PEGI Project National Forum Summary and Report

Omni Shoreham Hotel, Washington, DC, December 9-10, 2018

EDUCOPIA
INSTITUTE

INSTITUTE of
Museum and Library
SERVICES

Publication Notes

Suggested Citation: Halbert, Martin, Katherine Skinner, Roberta Sittel, Deborah Caldwell, Marie Concannon, James R. Jacobs, Shari Laster, and Scott Matheson. *PEGI Project National Forum Summary and Report*. Atlanta, Georgia: Educopia Institute, 2019.

TABLE OF CONTENTS

Preservation of Electronic Government Information PEGI Project National Forum Summary and Report

Omni Shoreham Hotel, Washington, DC, December 9-10, 2018

INTRODUCTION

The Preservation of Electronic Government Information (PEGI) National Forum was designed to explore and address concerns that have been raised by a variety of librarians, archivists, and researchers regarding long-term access to historically-significant born-digital government information in the U.S.

Today, U.S. government agencies mostly meet their obligations to inform the public of their work through the dissemination of information products online. Most levels of government have one or more agencies that bear responsibility for preserving records of government activity, including both public and internal documents and resources. However, the volume and scope of official information published online is vast, and responsible agencies are insufficiently resourced to comprehensively and systematically collect content, preserve it, and make it available to the public.

In the print era, public information was disseminated, accessed, and preserved through a distributed network model. Publications were provided directly from government agencies to libraries and other cultural heritage institutions through deposit, request, and/or purchase models. The information contained in these print publications has, for the most part, remained available and usable over time. The practice of distributing risk across independent institutions and geographical locations significantly decreased the likelihood that access to this content would be compromised, whether accidentally or intentionally. Moreover, this distributed print-content network encouraged the collection and description of information resources that were not necessarily available through a central distribution point, which resulted in an overall increase in the scope of resources available for the long term.

While no comparable, comprehensive model has emerged for born-digital government information to date, attempts to understand the problem rely on lessons from the past. The PEGI Project is one of the most recent efforts by libraries, agencies, and cultural heritage institutions to plan for, or respond to, the shift to electronic government publishing. The library and archives community already has a storied history of addressing the challenges inherent in providing access and preservation. In September 1999, when the U.S. Government Publishing Office (GPO), then known as the Government Printing Office,

established the Permanent Public Access (PPA) Working Group,[1] this group began to explore central questions that still remain unresolved two decades later. The PPA Working Group, active from September 1999 to November 2000, reached out across industries to find advisors, and included representatives from federal agencies, public interest groups, and other organizations interested in issues "regarding the preservation of, and access to, government information published electronically." Among their goals, the PPA Working Group listed:

- Identifying "at risk" electronic information and developing collaborative solutions or partnerships to ensure its permanent public accessibility.
- Providing a forum for sharing information among the participants and informing a wider audience about U.S. Government PPA activities.
- Formulating policies and programs to assure ongoing access to Federal Government information.

Although the PPA Working Group bears a remarkable similarity to PEGI in its goals and its multi-sector member structure, there has been no formal continuity between early and current efforts to address these problems. The only remaining documentation of the PPA Working Group known to us at this time are screen captures of its website on the Internet Archive's Wayback Machine.

The PEGI Project was the outcome of a series of meetings led by Dr. Martin Halbert, then Dean of the University of North Texas Libraries, beginning in 2016. Known collectively as the Digital Preservation Summit, a series of two meetings were held in conjunction with the 2016 spring and fall membership meetings of the Coalition of Networked Information (CNI). The first Digital Preservation Summit meeting held in San Antonio, Texas, in April 2016, engaged national leaders in a structured, facilitated dialogue. Facilitated by Dr. Katherine Skinner, the 28 attendees of this meeting intentionally represented diverse sectors and stakeholder groups, including federal agencies, university libraries, public archives, and nonprofits involved in digital content management. Together, this group explored the development of a shared agenda to address preservation and access for born-digital government information.[2] Attendees from the first meeting worked with Dr. Halbert to plan a subsequent meeting, which was facilitated by Michelle Gallinger and hosted in conjunction with CNI's Fall Membership Meeting in December 2016. These meetings would serve as one of several models for the PEGI Project National Forum.

Following the second Summit meeting in December 2016, Dr. Halbert and UNT Government Documents Librarian Roberta (Robbie) Sittel applied for and received an Institute of Museum and Library Services (IMLS) grant to continue this work, with additional support provided by participating institutions. The

---

[1] "Permanent Public Access to U.S. Government Information Working Group," U.S. Government Printing Office, site capture by the Internet Archive on September 26, 2006:
https://web.archive.org/web/20060926083022/http://www.gpo.gov/ppa/index.html

[2] Halbert, Martin, Katherine Skinner & Robbie Sittel. "Digital Preservation of Federal Information Summit: Reflections," April 2016: https://digital.library.unt.edu/ark:/67531/metadc826639/.

second Summit meeting also led to the initial formation of the PEGI Project Steering Committee.

What differentiates the PEGI Project from previous efforts is our commitment to harness momentum from the many organizations that care about these issues and are ready to work together toward the common goal. For two years, the PEGI Project has held seven meetings and two webinars, sending project representatives to library, archives, and academic discipline-focused conferences to ask three central questions: 1) what government information should be prioritized for preservation, 2) what is at the greatest risk for loss, and 3) what opportunities might there be for collaborative action to preserve government information. To gather ideas and chart a way forward, we invited representatives from potential collaborator organizations to attend the PEGI National Forum in December 2018.

THE NATIONAL FORUM

The PEGI Project National Forum, held at the Omni Shoreham Hotel in Washington, D.C., served as the capstone for a series of "Mini-Forums" convened as part of the IMLS-funded project to address national concerns regarding the preservation of electronic government information.

