government agencies has become only more problematic considering the ease with which electronic records can be altered, encrypted, or destroyed." [8] Analysis of electronic

government information indicates that "More than ever before, most born-digital information fits into that broad category of what was once called 'fugitive documents'" which are not preserved through standard mechanisms.[9]

There are far too many potential collections of at-risk government digital information to be addressed in a modest project of the scale that we are currently proposing; indeed, a significant problem is precisely the need for scoping the range of government information resources that are currently being lost.[10] The significance and size of this problem requires that we first identify, understand, and document the scope and range of electronic government information now being lost routinely, and this is the approach the project will focus on.

Following the recommendations of the Leviathan report, this project proposes focusing on activities of triage, drilling down into agency workflows, differentiating the audiences for electronic government information, mobilizing collaborative efforts, and undertaking advocacy and outreach efforts to raise awareness of the importance of preserving digital government information.[11]

To these ends, the project partners propose focusing on making accessible the following collections, and registries of collections:

1. **Agency Information Triage Database:** The project will analyze and assemble data concerning federal agency information of historical significance that are currently not being preserved, together with recommended steps for collaborative efforts between the relevant agencies and cultural memory organizations interested in digital preservation and access efforts. This collection of documents will be informed by the Agency Workflow Collection described below.

2. **PEGI Registry:** A registry of current preservation sites for electronic government information will be created in the course of the project. There is currently no comprehensive registry of such sites and relevant technical information, including application program interfaces for accessing the content preserved through distributed means, formats, systems deployed, etc.

3. **Agency Workflows Collection:** As preparation for creating the Triage Database, information about targeted agency workflows will be gathered and documented in a standardized format. Agencies that will be considered for analysis will begin with the 44 executive and legislative branch agencies identified at "high risk" of information loss by NARA in the 2015 Records Management Self-Assessment.

Project Goals

The PEGI project will undertake several goals to begin to collaboratively address the issue of preserving electronic government information, as follows. These are our initial goals, as of early 2017 and may be expanded as we garner more information to inform the project.

Goal: Comprehensive Environmental Scan

The project will conduct a comprehensive multi-modal environmental scan of at-risk federal digital content. The most priority in the project will be devoted to agencies identified by NARA as being at "high-risk" of information loss, although some attention will be given to all federal agencies. The scan will investigate three aspects of electronic government information content: A) content creation workflows, including fundamental questions about who creates electronic government information within specific agencies, the nature of their electronic content creation workflow, and how electronic content creation differs from traditional print workflows, B) content dissemination workflows and schedules will be closely examined to understand how and when electronic government information becomes at-risk because of non-transmittal to preservation and publication sites such as NARA and GPO, C) content users will be interviewed and analyzed, so that the ways that the information is either potentially or actually used now and in the future by citizens is better understood and documented. In addition to these three areas the report will address general questions and issues related to digital-born information and define categories of digital born information.

Along with these three tracks the project team will look to define what digital content is at-risk and what preservation efforts are currently underway and by whom (federal, state, community efforts). The project team also aims to define what preservation standards exist and how metrics can be applied to preservation efforts to gauge success that collected materials are meeting a need within user communities. Concomitantly, the Government Publishing Office (GPO) will be initiating a synergistic internal study to identify Federal agency digital publishing policies, workflows, practices, personnel, and information products. This internal study should bolster the work of the PEGI project, and the two investigative activities will be fully aligned.

The environmental scan will produce several outputs. A summary report of the findings of the environmental scan will be developed in collaboration with Project Steering Committee and (wherever possible) representatives of the GPO and NARA. As described above, additional major outputs of the environmental scan will include the Agency Triage Database, Agency Workflows Collection, and PEGI Registry.

Goal: Recommendations for Information Creators

Guidance documents will be produced by the PSC and project staff for agencies and other publishers of government electronic information. These documents will take the form of recommendations for best-practices in information lifecycle management that could improve the preservability and accessibility of electronic information produced by the content creator. Recommendations will address gaps in agency workflows that reduce the likelihood of effective preservation and access to electronic government information being produced. These documents will be prepared in a neutral tone and format designed to enable objective conversations and information exchange with agencies regarding the topic of improving preservation and access.

