

Capturing E-Publications (CEP) of Public Documents

Fourth Semiannual Report: March 2005 to September 2005

| | |
|-----------------------------------|--|
| Organization Name: | <i>Illinois State Library (ISL)</i> |
| Project Name & Number: | <i>Capturing Electronic Publications (CEP) LG-02030120</i> |

| |
|---|
| Program Influencers |
| <ul style="list-style-type: none"> • Graduate School of Library and Information Science (GSLIS); • Participating state libraries; • State of Illinois agency personnel including webmasters and public information officers; • Illinois Librarians, including government document librarians; • ISL and University of Illinois at Urbana/Champaign (UIUC); and • State government information seekers |

| |
|--|
| Organizational Mission |
| <p>It is the mission of the Illinois State Library to provide state government officials and employees with the information they need to make informed decisions as well as to develop and promote libraries in order to enrich the quality of life for the people of Illinois.</p> <p>Inherent in this mission is the State Library's advocacy of the right of Illinois citizens to read and have full access to information.</p> <p>The mission is accomplished by sharing library resources through the Illinois Library and Information Network.</p> |

| | |
|---|--|
| Program Purpose | |
| Summary of key proposed services | <ol style="list-style-type: none"> 1. Refine existing methodology for permanent retention and public access to state government electronic publications; 2. Facilitate the replication of this methodology by other states; and 3. Create increased access to Illinois government documents through MARC records. |
| Target population | Illinois State Library; other state libraries and state archives; citizens seeking information; and Illinois librarians |
| Target outcomes | <ol style="list-style-type: none"> 1. Other state libraries and archives will implement the system. 2. Illinois Document Depository librarians will use the public access of the ISL system. 3. Other state libraries and archives will use the system. 4. MARC records will be created for Illinois state documents in electronic format. |
| Planned Program Activities | Progress in the fourth reporting period of the grant |
| <i>Creation of the metadata generator</i> | The metadata generator has been completed. However, adaptations for authority control of publisher name changes are being developed. |
| <i>Hiring and contracting staff</i> | New contracts are in process for October 2005-September 2006. |
| <i>Change monitoring and caching e-publications</i> | Change monitoring and caching e-publications is ongoing for Illinois, Utah, Arizona and Alaska in their respective states. The system was implemented in North Carolina in late September, as opposed to being handled for them by UIUC. Montana's system has been hosted at UIUC until the end of grant funding, September 30, 2005. Wisconsin has never led anyone to believe they |

| | |
|--|--|
| | <p>have the resources to implement or maintain the system. Now that the grant funding has ended for UIUC to operate the system for Wisconsin, change monitoring and caching has ceased for Wisconsin e-publications.</p> |
| <i>MARC cataloging</i> | <p>Our cataloger for this project has returned to work following recuperation from serious health problems. She is prepared to catalog documents as they are deposited. No cataloging has occurred in connection with CEP as yet because no documents have been deposited to date.</p> |
| <i>Maintaining server list</i> | <p>Alaska: Many new pages appear in existing agencies, so that they are automatically harvested. However, the media is monitored for new programs with possible new Web pages.</p> <p>Arizona: Maintaining the spider lists is a fair amount of work, although a lot more time could be spent tweaking the spider by giving it hints about good and bad directories. But one of the nice things is that the software does a reasonable job in most instances, so if you don't have time to tweak the lists are still in fairly good shape.</p> <p>Illinois: The graduate assistant at UIUC running our harvesting spider notifies ISL when any new links appear from the spidering process. Also, the news media is monitored for new state programs that may have Web sites.</p> <p>North Carolina: NC acknowledges that it is a challenge to keep up the server list, but they believe their core list good. The biggest difficulty is keeping several lists current--for CEP, for a pilot project with Archivelt, and for their NDIIPP project. They are currently relying on students and the grapevine to help in this area.</p> <p>Utah: They feel that they have a very workable process of maintaining server crawling lists. They are able to contact the state CIO for a list of new state Web sites.</p> <p>Wisconsin: They help keep state Web site lists for the state search engine, so their lists are as current and complete as possible. Wisconsin has added local and university sites for the state portal. The toughest issue has been discerning which contract organizations and non-profits that work with state should be included.</p> |
| <i>State agency training</i> | <p>A training session was held in late April 2005. Illinois Secretary of State employees attended the morning session, including ISL employees. ISL Technical Services staff attended the afternoon session.</p> |
| <i>Mentoring other states</i> | <p>Alaska reports that they have been very pleased with support received during the grant, especially with data backup issues and conserving disk space. Larry Jackson was very responsive.</p> <p>Arizona says Jackson has been a dream.</p> <p>Montana commented, "The level of support and customer service from Illinois was exceptional from start to finish. Everyone from the administrators of the grant project to the graduate students who set up and monitored the spiders gave us prompt assistance."</p> <p>North Carolina felt that Jackson was very responsive to their needs.</p> <p>Utah had minimal contact with Illinois support because they were so tied up with other projects. However, they had no complaints because all of their questions were answered satisfactorily.</p> <p>Wisconsin: Jackson was knowledgeable and helpful.</p> |
| <i>Coordinating steering committee</i> | <p>The steering committee met by conference call October 12, 2005. Those states that could not participate responded by email concerning their project experiences.</p> |
| <i>Keeping statistics</i> | <p>Statistical reports are online:</p> <p>Alaska http://pep.library.uiuc.edu/~cep/stats/AK/LatestStats.html</p> <p>Arizona http://www.cyberdriveillinois.com/departments/library/who_we_are/pdfs/ArizonaStatistics.pdf</p> <p>Illinois http://history.lis.uiuc.edu/~cep/stats/IL/LatestStats.html</p> |