The seven prior forums, or Mini-Forums, took place at professional conferences throughout 2018, in order to gather perspectives from a variety of stakeholder groups. The National Forum differed significantly in scope and intent from these Mini-Forums. As the final project event, the National Forum functioned as a springboard for launching PEGI's first collaborative initiatives and thereby maintain the momentum generated by the project's events. With this in mind, it convened highly engaged professionals and leaders who could represent their respective fields and industries effectively, and who exhibited potential to actively participate in the work that lies ahead. Invitations were extended to federal agency staff, state government leaders, archivists, librarians, open government advocates, policy analysts, data specialists, scientists, journalists, digital preservationists, historians, and others. The list of participants is included in Appendix 1.

Selecting participants with the requisite knowledge, skills, and connections was one of the PEGI team's greatest challenges. As a moderated event featuring orchestrated small-group and full-room problem-solving activities, attendance needed to be capped at 60 in order to achieve the best results. The focus of interests in the constrained invitee list skewed heavily toward US federal government information, though there was some participation from representatives of US state and local governments, as well as three Canadian citizens. United States tribal governments were entirely unrepresented in this gathering. While we have sought input from scientists, historians, and public librarians, more work remains to be

accomplished in building relationships for future collaboration.

In the lead-up to the National Forum, we shared preparatory information with attendees for review and commentary. We developed an initial draft of the final project report for distribution to all National Forum attendees, together with the Environmental Scan report earlier commissioned by the project team and produced by Sarah Lippincott.[3] We received feedback from participants on these documents, both before and during the event.

The National Forum was structured as a progression of presentations and discussions, concluding with sessions that sought to identify next steps. Developed by Dr. Katherine Skinner, the facilitation plan mapped out deliberate interactions that reflected the widely divergent perspectives anticipated among attendees. Rather than unstructured conversations that could easily encounter sticking points and roadblocks familiar to those in the room, the facilitation plan raised these issues as themselves worthy of observation, reflection, and discussion. As a consequence, the activities helped to build consensus among collaborators about shared priorities, such as goals for collective action, and shared challenges, such as vocabulary and metrics. The National Forum was an intense and dynamic event, with a high level of positive energy and activity by all attendees. The rich discussions of the event are summarized below, along with the facilitation process used to provoke these conversations. The event agenda is included as Appendix 2.

---

[3] Sarah K. Lippincott, Environmental Scan of Government Information and Data Preservation Efforts and Challenges. Atlanta, Georgia: Educopia Institute, 2018.

EVENT SUMMARY & OBSERVATIONS

As a strategy to draw out rich discussion, we first assigned participants to table cohorts based on shared identity. One table included federal agency representatives. At another table we seated preservationists. Further tables held researchers, practitioners, and so on. These groupings for the initial table conversations were intended to build an atmosphere of engagement and trust between the attendees by prioritizing the cultivation of in-group knowledge first, before encouraging cross-group exchanges. These cohorts formed knowledge-group identities and distilled and documented their specific, concentrated areas of knowledge. Each cohort reported out to the full group the initial observations they had explored. After this important foundation was laid, attendees were assigned to new, mixed cohorts (each with a range of stakeholder and sector perspectives represented) in order to help build bridges between stakeholder groups. The final portion of the event took place with self-selected tables that roughly corresponded to participant-identified affinity groups. As a direct result of this work to first define and then transcend the differences between stakeholder groups, shared values and common themes emerged and were embraced by the full assembly. These included a need for greater advocacy for digital preservation, a related need for a shared vocabulary and strong use cases, and a growing consensus on content that is most at risk.

DEFINING THE CHALLENGE

Following the introductory activities, tables began exploring a prompt to consider what government content is at risk of loss. While the prompt focused on content, discussions strayed into factors that affect how risk of loss could be determined, and the ways in which these risks are associated with how information is created and disseminated. As part of the prompt, participants were also asked to reflect on what they are *not* worried about. Interestingly, almost everything that someone was not worried about prompted a cause for concern by someone else.

A few challenges arose repeatedly in these discussions:

1. It is hard to know what is out there. Government agencies are not consistent in how information is published. Some agencies lack adequate information technology infrastructure, while others fail to publish information for reasons of funding or political will.
2. Beyond factors of technology and infrastructure, information of public interest is not always public. Freedom of Information (FOI) laws are intended to protect the public interest in accessing information, but disclosure practices are not consistently favorable and information

that is released is not typically managed for long-term access.

3. While records schedules address appropriate management of both published and unpublished information of potential public interest, agencies may not have sufficient resources to fully comply with records management requirements. Additionally, records scheduled as "temporary" have a ticking clock for public discovery and access.

4. There are political risks affecting the continued creation of public information, particularly data. If a program is cut, any data from that program that remains publicly available may nevertheless quickly age out of usefulness. Much of our current information policy in the U.S. is controlled by the administration in power, and it can be difficult to advocate for changes.

5. Government information at the federal level falls into a few rough categories: works created by federal employees, works created by federal contractors, works created with federal funding, and works created as derivatives of any of these. A variety of copyright and intellectual property rights may apply, particularly when copyrighted information is incorporated.

6. Format is a significant component of risk. Some information is locked into proprietary formats, while dynamic content can be difficult to preserve in a form that retains its "live" characteristics. Metadata and documentation are essential for using many types of information, and both of these elements can also be at risk of loss, jeopardizing the ability to understand or render the objects they describe.

7. Although much of the discussion centered on federal information, it was acknowledged that information created by subnational governments is almost always at greater risk because fewer organizations seek to collect and curate it, and the institutions mandated to do so are often under-resourced.