Goal: Educational Awareness and Advocacy Program

An educational awareness and advocacy outreach program will take place in 2018, the second year of the project. The aim of this program will be to start new dialogues on the issues in preserving and providing access to electronic government information, engaging the broader library community, relevant academic and public scholars, federal agencies, and other relevant stakeholders. The program will include an email advocacy campaign and several events, all of which will be interactive discussions with different stakeholder groups and PEGI project participants, held with the aim of raising awareness and advocating for improved preservation and access to electronic government information. The outreach program will include webinars, and panels at the American Historical Association conference, ALA Annual conference, and the annual Depository Library Council meeting.

Goal: PEGI Collaborative Agenda

The project will analyze and develop recommendations for a collaborative agenda for future work to continue improving preservation and access to electronic government information. These recommendations will be vetted and discussed in several public forums -- including both library conferences as well as other wider public fora -- in the course of the project. In particular, the collaborative agenda will take into account the Triage Database findings and the recommendations for information creators. Additional external groups that may wish to participate in the project will be contacted to assess their interest, for example, the DataRefuge initiative [12] and ARL Libraries Network. The collaborative agenda will seek to identify a sequence of next steps that could be collaboratively undertaken to make more electronic government information public, preservable, and preserved in multiple environments that include distributed sites in academic libraries and other heterogeneous locations that are indexed, contextualized and usable.

Plan of Work

This project will take place over two calendar years, from January 1, 2017 through December 31, 2018. The people working on the project will primarily fall into two categories: 1) members of the project steering committee, and 2) project staff.

The **Project Steering Committee** (PSC) will be a small group of well-qualified professionals committed to the aims of the project, who will coordinate their work via a combination of teleconference and in-person meetings (all PSC travel will be self-funded by their institutions). The PSC will include the following individuals: 1) Committee chair will be Martin Halbert, Dean of Libraries, 2) *Roberta Sittel*, UNT Government Information Librarian and department head, works with the UNT Digital Libraries on signification digital archives such as UNT's CRS Archive and CyberCemetery and has facilitated the preparatory activities for this project, 3) *James Jacobs*, Government Information Librarian at Stanford University and founder of Free Government Information (FGI), a blog that advocates for preservation and access of government information. Mr. Jacobs works with the LOCKSS-USDOCS and the End of Term crawls, two projects that work to preserve digital born government information. 4) *Marie Waltz,* Special Projects Manager at the Center for Research Libraries, is responsible for several CRL digital and print archival initiatives, 5) *David Walls*, GPO Preservation Librarian, has participated in all preparatory meetings, 6) *Shari Laster*, Government Information and Data Services Librarian at University of California, Santa Barbara is a past-member and chair of the Depository Library Council (DLC), advocacy arm of the FDLP. Ms. Laster is also an active member of ALA GODORT and an outspoken advocate for access to government information. 7) *Marie Concannon* is the Regional Depository Coordinator and head of the Government Information and Data Archives Research & Information Services Division at the University of Missouri. Ms. Concannon is also a past-member and of DLC and works to coordinate preservation efforts of Missouri state government information. 8) *Scott Matheson* is the current chair of the Depository Library Council and Yale's Associate Librarian for Technical Services. As a member of the American Association of Law Libraries, Mr. Matheson is committed to the preservation and access of legislative and legal materials, 9) Lynda Kellam, Data Services and Government Information Librarian at the University of North Carolina at Greensboro's University Libraries; She is the author of *Numeric Data Services and Sources for the General Data Librarian* (2011), co-editor of *Databrarianship: The Academic Data Librarian In Theory And Practice* (2016), and has presented extensively on data services..

The **Project Staff** will be comprised of government documents specialists. Two librarians and two staff will be contributed by the UNT government documents department, with part-time commitments amounting to roughly 0.5 FTE overall. Two half-time graduate researchers will be hired for the project from the UNT College of Information. These graduate researchers will be selected for their knowledge of government information and ability to conduct field investigative work. The graduate researchers will be supervised by Roberta Sittel. Much of the work of the project staff will be researching and documenting the workflows and practices of federal agencies in producing electronic information.

Year One (2017) Activities: The first year of the project will be focused on gathering data and preparing for the educational awareness and advocacy outreach activities in the second year of the project. In addition to regular email correspondence, the PSC will meet together in the following months and locations during 2017: 1) March, via teleconference, for an initial organizational meeting, 2) May, at the Open Access Symposium in Frisco, Texas, 3) June, at ALA Annual in Chicago, to review environmental scan data gathering activities and begin planning project educational awareness and advocacy outreach activities, 4) October, at Depository Library Council (DLC) in Washington, to review initial project research data and continue planning educational awareness and advocacy outreach efforts. Project staff will begin research efforts in the first half of the year, with three 2-day data gathering trips to Washington to be conducted before the end of the year for interviews and other on-site work in the District of Columbia region. One of these trips will likely coincide with DLC in October for information sharing purposes.

