US House Legislative Data and Transparency Conference
May 22, 2013
https://cha.house.gov/2013-legislative-data-and-transparency-conference

Panel on preservation w James Jacobs, Lisa LaPlant and Marc Levitt

Comments by James Jacobs, Government Information Librarian, Stanford University Library

Thanks Reynold and all of the conference organizers. I'm really glad to be able to participate in today's conference even from afar. I've been an active participant in the area of govt information and libraries for a while so hope I can add something to the day.

What is digital preservation?

Digital preservation is the combination of policies, strategies, infrastructure and actions of an organization or group of organizations that guide preservation efforts and ensure access over the long-term to digital content that is under their control.

There are many libraries, archives and other memory organizations working on digital preservation at this time. The role of dedicated memory organizations is essential to long term preservation and access. It is, simply put, what we do. We as a society need dedicated memory orgs to ensure preservation and open, free access to the raw information streams so that others can use the data. Libraries and archives will ensure that future govtracks will be able to be built and that there will be bridges spanning the digital divide.

These organizations – including govt agencies like Library of Congress, GPO, and NARA, Internet Archive, national- and academic libraries and archives around the world, as well as membership groups like the National Digital Stewardship Alliance (direct descendent of NDIIPP), Intl Internet Preservation Consortium, Digital Preservation Consortium, Digital Library Federation, Digital Preservation Network (DPN) etc. have all put significant time and effort into building systems, policies and procedures to assure long-term digital preservation.

The accepted standard practice for digital preservation is the Open Archival Information System (OAIS), a model that provides a framework, foundation, and consensus for building the elements and processes for long-term digital information preservation and access. Key to OAIS is the reliance on a designated community of interest and contingency planning.

In the interest of time, I'll cut right to the chase. Digital preservation needs to be the following:

- Planned for as part of the information life cycle, NOT after-the-fact or through serendipity. Google is not a preservation plan.
- Distributed across organizations. Many hands make light work but they also assure preservation regardless of budgets, natural disasters, technology failure, or issues of Cybersecurity.
- Redundant. We can't have digital content in one silo and call that preservation.
- OAIS compliant. That means being well-structured and standards-based so that failure of one particular organization does not mean irreparable loss of content. The implementation of standards greatly reduces operating costs.
- Lastly, digital preservation of govt information needs to be held outside of the .gov domain in order to ensure tamper-evidence and govt transparency.

*this is a mix of commonly agreed upon needs, and my own opinion ;-)

With these needs in mind, I'd like to briefly talk about the current state of digital preservation of federal govt information:

--More and more historic content is in the process of being digitized and stored in Hathitrust, Internet Archive, Google books, as well as over 150 completed and ongoing digitization projects listed on the FDLP registry at registry.fdlp.gov. Digitization will continue for the foreseeable future. Though there are issues and problems with

digitized content, that content is not the primary concern of my talk today.

--Of more importance to me is born-digital content, because that is the content at most risk in terms of preservation:

GPO's FDsys content management system hosts a large amount of Congressional, executive and judicial branch content from the early 1990s – present, with some collections going back even farther. Lisa LaPlant will be up next to go in-depth into FDsys. I'll just say that FDsys is OAIS-compliant. Content on FDsys has robust metadata to aid in preservation and access.

In addition, with the consent and help of GPO, all FDsys collections are harvested and preserved by the 37 libraries currently participating in the LOCKSS-USDOCS program. LOCKSS is Lots of Copies Keep Stuff Safe, a long-standing and award winning open source distributed digital preservation system built and maintained by the Stanford University Libraries. I won't go into long explanations about the technology behind LOCKSS-USDOCS, but think of the program as the digital Federal Depository Library Program (FDLP) where digital content is distributed out from GPO to participating libraries where it is preserved for the long-term. Conceptually, this mirrors the historic process of the FDLP and its 200 years of distributed access and preservation of govt information. I want to stress that LOCKSS is just one current digital preservation system in place at the moment. LOCKSS is very important, but it need not be unique.