8. It was generally agreed that information made publicly available in a trustworthy repository is at low risk.[4] Trustworthiness is important in the public sector for the purposes of accountability, but making the continual improvements necessary to maintain that trust requires ongoing resource commitments.

Toward the conclusion of this activity, conversation turned toward identifying goals for efforts that collect, describe, and provide long-term access to born-digital government information. Participants coalesced around a few broad goals in particular:

1. **Coordinating public advocacy and awareness around these issues.** Making any major improvements in information policy, including securing funding for these activities, requires public support. Members of the public also need to be better informed about how access to government information affects priorities they care about.

2. **Improving how we identify and prioritize what to collect and preserve.** Right now many collection activities are driven in ad hoc, reactive ways that do not always take into

---

[4] Participants in this conversation referenced both the standard auditing processes for ISO 16363, the Trusted Digital Repository (TDR) Checklist, as well as a more general sense of public trust in official repositories.

consideration investments necessary for long-term access. It is important to work with and listen to those who use (or might use) government information, as part of the process of developing and refining collections criteria. It is equally important that any such process focus on inclusion and on addressing systemic issues of power and representation as they pertain to the public interest.

3. **Proactively seeking solutions that reflect public access needs.** Many existing efforts at this time take only limited steps to address inequities in access to information, such as those arising from the digital divide or other barriers to finding, interpreting, and using government information and data. Preservation without equitable access is not ideal, and aspirational goals concerning government information access suggest the need for measures to address information privilege as an explicit focus for efforts.

4. **Improving coordination among existing activities.** Stronger and more creative forms of communication among and between existing projects and with institutions that have a mandate to preserve government information are crucial to avoid unnecessary duplication of effort and to enhance our progress fieldwide. Articulating shared priorities would be easier with agreed-upon standards. Interoperability and reusability are also of interest to everyone, and working together can improve these practices.

5. **Addressing the sustainability of efforts to preserve information.** Proactive planning is important for sustained efforts; projects that rely on a crisis-mode framing may not be able to attract long-term investment. It is impossible (and likely undesirable) to keep everything forever, so prioritization is crucial to a sustainable strategy.

6. **Improved information policies as well as improved implementation of existing policies**. A government-wide mandate to connect publishing and preservation strategies could be transformative, but consistent guidance and application of existing policy within agencies would also help with public access over time.

BRIDGING VOCABULARY

Identifying "trouble terms" that arise related to preserving government information proved relatively easy for all stakeholders. Dozens of terms were flagged by participants, and many of the same terms arose at nearly all the tables. Assigning definitions to these terms was, unsurprisingly, a much greater challenge. Many of the definitions proposed at tables focused less on describing what the term means, and more on scoping how these trouble terms can be used, or when it might be better to choose another, more specific, way to express something. Participants

noted that sometimes terms are used differently in different contexts, especially when specific statutory definitions apply.

One example of a term that differs between general or technical usage and its statutory definition is *machine readable.* The general technical usage means that data is in digital format and can be processed using a computer. The more precise legislative definition, "data in a format that can be easily processed by a computer without human intervention while ensuring no semantic meaning is lost," includes the concept that there is semantic meaning around the data that must also be interpretable to the machine. Some stakeholders might consider a given electronic document to be machine readable even if it does not meet the statutory definition that agencies must follow. As another example, *metadata* has a statutory definition that includes both descriptive, technical, and structural information. Many information professionals use the term *metadata* to refer to separate, distinct concepts or assets. Both of these examples are present in the OPEN Government Data Act (Title 2 of Pub. L. No. 115-435 (2019)).

Terms that are defined in standards are generally understood if the standard is widely used across stakeholder groups. The Open Archival Information Systems (ISO 14721:2012) term *user community* is one example. Because the OAIS model has penetrated diverse communities of practice, there was general agreement on the term as a common and useful one. In other standards more unique to specific stakeholder groups, there was less agreement on the terms. For example, *metadata* is used in many different ways, often with modifiers such as *descriptive*, even within a single stakeholder group.

While changing these definitions or associated professional norms is unlikely, acknowledging the difference is vital to working together.

Other troublesome words are less technical and more philosophical, like *authenticity* or *government information*. While table discussions reached general consensus on the broad meanings of these terms, specific examples highlight how imprecise they are. There can be varying understandings of what made a particular piece of information authentic, from chain-of-custody and technical anti-tampering measures to the authority of an issuing body. Even the "government" part of *government information* is not understood uniformly. Attendees generally agreed upon the Copyright Act definition of *government information*, "information produced by government employees in the course of their work," but the government-funded work of contractors or grant recipients, or privately produced work incorporated by reference, were areas of disagreement.

Another category that lacked mutual understanding is the set of technical terms used differently by different stakeholder groups. Key among these are *access* and *preservation*, each which has multiple commonly used senses aside from the technical definitions. Librarians and archivists have different, and equally valid, professional definitions of what constitutes access. *Preservation* was understood similarly, but with different end goals and vastly different underlying processes and requirements, depending on whether the term is described by those specializing in digital preservation, by librarians working in open

collections, or by archivists or conservators. While changing these definitions or associated professional norms is unlikely, acknowledging the difference is vital to working together. One point of agreement is that the ability of future users to understand and use digital content is dependent on the information they have available about the context of the original content. While this is true of physically preserved materials, the context of a book, pamphlet, or set of file folders is less variable than the context in which digital content was originally made available, and so that content may have differently complex preservation considerations.

Finally, there are words that function in different but related ways. In many cases, these are words like *archive* or *document* which are both nouns and verbs. There seemed to be more general agreement on the noun forms, while the verb forms, which refer to processes, showed more divergence in participants' understanding.