Year Two (2018) Activities: The second year of the project will be focused on analyzing and documenting project findings, and conducting educational awareness and advocacy outreach activities. Again in 2018, the PSC will meet together several times in the following months and locations: 1) late January, at ALA Midwinter in Denver, to do final preparations for educational awareness and advocacy outreach activities and work on editing project documents, 2) June, at ALA Annual in New Orleans, to undertake educational awareness and advocacy outreach presentations, including publication of the environmental scan summary report, and 3) October, at Depository Library Council (DLC) again in Washington, to conduct final outreach presentations and final project completion activities, including publication of the final project report and recommendations for future collaborative work. Project staff will complete research efforts in the first half of the year. One of these trips may be planned to coincide with ALA Annual in June. In addition to the events listed above, the educational awareness and advocacy program will hold targeted discussion panels as follows: 1) early January, at AHA conference in Washington, to engage historians, 2) webinars will be held in March and August to engage as many different stakeholder groups as possible, and 3) in October a public forum panel with federal agency representatives will be held in Washington in conjunction with Depository Library Council. An email advocacy campaign will take place to contact as many stakeholders as possible in advance of these events, and to get RSVP confirmations to ensure attendance.

Success Metrics and Objectives

The following performance objectives will allow the PSC to evaluate the project and its impacts: 1) number of entries produced for the Agency Triage Database, 2) number of documents produced for the Agency Workflows Collection, 3) number of documents produced for the PEGI Registry, 4) number of guidance recommendation documents produced for information creators, 5) number of

educational awareness and advocacy outreach events and presentations held as part of the project, 6) completion and publication of the environmental scan summary report, and 7) completion and publication of final project recommendations for collaborative future actions.

---

[1] Halbert, Martin; Skinner, Katherine & Sittel, Robbie. *Digital Preservation of Federal Information Summit: Reflections.* UNT Digital Library. http://digital.library.unt.edu/ark:/67531/metadc826639/. Accessed January 10, 2017.
[2] Center for Research Libraries. *Leviathan: Libraries and Government in the Age of Big Data*. CRL Focus on Global Resources, Summer 2014, Vol. 33, No. 4. http://www.crl.edu/focus/summer-2014. Accessed January 10, 2017.
[3] Reilly, Bernie. *Governments and the Digital Record: The Historian's Perspective.* Report on a Panel Discussion on Government Information and Societal Memory convened by the American Historical Association, January 2014. http://www.crl.edu/sites/default/files/d6/attachments/pages/AHA%20Meeting%20of%20Historians_final3.pdf Accessed January 10, 2017.
[4] General Records Schedules can be reviewed at https://www.archives.gov/records-mgmt/grs.html; and Agency Records Management Schedules are available at https://www.archives.gov/records-mgmt/rcs/. Accessed January 10, 2017.
[5] NARA. Records Management Self-Assessment (RMSA) Website. https://www.archives.gov/records-mgmt/resources/self-assessment.html Accessed January 10, 2017.
[6] Peet, Lisa. "Government Website Harvest Enlists Librarians, Educators, Students." Library Journal Online, December 13, 2016. http://lj.libraryjournal.com/2016/12/industry-news/government-website-harvest-enlists-librarians-educators-students/ Accessed January 10, 2017.
[7] https://www.lockss.org/community/networks/digital-federal-depository-library-program/ Accessed January 10, 2017.
[8] Reilly, p. 5.
[9] Jacobs, James. *Born-Digital U.S. Federal Government Information: Preservation and Access.* Prepared for the Center for Research Libraries Global Resources Collections Forum, April 24-25, 2014. p. 13. Accessed January 10, 2017. http://www.crl.edu/sites/default/files/d6/attachments/pages/Leviathan%20Jacobs%20Report%20CRL%20%C6%92%20%283%29.pdf
[10] Jacobs, pp. 2-6.
[11] Center for Research Libraries. *A New Strategic Framework for North American Research Libraries.*http://www.crl.edu/focus/article/10702 Accessed January 10, 2017.
[12] http://www.ppehlab.org/datarefuge Last accessed January 10, 2017