| | |
|---------------------------------|---|
| | <p>Montana Not available since the end of hosting at UIUC.</p> <p>North Carolina http://depo.library.uiuc.edu/~cep/stats/MT/LatestStats.html</p> <p>Utah Not available at this time.</p> <p>Wisconsin http://pep.library.uiuc.edu/~cep/stats/WI/LatestStats.html</p> |
| <i>Reporting and evaluation</i> | This document constitutes the fourth report using the OBE method. Data is taken from computer statistics and personal reports of experience by phone or email. |

| Planned Program Services | Progress in the fourth reporting period of the grant |
|---|--|
| <i>Metadata generator</i> | The issues of communication between the generator and the depository have been resolved. Final testing occurred in April with all systems working without problems. Enhancements have not ended as development continues. |
| <i>Website</i> | http://www.cyberdriveillinois.com/departments/library/who_we_are/cep.html ; also see http://www.isrl.uiuc.edu/pep/ |
| <i>Agency training sessions and materials</i> | Web-based training tools for Illinois continue to be modified in response to continued development and enhancement of the system. Both Web-based and paper-based materials will be available for agency personnel. Agency training has been delayed for two reasons. 1.) The principal investigator recommended some changes to the depository for better management of migrated digital objects. 2.) Computer programming changes were deemed necessary to ensure authority control over the names of publishing agencies. The CEP project grant period has been extended to September 2006 for the purpose of agency training. |
| <i>Mentoring</i> | DVDs have been shipped to all participating institutions with a back up of the latest version of the software and a CVS compressed copy of the latest version-controlled harvest. The participant institutions are discussing maintaining discussions in the future. |
| <i>Steering committee meetings</i> | Conference calls for representatives from participating state have been held to share experiences and discuss issues or questions. Feedback from the latest call and email is the basis of the narrative, which follows. |
| <i>Refinements to the software</i> | <ul style="list-style-type: none"> • The statistical package is operational for participating states except Alaska and Utah. Details follow in the narrative. • Additional manager functions and bug fixes are included in the software distribution. • Illinois is OAI compliant, and OAI server software is included in the software distribution to all the participating states. . |
| <i>Implementation in other states</i> | Implementation has been completed in Alaska, Arizona and North Carolina. Upgrades have been installed for Arizona and North Carolina. Upgrades for Alaska and possibly Utah are planned for the extended grant period. Montana and Wisconsin have not had resources to implement in their states. |

