

Sample Talk for Preserving Web Sites and E-Pubs
March 2007

My name is Daniel Cornwall and I am the Government Publications/Technical Services Librarian for the Alaska State Library. I have a number of duties including being the coordinator of the Alaska State Publications Program. Over the years we have seen the proportion of state publications issued in electronic format rise to about half of all state publications. Some departments are more electronic than others.

We did our first “web preservation” project in December 2002 when we took a web snapshot of the Knowles/Ulmer Gov/Lt. Gov pages and issued that on a CD to the Alaska Depository Libraries [Show CD]. There are Alaska State Depository Libraries in Anchorage, Fairbanks, Juneau and Washington D.C.

Since mid-2004, we have taken monthly web snapshots of all publicly accessible agency web sites known to the State Library. We use open source software called Capturing Electronic Publications available from the University of Illinois – Urbana Champagne.

We will shortly be issuing DVDs of the Murkowski/Leman Gov/Lt. Gov pages based on our own web gathering and on data provided to us by the previous administration. Both the Knowles/Ulmer and the Murkowski/Leman web snapshots are as complete as we could make them. With the vast majority of agencies, this is not the case.

Web site preservation is a sideline for the Alaska State Library. Our statutory mandate in AS 14.56 is for publications. Our web snapshots are incomplete as we’ve had to forgo graphics files to conserve storage space. Our web snapshots are stored in compressed format behind the state’s firewall. They can be reconstituted on request.

So why take web snapshots if not to exactly preserve the exact look and feel of agency web sites? For two main reasons – current awareness and as a fallback. The software we use produces a list of files added to the agency website since the last gathering run [Show SLIDE1 of E-Mail (ISER?)]. We are not aware of any other resource within the state that does this. I and one other library staff member comb through these monthly lists looking for new electronic documents. It is not a straightforward and quick process. Many new files are forms, posters, regulations and other items that are not part of the State Publications Program. Other files are fragments of larger documents and must be reconstructed to be preserved. In other cases, we already have the printed version of the online publication. We gather web snapshots from roughly 80 agencies, divisions, state corporations, etc each month. On average it takes about 20 hours each month to identify new electronic publications. We are not unique in these challenges. The federal Government Printing Office released a web harvesting white paper in February essentially saying getting files is easy, but identifying electronic publications is hard.

The other reason we take web snapshots of agencies is so that we can go back to an older copy if a publication slipped through the cracks and was taken off the agency web site before it could be brought within the Alaska State Publications Program. This practice

may also allow the Alaska State Archives the ability to mine prior versions of web sites for electronic records.

Having explained a little of what we do to gather and identify electronic state publications, let me cover what we do with them. Ironically, first and foremost we obtain a print copy of any “document-like” object. First we try to get a published copy from the agency. If that isn’t available, we will print off a copy of the document. Preservation of digital items is still in its infancy and no one knows what will work. Having a tangible copy is currently the only iron-clad guarantee that a document will stay accessible. Even the Census Bureau is taking a two-pronged approach to preservation for their 2000 Census questionnaires that of producing an electronic file and microfilm. They did this to ensure that something would be available to the genealogists of 2072 when the 2000 Census questionnaires become available to the public. I feel confident that if the feds had a guaranteed method of digital preservation, the Census Bureau would not have gone to the expense and trouble of producing microfilm.

But we also have an electronic approach. For near and medium term accessibility, we place a copy of every electronic state publication cataloged since July 2005 on our library web server. We have found that this practice dramatically reduces the instances of broken URLs resulting from agency web site changes, division mergers, etc. Aside from stabilizing a publication’s URL, placing cataloged electronic state publications on our web server allows Alaska State Publications to be collected and stored in the LOCKSS system.

LOCKSS stands for Lots of Copies Keep Stuff Safe and is a network of library servers around the world, but mostly in the United States and Canada. Basic information about LOCKSS can be found at <http://www.lockss.org> [SLIDE 2 – LOCKSS site] . The Alaska State Library started working with LOCKSS as part of a Government Printing Office pilot study of preserving federal e-journals, but quickly recognized its potential for preserving electronic state publications.

LOCKSS works by caching identical copies of materials in multiple computers and using a polling system to verify that stored content is unchanged. If a copy of a publication in one server is accidentally corrupted or intentionally damaged, that server will lose the “integrity vote” and will repair its content from other LOCKSS caches. This cooperation between the LOCKSS caches avoids the need to back them up individually. The LOCKSS software team is based at Stanford University and is currently researching ways to migrate content from one format to another without user intervention. This, combined with its automated error checking makes it a promising tool for long-term preservation.

The Alaska State Library began using LOCKSS to store Alaska State Documents in October 2005, reaching back to publications cataloged in July 2005. Since then, over 1000 documents have been stored in our LOCKSS cache [SLIDE 3 – ASL LOCKSS] and in the caches of over 30 institutions. On our LOCKSS cache, a year and half worth of state documents, plus a year’s worth of ten federal e-journals has taken up just one percent of our storage space.

No other state currently uses LOCKSS for the preservation of electronic publications, but we have fielded questions from several other states and one Canadian province.

Challenges remain. [SLIDE 4 – Crystal Ball] Like the Government Printing Office, we have trouble keeping up with all the new electronic material coming from agencies. In a perfect world, state agencies would clearly mark files as electronic publications and deliver them to the library, greatly facilitating their description and future preservation. We in the information preservation community do not yet know what approach will be best so we will need to hedge our bets and continuously educate ourselves about best practices in digital preservation. But we feel that we have made a start.