Roles for Libraries Big and Small in the Digital Preservation of Government Information

Elizabeth Cowell
University of California, San Diego
San Diego, CA

My name is Elizabeth Cowell and I am the Head of Data, Government and Geographic Information Services at the University of California, San Diego. I would like to thank the GPO and the Depository Library Council for giving me the opportunity to talk to you today. While it might seem like I'm picking on the GPO, that is not my intention. I am advocating for an alternative way to provide permanent public access to government information, keeping in mind that the most important thing we do as Federal depository libraries is protect the public's right to know.

Outline

In my talk, I'd like to do three things:

Tell you why I believe that the depository system must work in the digital age.
Outline the functions that are essential to maintain a depository system and raise the question of who can best fulfill each role.
Explore practical steps that we can take today to achieve the FDLP of tomorrow.

The two speakers who follow me will describe projects that incorporate many of these ideas into action. Trisha Cruse from the California Digital Library will talk about their project, Counting California, and then Chuck Eckman will talk about the LOCKSS project at Stanford University. LOCKSS meaning Lots of Copies Keeps Stuff Safe.

1. Why…

While it may seem odd at a meeting of members of the Federal Depository Library Program to explain the concepts of "depository" and "library," some argue that "deposits" are unnecessary, that we are in a "post-depository era" and that the basic functions of libraries are unnecessary or have changed or will change because of the shift of publishing from print to digital.

Let's review some criteria for judging the success of a system of public access to government information.

First, it must serve *multiple* communities of interest adequately.

Second, the government has the responsibility to make government information widely accessible to ordinary citizens and to do so in a way that permanent public access is assured.

Third, the government must not have exclusive control of access to government information.

These criteria lead us to one solution: GPO should deposit publications, regardless of format, with depository libraries.

This conclusion is important for the public because distributing copies of digital documents will assure the public of several things. It enables different communities of interest to have their own collections of materials collected, organized, presented, serviced, and preserved to meet their needs. It provides an immediate and workable way of providing short-term public access to

government information. Because it ensures the short-term preservation of digital documents, it assures the public that those documents will be available for long-term preservation when more permanent digital preservation solutions are available. Finally, it provides the public with a simple, understandable and workable assurance of the authenticity of digital publications without giving over to the government control of that information.

The model proposed by GPO doesn't meet the basic criteria in several ways.

It proposes a single electronic collection rather than multiple collections. No one-size-fits-all collection can adequately meet the information needs of all communities of interest. This should be self evident to us. Our libraries are not identical because we serve different communities. A document that in my library may be essential may be of no interest to your community. Lawyers require different collections, organized in a different way, than K-12 libraries, or environmental libraries, or agricultural libraries.

GPO cannot ensure permanent public access to public information because it has neither the Congressional mandate nor the funding to do so. If it did, such mandate and funding could be changed by Congress at any time. Even GPO's very existence is not guaranteed. To state it as simply as possible, no model that relies on a government agency can be considered a permanent solution. Good intentions, excellent plans, and workable policies are only as permanent as the funding that enables them. The GPO model relies on government information remaining in the control of GPO and other government agencies. This model even tells us that if government information is obtained from any server other than a government server, it cannot be guaranteed as "authentic" thus imposing a new definition of authenticity. It is as if, in the print world, GPO said it is OK for users to look at that government document on your library shelf, but only one that the user gets from a government bookstore is authentic.

Let me address the concerns of many that getting digital materials is beyond the abilities of our libraries today.

The digital world is simply one more change in formats. We've had to take on microfilm, microcard, microfiche, floppy disks, CD-ROMs, and the Internet. Acquiring digital materials in "non-tangible" form is our next step. It is true that most of us are not ready to start a big project next week, but what we can do next week is begin the process so that next year we'll be better equipped than we are today.

The stakes are high. If libraries do not accommodate digital materials, libraries will simply not have access to materials in the future. If we cannot select and acquire, we will not be able to organize and preserve. "Pointing" to materials that move around, change, and disappear is not the job of a library. The job of a library is to help users identify, locate and use information and the best way we have of doing that is acquiring the materials our communities need and organizing and preserving those materials for them.

