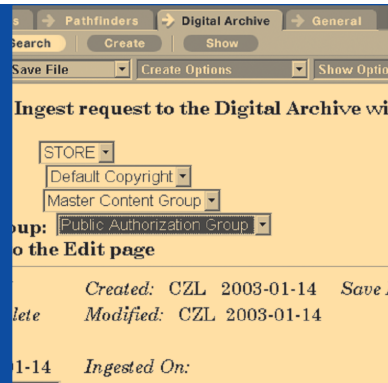


U.S. Government Printing Office &
Connecticut State Library

Connecting Libraries with Government Information

— Preserving electronic public documents



*Learn how federal and state librarians capture,
catalog and preserve Web-based documents.*

The Government Printing Office and the Connecticut State Library in Hartford provide public access to official documents

“These new digital preservation tools represent a whole new world.”



George Barnum is the electronic collection manager for the U.S. Government Printing Office (GPO), Library Programs Service, in Washington, DC.

Mr. Barnum discusses long-term access, management tools and preservation for digital materials at the Government Printing Office.

The GPO: Managing the “Printed” Word

What is GPO’s role in government electronic publishing?

The GPO was the center of the federal paper document universe for a long time. Our role was, and continues to be, printing documents and getting information into the hands of users. The GPO has the responsibility for gathering official documents of all three branches of government, cataloging them and making them available to the American public over time through the Federal Depository Library Program.

In the world of printed documents, distribution to designated depository libraries was a part of the greater production stream. When agencies secured printing or printing services from GPO (as the law requires), we added or “rode” sufficient copies to the print order to accommodate library distribution. That system worked relatively well in its native environment.

In the digital environment, that single stream for acquiring content is only one of many. Agencies have no requirement to come to GPO when making information available on the World Wide Web. Our goals remain the same, but the channels by which we obtain the information are far more complicated. Agencies are changing the systems by which they make information available, which was formerly a big paper trail process that ended up on our doorstep. So while there’s the same or greater expectation that information will be available and will be official, it’s far more complex for us to find and process the publications.

What types of documents is GPO responsible for?

The collection development scope for the depository program is huge, and is defined in law: Government publications must be produced at public expense, have public interest or educational value, not be for strictly administrative or internal use (such as office procedures) and not be classified for reasons of national security.

This statutory framework now has to be applied much more critically than when the printing procurement process delivered publications to our distribution line. To build our electronic collection we must first attempt to determine if a Web resource meets these criteria and therefore falls within our scope. Once the selection decision is made, we archive publications for which no tangible equivalent was distributed to depositories. Since 2000 we have archived over 7,000 such publications.

How was GPO involved in the Digital Archive Project?

A few years ago we worked with OCLC and the Department of Education on a project for electronic publications that started out with a preservation emphasis, but didn’t really end up there. At its conclusion we sat down with OCLC staff to do the typical “postmortem,” in which we finally articulated our principal goal, summarized as “Build us an archive and the tools to use it.” From that starting point we began to draft a set of high-level requirements—which later led to operational specifications—and ultimately to the Digital Archive. It’s a repository for web-based documents that provides an environment and tools for storage, access and preservation. We began pilot testing, with the other five pilot libraries, in mid-2002. Documents are archived in HTML, PDF and a few other supporting formats. We created a flexible process for bringing objects into the archive that can be adjusted as needed.

How is GPO’s workflow changing?

Again, as our stream for acquiring publications changes, our work patterns change. Formerly acquisition, distribution, cataloging and processing all took place in a rigidly set order. Now, because we “acquire” in so many ways—from suggestions made by people in our user community, to cues we find in printed documents, to notification from publishing agencies—the order in which we catalog and archive may be

quite fluid. We need to discover, capture, establish reliable access, and preserve files to sustain access over time. With the Digital Archive tools, all the individual steps can be moved around as needed.

The key to this preservation is the metadata we create specifically to support the task, separate from other kinds of metadata used for retrieval.

Tell us more about preservation metadata.

We began hearing that term sometime in 1999 or 2000 and it still means different things to different people. The essential concept is that there is a set of attributes that should be recorded in order to give people and machines information that will be required to perform preservation processes like emulation or migration. It's really related to what's described as "administrative" or "technical" metadata, in that it's not used for access or retrieval.

OCLC and RLG have just convened a second preservation metadata working group that will continue to develop definitions and specifications.

You've talked about individual documents, what about Web sites?

By statute our concern is with publications, so that's our focus. We do, in some cases, catalog and harvest entire sites as a precursor to mining them for individual publications.

What happens if a document is removed from the Web?

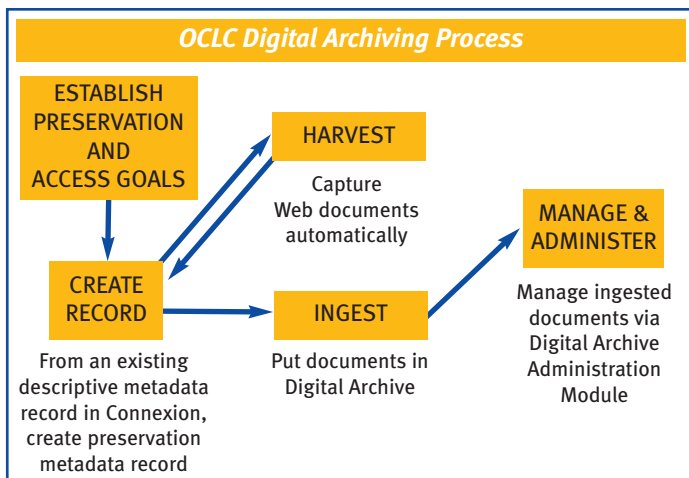
When we discover that we're pointing from our catalog to a URL that no longer works, we first must determine why. If we aren't prevented from providing continuing access (for reasons of national security, for example) we will serve up the archived copy, and we inform the user that that's what they're getting.

