Web Documents Digital Archive Pilot Project (OCLC): Arizona

Janet Fisher
Arizona State Library, Archives and Public Records
Phoenix, AZ

The State Library in Arizona has recently signed on to the OCLC Web Preservation Project. The responsibility for this falls to another person in my agency, our Director of Electronic Government Information. I am here today to present some of his ideas and to describe our thoughts and efforts to capture Arizona's electronic government information for the future.

The Arizona State Library, Archives and Public Records is mandated to preserve state agency materials, in all formats.

Currently, the agency has three responsibilities for state information. We:

- retain permanent copies of selected unpublished state government records in our Archives Division,
- assist agencies in creating retention schedules and housing those retained files for the required time periods in our Records Management Division, and
- obtain copies of state agency publications for long term retention in the Law and Research Library Division.

We are in a unique position to have all of these activities under the leadership of one agency. The coordination of projects and any difficulties dealing with overlaps between these programs is more easily solved within one agency than it might be if we were in separate agencies.

In the Law and Research Library Division, we have been working with print publications for many years. With varying success, we have been collecting, cataloging and providing access to these print publications.

For the past year, we have been piloting a Government Information Locator System (GILS) program which spiders 105 state Web servers (approximately 190,000 Web pages).

But we have been searching for a way to preserve the state agency Web publications. We have looked at the work of the National Archives and at the efforts of other states; and we have looked at the private sector. We have come to this research project of OCLC to continue our growth and to start working on a caching project for state information.

We do not want to save everything in Web space.

- Given the challenges of preserving digital materials, we do not want to spend our resources saving electronic copies of information that is already being preserved in another, more stable, format (e.g. paper).
- We do not want to save ephemeral material: the electronic equivalent of the "wet paint sign."
- We are not interested in saving every view of the Web.
- We are not interested in saving transactions performed on the Web – for example, renewing a driver's license – because those transactions will be captured in another record series (the underlying database).

As an agency we want to preserve information that is accessible to the public only on the Web, that provides evidence of what an agency published and which may be used by the public in making a decision, and demonstrates an agency's accountability.

Some of the problems that we are encountering are that:

- Some permanently valuable publications are going on the Web without a paper copy being deposited at the Library.

- The distinction between records and publications is blurring on the Web, so our need to retain both is good, but length of retention is different. For example, records have a specific period of retention time as defined by a retention schedule (perhaps 3 years, or longer, before destruction), where a publication may be kept for a much longer period of time.

- Agencies need some way to demonstrate the state of information provided to the public via the Web in case of litigation (evidential value).

Throughout all of this, we must be able to meet out statutory responsibility to preserve state agency publications. We can't rely on agencies to preserve their Web publications; that's our mandate, not theirs. We also have to recognize that agencies are not prepared to provide reference service to their old Web sites, which may be stored offline.

We have joined the OCLC Preservation Project to see how we can capture and provide access to these publications. One thing we need to determine is how much and how often we capture the information. Rather than trying to capture every page, we are looking at a way to scan and capture (or get a snapshot of) state information on state servers at specific periods of time.

How would we describe the cache? What is the scope of the cache?

The cache should include only documents that meet the following five criteria:

1. Reside on a state server. Our primary list of servers to work from are those that are currently spidered by our GILS.

2. The information has been created by or on behalf of a state agency as a product of that agency's mission. As a result, the state should have all rights to the content of those pages.

3. We have considered giving precedence to Web pages that are roughly equivalent to printed publications. Such Web publications are typically more than a few pages and may be organized into chapters. They are not ephemeral; their value is measured in months rather than days or weeks.

4. Web publications that are preserved in another format may be deselected. For example, Web versions of student/faculty directories may be excluded if a print version is preserved.

5. The whole document should be preserved, including graphics and audio clips. Links to other pages in the cache should be modified so that they continue to point to a contemporaneous version of the linked page. Links to pages that fall outside the scope should be flagged as a link outside the cache that may no longer work and on which the information may have changed.

Ideally, all Web publications that meet the preceding criteria would be cached. We are looking at capturing state Web materials four (4) times a year, at a minimum. When we go into full production, we will investigate monthly captures, and we will ask that agencies leave pages up for at least a month to give us a chance to capture them.

One of the points in a methodology for Web space is to involve Webmasters in the process and to stress the use of metatags. We have already begun training of state Webmasters in metatagging for their sites for successful searches using our Government Information Locator System. In addition to helping point to current Web locations, the metatags can be used as identifiers for these documents in the future (and may include additional information to describe the electronic information)

We have considered tags that would facilitate automated caching of those pages for enduring access. Those tags we have considered include:

- Unique Identifier (a unique alphanumeric code, similar to ISBN or ISSN, to identify a document as a whole)
- Version Control Number (each revision of a document would receive a new Version Control Number, not a new Unique Identifier)
- Sequence Number (a number indicating the preferred order for linear presentation of a document.)
    - For a print-like document, the sequence number is roughly equivalent to page numbers.
    - For hyperdocuments, the number may reflect a branch structure or some other organizing principle.

  The sequence number establishes the order in which pages should be printed or output to microfilm.
- Retention (indicates if the document should be cached through three possible values: ephemera, deselected, permanent)

How will we retrieve materials from the cache?

If the cache is stored on a Web server, the Government Information Locator System can index and provide access to the contents of the server in the same way it indexes and provides access to other Web sites. We have also considered using a proxy server combined with a database to ensure that links point to other documents contemporaneous with the document being viewed, rather than the most current version. For example, clicking on the Home button on an archived page would take you to the Home page current when the archived page was served.

It is our hope that such a logical methodology will be viable and assist us in the retention of Web-based state information, for the short term. The puzzle of electronic information – here today and gone tomorrow – is one that needs to get solved. We need to be able to say it is here today, and those materials of enduring value will also be here tomorrow.

These are the concepts we are carrying forward as we join in with others to face the challenge of preservation of electronic documents.

An additional effort we are making in Arizona is to come to terms with electronic records. We have convened a group of representatives from state agencies and local governments, called the "Arizona 'Lectronic Records Task Force" (ALERT) to look at public records, not publications. This group is working on a methodology and standards for handling electronic records. We are starting with the need for a common vocabulary, and building discussions about retention of electronic records.

The challenges brought with born digital information do not have one answer, and so we are trying to address various angles and groups who are involved in the creation and later referral to these information sources.

We are looking at the Web caching possibilities of this OCLC Web Preservation project as one of the pieces in the puzzle of preserving state government information. We look forward to the experience we gain, and lessons learned in this process.