Essentially, the choice we have is to remain libraries or not. Either we take on the responsibilities of digital formats, or we become museums of old technology and information. Either we ensure current and long term access to materials our users need by acquiring and organizing and preserving, or we abrogate that role, and hope someone else does it. But if we do not do it, who will? The government with a single collection for all – subject to budget cuts and political influence? The private sector that will charge us again and again for access to the information we've already paid for as taxpayers? Who but libraries can fulfill the function that libraries have always filled?

Roles and Functions

What are the different functions/roles?

- Identify
- Describe
- Organize

- Preserve
- Provide service

Who best to do each?

- ID, describe, distribute: GPO
- Select, organize, preserve, service: Libraries

How

Selection – How would you select digital materials? An easy way to approach this is to think of this as adding another format to your collection. You have already done this work by setting up your selection profile in the current environment. Think in terms of the size of your depository. Depending on the percentage you are currently selecting, you would receive and take responsibility for the associated documents electronically.

Format choices – We should think of "print" as a format choice whenever appropriate. This would lead us to several possible actions when a digital document is available as a printed publication (e.g., PDF documents).

- Buy it.
- Print out the digital version.
- Continue to lobby for a simple solution to this problem: If an agency sees the importance of a document being in print format, depositories should have the option of selecting the publication in print format.

When a digital document is not available as a print publication, but is "print-like" (e.g., PDF), there are several options:

- Store the digital publication on a Web server; link to it, providing an OPAC record that points to the original (or PURL) and to your local copy.
- Store the local copy on a public service PC.
- Print a copy

When a digital document is a complex HTML document, consisting, for instance, of multiple chapters, embedded graphics, etc., there are at least three options:

- Print it.
- Download it to a public service PC using "offline browsing" software.
- Use digital library software to acquire the files and manage their location; provide an OPAC record that points to the original (or PURL) and to your local copy.

When digital information is totally dynamic and Web "pages" are actually built "on the fly" by user request, query, profile, etc., there are at least two options:

- If the results of queries seem predictably static, capture the documents most often used by your community of users and save or print it as above.

- If the results are too dynamic (change too fast or have no particular "document" as output), obtain the database that drives the site and integrate that into your digital library.

Hardware – Questions of hardware and software are all questions of scale: from printing out documents, to acting as a node in a distributed system like LOCKSS, to large scale digital library projects like Counting California. Do you need a double-sided printer, a workstation that can handle the free LOCKSS software or a server to manage large collections of digital information? It is likely that the more complex and difficult a project is, the fewer libraries will undertake it. Not every University library is running its own Counting California project.

Software – We hear and read about all kinds of software. Greenstone, eprints, icollect, getbot and more. It might be helpful to start thinking about this issue in terms of what kind of project you want to do in your library. Are you going to focus on single documents as they come up or full agency Web sites? If it's the former, there is a whole genre of software that allows you to grab documents off of Web sites and store them on your desktop. They are similar to the software

that is used in canned Internet demos. Icollect and getbot are examples of this type of software. Projects like the one described by Barbie Selby dealing with documents from the Civil Rights Commission use this type of software.

If it is the latter type of project, the other branch of software focuses on server-based, digital library projects. Software like eprints and greenstone are examples of this. Counting California uses greenstone to grab information from Websites that is then stored, combined with other data from other sources and manipulated for presentation to a local community.

Roles for the Government Printing Office

In this model, the GPO would focus on the roles that it currently carries out, the identification, description and distribution of government information. By distributing the digital publications along with descriptive information (metadata) about the publications, the GPO will have taken a huge leap towards protecting the public's right to know. Without any other actions, just the multiple copies alone can improve the odds that if one copy is damaged or tampered with, users can go to other locations to find the same information. If the GPO goes the extra step of adding a 'hash' or 'digest' to the metadata record, there will be the additional level of integrity. Because the hash is produced by the GPO, the source of official government information and maintained by depository libraries, trusted repositories of government information, this integrity check becomes an authentication tool as well.

Reality

The GPO is not currently distributing digital publications to depository libraries. This does not mean there is nothing to do in the interim. In fact, libraries can and have been doing things to preserve particular documents for some time. Our assumption here is that an organized dissemination method would allow more libraries to participate and ensure preservation of this material in a more systematic way.

Libraries can identify publications they feel need to be preserved locally. Tim Byrne's presentation[1] at DLC last year and Raleigh Muns' presentation[2] in 1995 are both examples of libraries identifying and collecting certain publications for local storage and use.