For a list of "trouble terms" identified by forum participants see Appendix 3.

MEASURING INPUTS & OUTPUTS

After a break, the groups reconvened to identify and categorize metrics that could directly or indirectly characterize government information preservation activity. Following this brainstorming session, each group quickly reviewed the table notes and created categories that capture frequent themes about measurement. A sticker voting exercise was used to identify five thematic clusters, which were assigned back to tables for further reflection and development.

Metrics are particularly difficult when the universe of what is being measured is undefined or amorphous, or when the denominator of the equation is unknown. Throughout this exercise, related discussions arose at all of the tables about the limits of measurability. One table noted that certain types of metrics, such as usage, may have a negative effect on long-term efforts, and that other desirable outcomes, such as a stronger and more sustainable democracy, cannot be easily measured. Nonetheless, it is helpful to have agreed-upon metrics, especially in a multi-organizational effort, to guide decision-making processes, measure progress, and identify attainable targets toward shared goals.

What follows is a synthesis of major topics that arose during this exercise. These emerged organically across multiple tables during the initial phase of this activity and then were examined more closely by one or two tables.

Value & Impact Stories

Stories about the value and impact of access to government information are important to the larger effort of born-digital preservation. News stories and ongoing press coverage describing the risks to long-term access to government data and information can support the public advocacy that is critical to sustaining collaborative efforts and seeking funding and other resources that are essential for a meaningful preservation program over time.

Usage

Another key metric is the idea of somehow measuring the use of digital government information. Tracking anonymized usage could help the community by creating a surrogate for the value of these resources, potentially increasing our ability to target what gets preserved and what requires additional attention for access. Possible current technologies that could be applied to the government publishing space include Digital Object Identifiers (DOIs) and Persistent URLs (PURLs).[5] Other systems like CrossRef[6] and Perma.CC[7] could be used to track the number of resolving links as an indicator to drive preservation efforts. For example, they could be used to generate data for "link rot" studies similar to those done by the Legal Information Archive (LIA)[8] to lead toward adoption of best practices that may decrease the amount of information that can no longer be accessed.

What is Preserved?

This metric focuses on discrete counts of digital materials. Tables discussed possible ways to count the amount of born-digital information that is preserved or even made available—including information with complete metadata—with suggestions that included counting the number of agencies contributing to a preservation platform, number of titles in preservation repositories, number of bytes being preserved, number of files preserved by file type, and the number of URLs duplicated in these systems.

Policy

Public policy related to government information preservation includes rules and regulations affecting information creators; those affecting institutions that are mandated to oversee government information, provide public access to it, and preserve it for the long term; and measures to provide funding for these activities.

---

[5] Digital Object Identifiers and Persistent URLs are two examples of registered persistent identifiers that are useful to creating so-called 'permalinks.' For more on DOIs, see: https://www.doi.org/; for more on the current implementation of PURLs used by the U.S. Government Publishing Office (GPO), see: https://www.fdlp.gov/requirements-guidance/instructions/709-purls.

[6] CrossRef is a digital object identifier registration agency that links together scholarly publications. For more information, see: https://www.crossref.org/.

[7] Perma.cc was developed as a solution to court opinions and other legal documents that need to provide links to ephemeral web resources. For more information, see: https://perma.cc/.

[8] Formerly the Chesapeake Project, the Legal Information Archive (LIA) is hosted by the Legal Information Preservation Alliance (LIPA), a non-profit consortium of academic, federal, state and public law libraries working on projects to preserve print and electronic legal information. More information: https://lipa.access.preservica.com/

At the federal level, there is no comprehensive information policy that addresses the access and preservation lifecycle for all public information products. For example, the National Archives and Records Administration (NARA) oversees records management and disposition, but does not oversee information dissemination practices, which are addressed to some extent by the Office of Management and Budget (OMB) and in other ways by the Government Publishing Office (GPO). Some aspects of information policy lack meaningful enforcement or oversight, while others—particularly uniform dissemination practices—are insufficient in coverage and scope. Appropriations also affect the ability of agencies to meet obligations related to publishing as well as long-term access.

While it is difficult to measure policy directly, participants identified different types of improvement in policy that could be tracked over time:

- How is information made available? Is it published on a standard domain?
- How is access maintained for the long-term? What are the official requirements? Is information published with appropriate structure and metadata that make description and preservation easier? Are there persistent identifiers created to locate resources?
- How are information dissemination activities overseen? How consistent are agencies in following rules and best practices? How easy is it to determine relative progress in meeting these requirements? Does each agency have a mandate, mission or standard operating procedure for born-digital publications? Which entities are ultimately responsible for stewarding access?
- How are these activities funded? Is the funding adequate? Are investments in digital preservation made proportionately to the need for long-term access?

Systems and Infrastructure for Preservation Activities

A final area of discussion explored the need for measuring the systems and infrastructure in place for preservation activities. To measure efforts at preserving the government information domain it is important to know which institutions have trustworthy digital repository (TDR) certification or follow another public auditing process. Other opportunities for measurement include a registry of official mirrors within and outside institutions, and seeking other ways of tracking archived content stored at host institutions. It is also possible to track which agencies partner with Data.gov for registering descriptive metadata for their data products, and which non-governmental organizations partner with agencies on preservation activities.

In the discussion wrap-up after the table exercise, many participants agreed that a primary challenge for framing shared metrics is that preservation activities have complex contexts and many dimensions for considering what constitutes progress. Most participants agreed that measuring the universe of electronic government information is difficult, but also valuable as a way to understand what is being preserved.

