Speaker notes
Government Records and Information: Real Risks and Potential Losses.
by James A. Jacobs. April 25, 2014.

Center for Research Libraries Global Resources Collections Forum:
Leviathan: Libraries and Government Information in the Age of Big Data

(2) outline
I'm here to report on the paper I wrote for CRL that is one of your background readings. I won't repeat the details that are in the report. Instead, I will

- Review the issues and findings of the report
- Provide a little historical context that is not in the report
- And Suggest a framework for addressing the issues

- i'm also going to ask you (in about 15 minutes) to write a couple of things down in a very short brain storming session. So you might want to get your note-taking device of choice handy and turned on and open.

(3) 3stages.org/crl
I have put links to my report, this presentation at this URL: 3stages.org/crl

At that URL, you will also find additional links that I will mention during the presentation.

(6) gaps

Let's begin with what we do not know. A great deal of the problem of preserving born-digital government information can be attributed to gaps in our knowledge.

For example: there is no directory or listing or catalog of digital government information.

in fact, we do not even have a list of of all government websites, from which we might attempt to compile a list of government information.

In addition, libraries and other memory institutions do not have a unified approach to identifying and sharing information about what we *have* preserved.

But There are some things that we do know...

(7) what we know (1)
First, we know that FDLP libraries have successfully preserved millions of volumes of non-digital government information.

And we know why and how that happened.

(8) life cycle
We can see, from this greatly over-simplified and incomplete "information life-cycle" diagram, that for many years there was an institution that took responsibility for each of these life-cycle stages.

Agencies created information and sent it to GPO.
GPO produced and distributed the information and sent it to FDLP libraries.
The libraries preserved the information and provided access and services for it.

For a long time, this worked pretty well.

We called information that was outside this model "**fugitive**." it had escaped and was hard to find and likely to disappear over time.

(9) what we know (2)
We also know that most born-digital government information is not being preserved by libraries.

And we know why:

Libraries have been relying on government agencies to preserve their own information,

We have enough experience with digital preservation to draw a conclusion about this:

Putting complete and sole responsibility for the preservation of *any* information on the shoulders of any *single* organization is risky.

- In such a case, preservation is at the mercy of changes in budgets and priorities and technologies of that single institution.
- It more risky, still, if the organization is primarily a content creator (as agencies are) and does not have preservation as a primary mission (and very very few agencies have preservation as a mission **at all**).
-And it is even riskier if the organization is itself governed by politicians who have a political stake in control of that information.

I have heard a lot of librarians say that this situation was inevitable with the shift to digital publishing, but I do not agree with that. Let me give you two examples of why most government information is not being preserved by libraries today.

(10) 1983

First example: this supreme court decision had the effect of breaking the life cycle model of preservation because it allowed agencies to NOT send their content to GPO. This meant many more documents became "**fugitives**."

Although libraries have attempted to identify and capture fugitives, they did not put adequate resources into that effort and the result has been fewer documents in libraries.

(11) 1993

Second Example: Ten years after the Supreme Court decision, this law was passed. It directed GPO to provide "electronic storage" and "on-line access" to government information.

Although Congress has continued to explicitly direct GPO to provide both DEPOSIT and ONLINE ACCESS, GPO CHOSE to limit deposit of digital information. In 1995, GPO asserted that it would be solely responsible for permanent access, even though Congress neither neither suggested nor authorized such a role.

Libraries accepted this with the result being that most born digital government information has not been going into FDLP libraries.

The point of these two examples is that the problems of preserving government information are not new and were not created by technologies. The problems are created by people making choices.

(12) what we know (3)

This brings us to the third thing we know: The scope of born-digital government information now greatly out paces what is being done to preserve it.

(13) histogram

We can see this graphically in this chart from my report.

> 10,200 items distributed by GPO to FDLP libraries in one year,
> 2.3-3 million items estimated to be held in the Federal Depository Library Program
> 160 million URLs harvested in the 2008 End of Term Crawl
>
> The EOT crawl attempted to harvest a snapshot of federal government information on the web at the end 2008.

(14) simple fact

If we had to draw one conclusion from all this it would be that: no one knows how much born-digital U.S. Federal government information has been created, or where it all is, or how much of it is being preserved.

The fact that we do not know this is clearly the most obvious of several gnarly issues that we need to address as we start to think about digital preservation.

Let's look very quickly at some of the other issues.

(15) Issues

Versioning

Because digital documents can be changed easily, it is important that preservation activities identify different **versions** -- both in order to preserve unique content, and in order to minimize preserving the same content many times unnecessarily.

Link Rot (· The need for persistent URLs)

You have probably all seen the Chesapeake Group's latest report on **link rot**. By studying the URLs of documents that it selected and preserved, it discovered that 51% of .gov links are no longer accurate.

Temporal context is an important and much overlooked aspect of preservation.

Users of preserved information need a way of using that information in its original context. A link in a document to another document should not only work, it should also link to the same content that the author linked to at the time the document was created -- not a later version that the author never saw that may have replaced the document the author linked to.

I've provided you with a links on the 3stages.org site to an interview with Herbert Van de Sompel, and an article by Scott Ainsworth on this topic.

E-government

The proliferation of **e-government services** has led some librarians to suggest that such services obviate the need for libraries to have collections of government information at all. I would suggest that the opposite is true.

To see this, we need only recognize that there is a difference between e-government, which is a *service* that uses government information and the *information itself*, which is a *resource*.

This means that E-government services often have the effect of hiding the resources behind the service -- thus making it harder for us to identify, evaluate, acquire, and preserve the resource itself.

The availability of e-government information services should be seen as an impediment to digital preservation, not an excuse to ignore preservation.