Counting California is an example of a more sophisticated presentation of locally held digital collections of government information. Trisha will give the specifics, but if Federal information (like the Federal Register) were available electronically for deposit, this system could be richer.

Objections

Authenticity/Integrity – This is a big concern in the digital environment, so I'm going to take a minute to go through some definitions and clarifications. I would like to state up front that the concept of authenticity is a social construct that relies heavily on the notion of trust. It is also important to keep in mind that there is not one great technical solution to this social problem.

Clifford Lynch states that "a verification of authenticity is the act or process of establishing a correspondence between known facts about the record and the various contexts in which it has been created and maintained and the proposed fact of the record's authenticity."

- Testimony by a trusted third party
- Copies
- Register/Digest/Hash

Trust is more important and reliable than anything technical or digital. So, I would like to pose the question, who do you trust?

---

[1] Byrne, Tim, "The Regional Role in Permanent Access to Electronic Government Information," Proceedings of the 9th Annual Federal Depository Library Conference, Washington, DC: U.S. GPO, 2001: 1-4.

2 Muns, Raleigh, "Mining the Electronic Documents for Local Collections," Proceedings of the 4th Annual Federal Depository Library Conference, Washington, DC: U.S. GPO, 2005: 53-66.

- Government administrations subject to the whims of party politics and budget changes?

- Corporate entities whose interest may go away as it is no longer profitable, whose focus is on short-term profit rather than long-term preservation?

- Federal Depository Libraries

A hash or digest added by the GPO can help answer the following questions:

Has the object been changed since its creation? If so, has this altered the fundamental essence of the object? Integrity – the digital content has not been corrupted.

If its integrity is intact, are the assertions that cluster around the object true of false? In other words is it authentic?

Too hard. - Yes it is hard. In fact there are more dire "up front" issues with digital information. Things are not likely to be saved by accident as they can be in the print world. Decisions have to be made up front about what to save and how to save it. What is integral to the document that must be saved? But, if we don't start now, experimenting with digital collections and asserting our rights under Title 44, how will we keep up with issues as they become even more complex?

Why us? It's just another example of the GPO passing costs on to the depositories. – This isn't true, we're just holding on to the roles that we have as libraries. In fact if we don't demand that the GPO distributed these files, the result will be a pass off in responsibilities. If we allow the GPO to maintain the only electronic collection, they will be focusing on that rather than improving their record of identifying and describing government publications. As these roles slip, we, as volunteers, will be required to do more and more of this work. If the GPO does its work well, a locally held electronic collection in DC is a good addition to the network of depository collections around the country.

What about GPO mirror sites?
Mirror sites provide additional copies of publications, so they are beneficial in terms of preservation, but they do not offer any variety of collections. They do not address the issues of multiple collections, customized presentation and service.

What about 'Partners'?
Can be a very useful part of the whole. Especially if they distribute what they have including the documents and associated meta data.

My administration won't do it. - Libraries have come to provide services for all kinds of new formats. This is going to change business as we know it. We had to buy fiche cabinets and readers even microcard readers that don't work anymore and new scanners that do. We've been dealing with format issues all along and none of our budgets have increased to do so. It requires a rethinking of our budgets, repurposing of staff positions and departments. Who had a full time systems position twenty years ago in their library?
Why??

It's what libraries do. We select, acquire, organize, preserve and make government information accessible.

It's how we provide the best service to our communities – Counting CA

Risks of government control – funding instabilities, changing administrative priorities, sinister motives. In the current environment of centralized access, innovation can't happen.

Risks of corporate control – Where does the information go when it is no longer profitable? Issues of proprietary software limiting access to 'free' information.

We don't have a choice. If we want to remain vital institutions preserving the public's right to know, we must make the decision to select,

preserve and provide access to digital information for our local communities. The one size fits all approach or access versus ownership approach to library collections will make us dispensable institutions and underestimates the value of locally selected and presented collections.

Who else is going to act as a watchdog for the public's right to know?

In conclusion, the most important function we provide as depository libraries is providing permanent public access to government information. Any system must fulfill the criteria of serving multiple communities adequately; fulfilling the governmental responsibility to provide permanent public access to government information to all citizens, and finally, the government must not have exclusive control.

I encourage all libraries to experiment with digital projects, remaining vital institutions in your communities while advocating for a secure system of distributing digital publications to depository libraries.