PATHWAYS FORWARD

For the final portion of the forum, each table group was given about an hour to design and propose projects that would address preservation issues raised in earlier discussions. Using flip charts, table leaders reported out to the full room each group's project title, structure, potential partners and possible activities, keeping in mind that it would be necessary to measure progress toward goals.

Some felt that any measurement of progress toward a goal must include dimensions of the goal itself, described in quantitative terms. Yet there are obvious difficulties in detecting and quantifying the universe of electronic government information. To measure progress toward an ever-growing and ill-defined body of material may not be possible, at least at the early stages of this endeavor. For this reason, groups tended to propose projects having limited scope and therefore more attainable goals.

The purpose of the "Pathways Forward" activity was not necessarily to create actionable blueprints, but rather to demonstrate the many opportunities for creative action that are available to us at this juncture. Even within the short time that groups had to complete this exercise, plans were drafted that have the potential to form some of PEGI's first post-grant work. We also hope that by sharing these ideas, we will inspire others to design projects as well and make further progress, either by acting independently or in collaboration with others. Details on the "Pathways Forward" proposals generated during this Forum are provided in Appendix 4.

FORUM WRAP-UP

The forum concluded with a full room discussion reflecting on the event, highlighting key observations and the most important takeaways. Participants uniformly indicated a high degree of energy and excitement in learning about all the other synergistic programs represented in the room, and being able to come together with like minded individuals who were similarly passionate about the importance of preserving electronic government information.

Dr. Halbert thanked all participants for their engagement over the two days, and indicated that a report of the event would be produced in the coming months and shared with both participants and the wider field. He noted that the project team would incorporate the valuable insights and proposals from the Forum participants in the final project report. Finally, all participants were invited to consider participating in future initiatives that might grow out of this National Forum.

As we discuss our findings from the National Forum and in other venues, we are scoping out the next steps, which we look forward to sharing for discussion in the final project report.

ACKNOWLEDGMENTS

Collectively, we wish to express our gratitude to all who contributed their time and energy to join our conversations, especially those who attended the National Forum. We are especially grateful to Dr.

Katherine Skinner, who developed the facilitation materials for the National Forum, and who has continued as an invaluable advisor to the PEGI Project. We also greatly appreciate the contributions of all of the table facilitators, who gracefully worked to keep the Forum's conversations moving, balanced, on track, and documented. We wish to extend special thanks to Dr. Alexandra Chassanoff and Sarah Lippincott for volunteering their time and effort. Their assistance was invaluable.

APPENDIX 1: NATIONAL FORUM ATTENDEES

Participants and facilitators took part in this event in their individual capacity; affiliations are as of the date of the meeting and are provided for identification only. Participation does not represent agency or institutional policies or views.

Participants

Stephen Abrams, Harvard University

Laurie Allen, University of Pennsylvania

Jefferson Bailey, Internet Archive

Tim Baker, Maryland State Archives / Council of State Archivists

Dr. Andrew Battista, New York University

Andrew Bergman, Sunlight Foundation

Rachel Bergman, Sunlight Foundation

Mara Blake, Johns Hopkins University

David Bleckley, Inter-university Consortium for Political and Social Research / University of Michigan

John Chodacki, California Digital Library

Heather Christenson, HathiTrust

Robin Dale, Institute of Museum & Library Services

Stephen Diggs, Scripps Institution of Oceanography / University of California, San Diego

Cindy Etkin, U.S. Government Publishing Office

Daniel Gillman, Bureau of Labor Statistics

Robert Gradeck, University Center for Social & Urban Research / University of Pittsburgh

Carla Graebner, Simon Fraser University

David Greisen, Open Law Library

Abbie Grotke, Library of Congress

Michael Halpern, Union of Concerned Scientists

Kathy Hart, Library of Congress

Margaret Janz, University of Pennsylvania

Lisa Johnston, University of Minnesota

Heather Joseph, SPARC

Patricia Kim, University of Pennsylvania

Dr. Emily Knox, University of Illinois at Urbana-Champaign

Purdom Lindblad, Maryland Institute for Technology in the Humanities

Brandon Locke, Michigan State University

Jennifer Marill, National Library of Medicine

Dr. Rachel Mattson, University of Minnesota

Edward McCain, Reynolds Journalism Institute / University of Missouri

Dr. Patrice McDermott, Government Information Watch

Dr. Sarah Melton, Boston College Libraries

Mary Moulton, U.S. Department of Transportation

James Neal, Institute of Museum and Library Services

Dr. Bethany Nowviskie, Digital Library Federation

Dr. Trevor Owens, Library of Congress

Dr. Dina Paltoo, U.S. National Library of Medicine / National Institutes of Health

Susan Paterson, University of British Columbia

Meg Phillips, National Archives & Records Administration

Dr. Matt Price, University of Toronto

Dr. Debbie Rabina, Pratt Institute

Dr. Justin Schell, University of Michigan Library

Katrina Stierholz, Federal Reserve Bank of St. Louis

Joseph Thompson, Maryland Library Association

Jessica Tieman, U.S. Government Publishing Office

Michelle Trumbo, Legal Information Preservation Alliance

David Walls, U.S. Government Publishing Office

Marie Waltz, Center for Research Libraries

Dr. Bethany Wiggin, University of Pennsylvania

Facilitators

Dr. Katherine Skinner, Educopia Institute

Dr. Alexandra Chassanoff, Educopia Institute

Sarah Lippincott, University of Massachusetts Amherst

PEGI Project Team Members

Deborah Caldwell, University of North Carolina at Greensboro

Marie Concannon, University of Missouri

Dr. Martin Halbert, University of North Carolina at Greensboro

James R. Jacobs, Stanford University

Lynda Kellam, University of North Carolina at Greensboro

Shari Laster, Arizona State University

Scott Matheson, Yale Law Library

Roberta Sittel, University of North Texas

APPENDIX 2: NATIONAL FORUM ATTENDEES AGENDA

NATIONAL FORUM ON PRESERVING ELECTRONIC GOVERNMENT INFORMATION

Objectives
- Strengthen relationships among key stakeholders engaged in the preservation of electronic government information
- Jointly develop a shared understanding of what preservation of electronic government information activities are currently underway
- Explore shared goals and identify potential shared measurement of progress toward those goals
- Identify and prioritize activities specific stakeholders might engage in to lower barriers
- Stress-test a possible "shared agenda" approach for ongoing cross-stakeholder collaboration