Are electronic documents replacing paper?

We lean heavily toward electronic-only dissemination in every case we can, because it's what our users are asking for and because it's usually more economical. We've come as far in the last five years as saying that between 60 and 70 percent of what's disseminated in the depository program is electronic. Our obligations to provide permanent access remain the same. Thus the need for new and better tools.

And what's next for the GPO?

Public Printer Bruce R. James has charged the GPO with being preeminent in federal information dissemination. We're actively developing new lines of business that complement what is already successful. We're now looking at ways for agencies to make a single "drop" of content which we can then turn into a variety of output products, both online and analog, perhaps helping agencies to fulfill other requirements (such as their deposit obligations for the National Archives) at the same time they fulfill the depository library ones. We won't be a single stream any more. And the basis of all this is permanent, reliable, no-fee public access.



The Digital Archive in Action

"We have more options and more ways of thinking and experimenting [with digital archiving and preservation] with the OCLC products."

—Stephen Slovasky



Stephen Slovasky



Julie Schwartz

State Creates "Connecticut Digital Archive"

Like the Government Printing Office in Washington, DC, the Connecticut State Library in Hartford provides public access to official documents. Julie Schwartz is the library's unit head for Government Information. Stephen Slovasky is the unit head of Bibliographic Services.

How did the Connecticut State Library join the OCLC Digital Archive pilot project?

JS—"In 1999, the staff here became intrigued by the OCLC Cooperative Online Resource Catalog (CORC). OCLC called this project a 'cooperative creation of a catalog of Internet resources'. Our library joined the CORC pilot project in 2000. OCLC has since incorporated CORC and other services into Connexion™."

"We first heard about the Digital Archive at a conference and I saw this as a great next step and a great fit for government publications. We were really fortunate to get into OCLC's pilot project, as well. It was good timing."

(continued on back)

Did your state archive project require special funding?

SS—“It required special funding, but we didn’t get it. [Laughs.] No, actually the Connecticut Digital Archive project is part of the library’s continuing mission. The tools developed are now part of the workflow. The library can absorb the cataloging and archiving of digitally distributed official publications because we have staff who have been accustomed to cataloging such material since the early days of the CORC project.”

What type documents do you archive?

SS—“One principle of the Connecticut Digital Archive is that we don’t archive anything that is an exact electronic representation of a print copy that we already have here. But defining which electronic documents should be preserved can initiate debate and discussion. Our guidelines say an electronic document selected for cataloging and preservation should continue a publication previously produced in a traditional format, or be a document that is ‘born and lives only electronically.’”

What does electronic-only or “born-digital” mean in this context?

SS—“The state documents we archive must report on or describe an agency’s activities; be legislatively, judicially or administratively mandated; or meet various other criteria. Many types of electronic documents fall within the parameters, either by mandate or staff decision. With budgetary restrictions as they are, there will be even more of them.”

What’s a good example of this type document?

JS—“One is a series of reports from the state Office of Legislative Research. These reports are issued perhaps several times a day when the legislature is in session and several times a month when it’s adjourned.

“These are two- to four-page reports prepared in response to legislative inquiries. We find them very useful for reference purposes, because they summarize key issues and contain statistical information. Often they’ll include a précis or a historical summary of laws on a particular topic.”

What is your workflow for capturing the documents in the Digital Archive?

SS—“We have a pretty effective workflow for harvesting these reports and uploading them for preservation in the Archive. The legislative reports constitute the majority of the nearly 3,000 titles we’ve archived to date. We catalog and archive them at the document level—that is, if content changes, both versions will be in there. It’s really the catalog that’s controlling the access to the archive.”

How does a listing appear in the Connecticut State Library’s Online Public Access Catalog (OPAC)?

JS—“Our OPAC (www.consuls.org) listing includes two links to each of these reports. One directs users to the legislature’s public Web page, the other to the Connecticut Digital Archive itself. This redundancy anticipates that documents eventually will be removed from the legislature’s site.”

Where do the documents actually reside?

JS—“The Connecticut Digital Archive copy resides on an OCLC server, not in Hartford. The transition between servers is transparent to the government and public users accessing the information. The OCLC Digital Archive also ensures long-term access to the data, even if the technology should change.”

SS—“We’re taking responsibility for keeping a report and making it available to whoever needs it, now and in the future.”

Is your workflow solidly in place now?

SS—“We’re about 95 percent nailed down in terms of workflow. Other things we have to take care of as they come in. We have more options and more ways of thinking about and experimenting with the different OCLC digitization and preservation products. It keeps getting better.”

OCLC Digital Collection and Preservation Services *at a glance*

Range of Services

OCLC offers a full range of digitization and preservation services to help you plan, digitize, manage and preserve your digital collections. These services include:

- A **Digital Archive**, to make it easy to preserve your collections
- Grant-writing assistance
- Digital collection management software such as **CONTENTdm**
- **Preservation Service Centers** for full service digitization

To Learn More

Visit us at:

www.oclc.org/digitalpreservation/

Contact:

OCLC Digital Collection and Preservation Services

OCLC Online Computer Library Center

6565 Frantz Road, Dublin OH 43017-3395 USA

Tel: 1-800-848-5878 ext. 6251

Fax: 1-614-764-0155

E-mail: libservices@oclc.org

www.oclc.org

The following OCLC product and service names are trademarks or service marks of OCLC Online Computer Library Center, Inc.: OCLC, Connexion, CONTENTdm, Olive Software. Third-party product and service names are trademarks or service marks of their respective companies. In addition, the OCLC symbol is a service mark of OCLC. OCLC grants permission to photocopy this publication as needed.