Relying on government for preservation (and free access)

As I have mentioned, most government agencies do not have a legal mandate to preserve their information for the long-term. That means that **relying on those agencies for preservation** is risky. Even those that do have preservation as a mission do not have the resources to do so adequately.

Even GPO's legislated mission of providing access to government information does not specify that it must keep everything online for ever for free. Although the *current* administration of GPO operates with the intention of preserving and making information freely available, there is no guarantee that a different administration would not make a different choice -- as other GPO administrations have.

Selection

The huge volume of digital information makes **Selection** more important to preservation than ever before. When libraries rely only on issuing agencies to preserve their agency information, they relinquish to those agencies the decision as to what is worth preserving and what will be discarded and lost.

Libraries will often define the scope of what needs to be preserved differently from agencies. And different libraries will (legitimately) define the scope of what needs to be preserved differently from each other --- and that is a good thing.

Collections need Services

Finally, **collections without services** are of little value to our communities of users. When we build dark archives with no access and little organization, we are not creating a value that our communities can see or use.

Libraries thinking of preservation should also plan for *access* and *services* that provide immediate value for their communities of users.

The background paper lists some of the major and well-known projects that preserve born-digital federal government information. What I want to do now is characterize two dimensions of preservation that we might take as lessons from those projects.

(16) [who should preserve?]

First, we can see that there are 3 typical models of who does the preserving:

·Government alone (NARA)
·Government with non-government partners (GPO / LOCKSS-USDOCS)
·Non-government without government cooperation (IA)

Experience with these models also reveals some important lessons:

We know -- or should know -- that relying on the government alone is risky for reasons I mentioned earlier.

We also know that trying to locate, identify, and acquire digital content w/o the cooperation of the government is, at best, difficult and prone to errors, gaps, unnecessary duplication, etc. [On the 3stages website, I have some examples of why this is so and provided a link to a very informative podcast that discusses this issue.]

I think it should be self-evident that the ideal preservation strategy would be for the government to cooperate with memory institutions.

(17) [Methods of selection]

The second lesson we can learn from existing projects is about the available **methods of selection**. Broadly speaking, there are three methods of selecting information for preservation:

-Broad web harvesting (IA)

-Focused selection (which can be either narrowly-focused web-harvests such as some of the Archive-It projects, or one-title-at-a-time selection, such as the Chesapeake project).

-"Digital Deposit" in which those who create and produce the information **create preservable digital objects** which they deposit with memory institutions.

To ensure adequate preservation for all our user communities, we will probably need a mix of these methods.

Finally, I want to suggest :

(18) A Framework for addressing the issues

- Preservation and Access
  Preservation should not be seen as an isolated activity. Rather it should be connected to access.
- Collections and Services
  Further, our preservation activities should be part of us building digital collections for use. As we design preservation strategies, we should include organization, discovery, access, service, and utility.
- Focus on user-communities first
  We can do this by focusing on our different user-communities first. We should complement a Provenance-approach (choosing an Agency or a web site or a domain for preservation) with a user-center approach. Rather than only looking at the web and wondering what we should preserve, we should look at our users and ask what they need.

  This gives the community of libraries an opportunity to expand how we collaborate by expanding how we think about our user communities. Instead of limiting a library's community to those that are geographically-local, we can identify communities that have similar content-interests and use-needs and collaborate on the building of collections and services for them.

- Unique collections for unique communities
  With the power of sharing collections on the web, we should collaboratively build collections that fit the needs of unique communities. For example: a group of libraries could build a shared collection for a specific user-community with shared collection needs: disciplines like law, medicine, agriculture lend themselves to such an approach as do smaller, specialized sub-disciplines like community health, epidemiology, nutrition, and food security.

  We could also build collaborative services that provide different kinds of functionality to different user-communities. For example, a demographer or sociologist or historian may want raw census data for analysis. On the other hand, most needs of most undergraduates for census data may be satisfied with aggregate tables of statistics and lookup functions. So we would have two different services for the same content.

- Participation of every library
  The above thoughts lead to the conclusion that every library should participate in digital preservation. The reasons for this go beyond preservation: This is about building the value of libraries by providing a combination of *collections and services* that are reliable and useful.

·Cooperation and Collaboration

This does not mean that every library needs to build a data center. This is not about technology! Shared collections and services can be built with different kinds of participation by different libraries. Some libraries could function as data centers; other libraries could develop services and create metadata, and so on; *every* library could select.

We do need more infrastructure to share what we are doing and tools for better discovery, access, and use, but the technologies for those exist as well.

And we've faced similar challenges in the past and developed mechanisms for addressing them.

- "DocEx" or the Documents Expediting Service, operated out of the Library of congress from 1946 till 2004, and successfully tracked down and provided copies of "fugitive" documents.

- The ambitious Farmington Plan, which, beginning in the 1940s and extending through the 1960s attempted to acquire one copy of every significant book published in numerous countries.

- crl

(19) Summary

In summary, I believe:

- That we can preserve born-digital government information (the technology exists)
- That every library can participate (the entry-cost is low)
- That we can add value to the information by building collections of use to our communities.
- And that we can add value to our libraries by providing collections + services for our communities.

The result will be more than *preserving* government information: Users will directly and immediately benefit by the value that libraries provide to them with these collections and services.

(20)
What I want to do now is ask you to participate a bit.

We have about 10 or 15 minutes left in this session.

I would like each of you to write down 2 things that you can do when you get home to start working on the preservation of born digital government information. They should be **action items.** Actions can include anything from asking questions and learning more, to participating in meetings and committees and groups by adding to their agendas, to developing plans and initiatives and projects -- and beyond.

I'll give you just about 1 minute to write down your 2 action items. Then, I will ask for volunteers to share your items with the group. We'll use these last few minutes to share ideas and start a new discussions.

begin!