Outcomes
- Deepened relationships and understandings among key stakeholders
- A draft *PEGI Project Final Report* that explores the viability of a "shared agenda" across stakeholder groups
- Progress toward a consensus on metrics that could measure progress toward government information preservation

Prework
- Environmental Scan of Government Information and Data Preservation Efforts and Challenges
- PEGI Project Final Report: Toward a Shared Agenda for the Preservation of Electronic Government Information (early draft version)
- Preservation of Electronic Government Information: An Urgent National Priority by Scott Matheson
- Put an "out of office" message on your email and plan to be mostly deviceless for the working portions of this convening

Agenda

Day 1: December 9

08:30 - 09:00	Breakfast and coffee	
09:00 - 09:30	Welcome, Agenda, Introductions	Robbie Sittel Martin Halbert Katherine Skinner
09:30 - 10:20	Session 1: Spanning Boundaries	Table Groups
10:20 - 10:45	Walk-and-Talk	Full Room
10:45 - 11:00	Break	
11:00 - 12:30	Session 2: Bridging Vocabulary	Table Groups
12:30 - 01:30	Working Lunch Q&A Environmental Scan CI Overview	Brief presentations Sarah Lippincott Katherine Skinner
01:30 - 02:45	Session 3: Measuring Inputs and Outputs	Table Groups
02:45 - 03:00	Break	
03:00 - 04:15	Session 3 (cont): Measuring Inputs and Outputs	
04:15 - 05:00	Day 1 Synthesis and Wrap up	

Day 2: December 10

08:30 - 09:00	Breakfast and coffee	
09:00 - 09:30	Reflections on day 1; Q&A Environmental Scan	Robbie Sittel Martin Halbert
09:30 - 10:15	Working Session 4: Potential Pathways Forward	Table Groups
10:15 - 10:30	Break	
10:15 - 11:30	Conversation Cafes	Rotating Table Groups
11:30 - 12:00	Reflections, Next Steps, Wrap up	Robbie Sittel Martin Halbert

APPENDIX 3: TROUBLE TERMS

As part of the National Forum, attendees developed a list of potentially problematic terminology, including terms for which there may be disagreement or imprecision in definition or use among stakeholder groups. This activity was included early in the forum in order to surface differing usage of terminology, and provide attendees with an opportunity to bridge their different contexts and backgrounds through group discussion.

Participants at the tables made preliminary attempts to develop definitions of selected terms, many of which (but not all) are presented here. The quoted statements are derived from notes taken by the table moderators, so some reflect direct statements while others synthesize discussion. Some footnotes have been added to provide context. Many terms identified by the forum attendees were not defined during the meeting, but are noted here as they were surfaced as potentially problematic or confusing.

Access, in the sense that users can get information
"Public access vs. open access vs. access"
"Machine readable"
"Public access and open access have policies and manifestos"
"Access depends on context"
"'Useful' access a better way to phrase"
"Determination of access is contextual"

Accessibility, in the sense that information can be accessed regardless of ability

Administrative metadata [see Provenance]

Archive *n.* and *v.* "Selected content goes into a repository. Organizational logic is applied in some way. Maybe best to use as a self-identified title or term?"

Authenticity [see Provenance]
"Pieces include [Section] 508 compliance,[9] document metadata, document level / site level information, technical and social information, context and system standards, PREMIS,[10] README [file]"
"Can say how we got it, but not how it was created"

Born-digital [see also Electronic]

Cataloging "Not the same as metadata"

Collection

---

[9] Section 508 compliance refers to federal digital tools that meet mandated accessibility standards. For more information, see: https://www.section508.gov/manage/laws-and-policies

[10] PREMIS is a preservation metadata standard. For more information, see: https://www.loc.gov/standards/premis/

Community or **User Community** "Access to whom, for whom, for what"[11]

Creator

Curation [see Stewardship]
"Philosophy and commitment vs. action or activity"
"Ongoing responsibility"
"Think[ing] intentionally which involves policy, budgets and staffing, institutional vs. 'people' level"
"Intentional approach"

Data / Dataset
"Tabular, statistical [information]"
"Static tables in PDF [documents]"
"Singular or plural? Individual or summary? Machine readable?"
"Digital (ensuing discussion about whether analog materials can be 'data'; the group conceded yes)"
"Too reductive, disagreement on meaning. 'Everything' vs. datasets. Is a turn-off to humanists; important for community engagement. Specificity of what we mean can make problems seem solvable. We prefer information, or more specifically, data set"

Discoverable

Document *n.* or *v.*

Electronic "Is this the same as digital? Online? Electronic in FDLP means physical media with digital information"

Enduring or Enduring Value

Ethics [see also Privacy]

Format [also Proprietary format]

Government Information
"Documents or data or communications produced or commissioned in whole or in part by or for government agencies or individuals"
"Legal/legislative definition may differ from library or common definition"
"Document activities, policies"
"Held and managed by government or other parties"
"Difference between federally funded research and government information"
"Difference between public and government information"

Information [see also Data]

Inventory

Knowledge "How can we provide access to data so that it becomes knowledge?"