The other major digital preservation activity involves Web harvesting. We have the '08 and '12 End of term .gov domain Web crawls (LoC, IA, UNT, CDL). In addition, dozens of institutions are using the Internet Archive's Archive-it subscription harvesting service or the CA Digital Library's Web Archiving Service to target and harvest particularly important govt information. A few quick examples: I'm using Archive-it to harvest CRS reports from 23 sites across the net (both .gov and NGOs); I'm harvesting FOIA documents including federal agency FOIA reading rooms and NGO sites that post FOIA'd documents like the National Security Archive and other non-governmental groups. There are many more examples of Web harvesting of .gov sites being done by other institutions.

Issues:

However, there are issues with these current methods of digital preservation.

The GPO used to be the Federal govt's primary publisher/distributer of govt information. But since the mid-1990's more and more Congressional committees and executive agencies have become their own publishers. The issue of fugitive documents – those documents which are in scope of the FDLP, but are not making their way into the program – is a fast-growing problem as these groups choose to self-publish on the Web with little or no thought about information lifecycles and long-term preservation let alone publishing- and Web standards. Preservation is therefore largely serendipitous, unplanned and after the fact if it's done at all.

Web harvesting has its own set of problems. It is difficult to ascertain quantity and quality of harvested content. Harvesting is expensive, incomplete, and ineffective especially where dynamic content is concerned – sort of like taking snapshots of the pages of a phone book instead of getting a copy of the phone book.

We need the data behind dynamic web sites more than we need the presentation of the data on web sites. And we need govt organizations to acknowledge the need to instantiate its information in a preservable format rather than just "make it available" piecemeal one page, database query, or API call at a time. And to do that, we need cooperation, collaboration and planning.

So I'll end with some ideas for what I think Congress should do in terms of digital preservation:

1. make use of existing FDsys infrastructure and avoid the expense of either duplicating or reinventing the wheel.

GPO's infrastructure is designed around OAIS to ensure ingest, preservation, and dissemination as well as discovery. This is their primary job, not an add-on or unfunded mandate (which is how I'm guessing some agencies might see the requirement for open data).

FDsys is already designed around the life-cycle of information and OAIS requirements. Using FDsys infrastructure would also bring Congressional information within the bounds of Title 44.

2. Congress and executive agencies should support funding for (maybe even contribute to?) Trusted Digital Repository (TDR) certification for FDsys. The TDR process would load test FDsys to assure that technologies and policies are in place for long-term preservation. This in turn should make it adminstratively easier for Congress and agencies to rely on FDsys and to be assured of long-term preservation and dissemination of their information and data.

3. GPO's infrastructure already supports replication not just to lockss, but to other future systems as well as near future systems like DPN. This has the advantage of bringing in stake-holder partners who will bring to the table their own financial, technical, human, administrative resources. This will also prevent single-point-of-failure risks of govt information and data, provide built-in sustainability, and provide a succession plan for any future failure or lapse in government funding. Proper preservation metadata can also ensure authenticity of replicated data and reduce the cost of maintaining fixity and authenticity.

4. FDsys can support redirects to agency-friendly permanent urls (committee.gov/data and committee.gov/publications or docs.house.gov which would fall in line with the agency.gov/data recommendation of the May 9th Executive Order on open govt data) allowing committees and agencies to 'brand' their data and get credit for their open-data policies.

Abby Smith Rumsey, in the executive summary of the 2010 blue ribbon task force on Sustainable Digital Preservation and Access, wrote that, “Access to valuable digital materials tomorrow depends upon preservation actions taken today; and, over time, access depends on ongoing and efficient allocation of resources to preservation.”

The House's ongoing and positive efforts toward XML structures and bulk legislative data need to facilitate both the downstream efforts of folks like sunlight, LII and govtrack AND assure long-term

preservation. The groundwork is already in place. With efficient collaborative action between the government and memory organizations today, we'll assure preservation of our nation's heritage long into the future. I'm ready to go. Will you join me?