---

[11] This definition is derived from the definition of designated community in the OAIS model (ISO 14721:2012)

License raised in the context of Canadian (or non-US federal) content

Machine readable

Metadata
"Should our definition describe the shift from ordinary cataloging to machine readable?"
"The description that leads to discovery and access"
"Metadata makes something understandable and useable"
"Metadata can facilitate aggregation and discovery"
"One person's metadata is another person's data. One person's data is another person's metadata."
"Cataloging and description is not necessarily the same as metadata"
"Anything that helps to make something more findable/useable"
"The term can do more harm than good"
"Lots of assumptions about how it can be used. Computer or person? What type of person?"
"Assumption that there is a person on the other end who will understand the conventions you're using"

Object

Open Access [see also Access]

Personally Identifiable Information
"Less data keeps getting more identifiable"
"Currently have a low bar but more technologies make it easier to identify"
"What constitutes de-identified data is not well defined"
"[We] would like a resource to...help...understand if data is identifiable"
"Can preserve, but not provide access to data, or limited access: potentially runs into funding limitations"

Platform

Policy

Preservation can refer to NDSA preservation levels as a descriptive model.

Privacy [see also Personally Identifiable Information]

Provenance [see Authenticity]

Public "People: public at large, for purposes of availability. Stuff: interest in or affect on people."

Public Access [see Access]

Publisher / publish

Record
"Lots of misunderstanding of 'record': it's not a subset, it is the superset"
"All Information Dissemination Products are records, not all records are Information Dissemination Products"

Registry

Repository [see also Archive]

Resources

Reusability [see also Machine Readable]

Risk

Stakeholder
"People who care"
"Impacted groups"
"Users / Producers / Creators / Owners"
"Public good / Public interest"
"Secondary/tertiary impacted groups"
"Institutional bodies / Decision makers"

Stewardship [see Curation]

Sustainability

Transparency
"Not hidden, not public"
"Knowledge of provenance"
"Supports accountability"
"Allows meaningful access to people with different knowledge levels"
"Spectrum of accessibility"

Trust / Trustworthy

Usability [see also Access and Accessibility]
"Information has documentation, is...accessible and available/discoverable to all, and can be reused.
FAIR data principles are a helpful model for usability"

User [see also Community]

Values

Version control [see also Version Control]

Vulnerable "A troublesome term that sits at the intersection of technology and power"

Website

APPENDIX 4: PATHWAYS FORWARD PROJECT PROPOSALS

Project proposals ranged from relatively simple, easily executable initiatives to those that aim at accomplishing more systemic changes around dissemination and preservation of electronic government information. The majority of the projects illustrate a need for partnerships with government agencies to articulate the value of these activities and to develop solutions for both identifying and preserving information.

PROJECT #1

TITLE: FRIENDS F FEDERAL FACTS

DESCRIPTION: We propose an umbrella organization "Friends of Federal Facts" to provide central administration and leadership for the formation of numerous topic-based federal information advocacy groups. For example, there could be a Friends of Federal Facts for the Census Bureau, or Friends of Federal Facts on Climate Change. Groups would function as advocates for the agency to continue important data collection and information publishing activities. They would represent users such as scholars, journalists, and scientists, in voicing the need for data in usable forms. They could provide ongoing feedback to agencies about the usefulness and usability of their information products.

POTENTIAL PARTNERS: These sub-groups could include both individual and institutional members. Membership could appeal to librarians, libraries, archivists, archives, journalists, civic groups, students, scholars, private sector employees, publishers, and members of the general public. Institutional members could include professional associations and commercial organizations including publishers. There could even be a few celebrities whose endorsement would help with visibility. Of course, the agencies themselves and government employees would be crucial partners.

POTENTIAL ACTIVITIES:

The umbrella organization "Friends of Federal Facts" could…

- Provide administrative support, ideas and suggested activities to sub-groups.
- Maintain an online directory of its sub-groups, with contact info and web addresses for each
- Guide and mentor anyone who might like to establish new sub-groups

The sub-groups could…

- Use communication tools such as listservs, social media, and blogs to update members and the general public on projects, news and developments at government agencies, and seek to enhance cohesion among members.
- Communicate with legislators about public needs for government information
- Publicize the good work of agencies, giving credit for work well done
- Hold events and meetings, in person or online, inviting special guests from the government

agency to speak or be interviewed
- Help agencies better understand those who need and/or use their information products
- Plan new projects according to need

PROJECT #2

TITLE: DOMAIN E-HEALTH ASSESSMENT OF GOVERNMENT PRESERVATION FRAMEWORK DEVELOPMENT (DEAGP)

DESCRIPTION: This project would involve testing or piloting the development of a framework for documenting government information being produced within defined subject areas, how it is being preserved and any risks involved in preserving that information. The framework would develop and employ a reusable methodology that could be applied to climate, health policy, biomed research, energy, transportation, criminal justice, immigration, international relations, air quality, child welfare, education, or other areas.

POTENTIAL PARTNERS: This project would require expert researchers and supporting organizations. Additional partners could include good government organizations, government information librarians, Civic Switchboard, technologically-minded leaders at information science programs, and the Digital Library Federation (DLF).

POTENTIAL ACTIVITIES:

- Develop a survey instrument to interview information producers, consumers and preservers
- Document information produced, indicate where and how it is being preserved, and conduct risk assessment
- Use the NARA records schedules and risk assessments pertaining to agencies involved in each subject domain
- Utilize information gap analysis. Identify patterns and elements that contribute to the problem of fugitive information.
- Draw up a stakeholder map, and identify missing pieces
- Identify any legal mandates involved, and again, identify missing pieces
- Identify and map out any subject based web archives or similar assets
- Make a plan for risk mitigation

PROJECT #3

TITLE: BUILDING A SOCIAL NETWORK AROUND THE DATA RESCUE TOOL

DESCRIPTION: This project will collaborate to build a data nomination tool.

POTENTIAL PARTNERS: In order to test and successfully roll out a data nomination tool, we will build a

network of users both inside and outside of the federal government. Potential partners could include Johns Hopkins University, Cloudburst, Earth Science Information Partners (ESIP), GPO, LC, NARA, ICPSR/DataLumos, and federal agencies.

POTENTIAL ACTIVITIES:

- Conduct an outreach and engagement campaign to agencies, practitioners, and potential partners (public libraries, universities, nonprofits), and seek other ways to reach out to communities of need
- A project team could pilot the tool and test functionality and usability, and seek feedback
- Funding could be secured through grant applications or endowments
- Progress could be promoted through press releases, and the team could develop use cases

PROJECT #4

TITLE: INCREASING TRANSPARENCY OF FEDERAL RECORDS DECISION MAKING

DESCRIPTION: This project will seek to develop a platform that increases the legibility of records retention schedules to improve public input into national preservation decisions. We want to work with NARA to move from the current system, which can be opaque or confusing to members of the public, to a system that is more open and clear.

POTENTIAL PARTNERS: NARA, other federal agencies, 18F, Digital Library Federation (DLF), Union of Concerned Scientists, American Library Association (ALA) / Government Documents Round Table (GODORT), Society of American Archivists (SAA) / Government Records Section (GRS), American Association of Law Libraries (AALL) / Government Documents SIS (GD-SIS), ARMA International, transparency coalitions

POTENTIAL ACTIVITIES:

- Undertake an information-gathering, relationship-building meeting with NARA, federal agencies, and other stakeholders, about technology and legislative needs.
- Use existing mechanisms such as Endangered Data Week to analyze possible avenues of improved communication and conduct a design charrette with multiple communities.
- Come up with proactive approaches to encourage public & Congressional engagement with schedule revisions throughout their development.
- Seek Congressional funding for records management needs; seek additional funding as needed for cooperative platform.

PROJECT #5

TITLE: SMART (Small/Medium-sized Agencies Repository Training) PROGRAM

DESCRIPTION: This project would create a best practices training program with guidance documents for small and medium sized agencies. The rationale for this target group is that small-to-mid-sized agencies are both more likely to have at-risk information as well as being more receptive and interested in such training. The project would start by documenting existing practices associated with management and preservation (or lack of preservation) for electronic government information in a representative variety of agencies. This would be accomplished by organizing and coordinating a cohort of *agency fellows* to undertake a series of case studies in a range of small federal agencies, and regional/municipal agencies. Each year-long fellowship would be aimed at documenting policies and practices in an individual agency. The case studies produced would be analyzed in a larger national context to identify emerging best practices which would then be distilled into guidance documents and training materials for agency staff members.

POTENTIAL PARTNERS: National Digital Stewardship Alliance (NDSA), OpenGov, The Carpentries, Code for America, AmeriCorps

POTENTIAL ACTIVITIES:

- Key stakeholders of small agencies would post/reposit policies, procedures, plans, metadata, contracts, and workflows.
- Analysis of needs in small agencies through case studies.
- Webinars, workshops.
- Metrics of success, Impacts, positive impacts report.
- Fellows in the agencies.
- Focused case study on NYC agencies.
- Toolkit for metadata for OpenDocs
- Use existing tools: OpenRefine (http://openrefine.org/ from Google)

POSSIBLE FUNDERS:

- Foundations

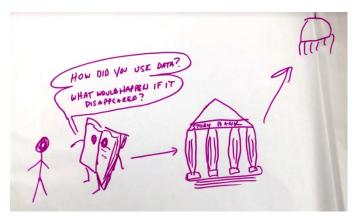SUSTAINABILITY STRATEGY:

- Seek continuing hosts for cohorts of fellows

PROJECT #6

TITLE: MIND THE GAP: CONNECTING USE WITH PRESERVATION

DESCRIPTION: This table focused on the problem of disappearing data and imagined a story bank that could hold specific examples of data loss gathered from the range of different "users" to convey the importance of good preservation—and to emphasize that provisioning access for use of materials is as significant as ensuring that we've collected and preserved the data.

POTENTIAL PARTNERS: Businesses, public libraries, schools and universities, libraries, journalists, influencers, civic data organizations, and government agencies at federal, state, and local levels.

POTENTIAL ACTIVITIES: Engage directly with businesses and users of government information to target, along with the stewarding agencies. Solicit user stories about data use, including potential fallout if the data were to disappear.

Figure 1: A drawing from the table notes for Project #6. The book says "How did you use data? What would happen if it disappeared?" The building is labeled "Story Bank," and the final destination is a government building.

PROJECT #7

TITLE: iDEPOSITORY—A COMMUNITY REPOSITORY

DESCRIPTION: This project will broaden the FDLP to engage with libraries at the local level through training and outreach, taking the program beyond the print world. Some potential goals are to facilitate a cultural shift so that "everyone is part of the network," and to seek to build digital communities that can center public library participation in access and preservation networks.

POTENTIAL PARTNERS: GPO and FDLP participating libraries, Internet Archive, public libraries, college libraries, digital preservationists, civic technology sectors, digital humanities centers, community networks, broadband access projects, and local governments.

POTENTIAL ACTIVITIES: Provide training and outreach to build digital communities, and encourage the formation of a distributed network of preservation. Seek ways to support Census Complete Count Committees and other e-government work in public libraries.

POSSIBLE FUNDERS: IMLS has several programs that may be applicable for this project.