| | |
|----------------------------------|--|
| <i>Reports and presentations</i> | <ul style="list-style-type: none"> • A presentation to the Annual Illinois State Library Government Documents Conference in May 2005 sparked interest in the depository. • The project manager presented to GODORT at the annual ALA meeting in June 2005. Over 100 librarians from across the country attended the session. • An article has been submitted to <i>Internet Reference Services Quarterly</i> based on the ALA presentation, which is scheduled for publication in November 2005. • In October 2005 Richard Pearce Moses spoke at the American Research Libraries forum on government documents. • Larry Jackson's paper, <i>Difficulties in Electronic Publication Archival Processing for State Governments</i>, was accepted for the 1st International Conference on Universal Digital Library, 2005. |
|----------------------------------|--|

| |
|--|
| <p>Target Population</p> <ul style="list-style-type: none"> • State libraries and state archives charged with preserving state government documents in electronic formats; • Citizens seeking state government information; • Illinois librarians using state government information to meet patron needs. |
|--|

Outcome #1 Other state libraries or archives will implement the system.

| Indicator | Data Source | Applied to Whom | Data Intervals | Target |
|--|--|---|--|--|
| Number of state libraries and or archives that complete two Web harvests of their state's documents and cache the resulting files. | Project staff will contact participating institutions by email or telephone. | All institutions that return signed participants' agreements. | Three months after system implementation | At least three of the participating institutions |

Progress:

- **Utah** has operated their PEP system since December 2004.
- **Montana** sites have monthly harvests from February to September 2005 hosted at UIUC.
- **Alaska's** first month of operation occurred September 2004 and has continued every subsequent month. Hardware was purchased expressly for this purpose, so plans to continue operations are firm. Alaska's challenge is to formulate plans in preparation for autumn 2006, when it's estimated the hard drive currently used for the project will reach no longer have unused capacity.
- **Arizona's** system has experienced successful harvests, change monitoring and caching every month beginning July 2004.
- **North Carolina's** system was hosted at UIUC and successfully operated on a monthly basis beginning August 2004. In late September 2005, the PEP system was installed in North Carolina along with all the data UIUC had harvested for them in over a year. They have 3 staff members who will share harvesting duties and received about 4 hours of training. They currently harvest 185 websites, pulling down approximately 98 GB per month. Their CVS archive spans 223 GB,

which filled 52 DVDs. UIUC will temporarily store a backup copy of their material.

- **Wisconsin** has been hosted at UIUC throughout the project. This participant has known throughout the project that they were unable to devote resources to a PEP system after the grant period. However, they are looking at uses for the data gathered during the grant.

Outcome #2 Other state libraries and archives will use the system

| Indicator | Data Source | Applied to Whom | Data Intervals | Target |
|---|---|---|--------------------------------|--|
| Number of state libraries and archives that place access to the system on their website | Project team verifies with each state by looking on the web | All institutions that return signed participants' agreements. | At the end of the grant period | At least two of participating institutions |

Progress: Development of tools for public access to PEP-harvested files has proved to be one of the project's greatest challenges. Therefore states have not placed access to the system archive on their websites. However, all the participating states are using the system or its harvested archive in some manner. As of July 30, 2005 the combined systems of PEP software running under the CEP grant had harvested 1043 websites with 463.9 gigabytes of download in 2.6 million files. The CEP Web site, which is located at http://www.cyberdriveillinois.com/departments/library/who_we_are/cep.html, lists participating states with links to the corresponding statistical tables. These tables list each state Web site harvested by agency, Web site size, file types found, and metadata type and quantity used. The statistics reports links above also connect to the same reports for evidence of use of the system by participating libraries.

Outcome #3 Illinois Document Depository librarians will use the public access of the ISL system for patron information needs.

| Indicator | Data Source | Applied to Whom | Data Intervals | Target |
|---|--------------|--|--------------------------------|---|
| Number of Illinois Document Depository Libraries which have staff who use the public access interface | Email survey | Illinois Document Depository Libraries staff | At the end of the grant period | 55% of the Illinois Document Depository Libraries |

Progress: The Illinois Document Depository Libraries met at ISL, and heard a presentation concerning PEP-based services and the coming depository on May 6, 2005. The depository interface was demonstrated. The depository librarians from across Illinois gave very positive comments. A survey will be done at the end of the grant to determine use.

Outcome #4 MARC records will be created for Illinois state documents in electronic format.

| Indicator | Data Source | Applied to Whom | Data Intervals | Target |
|--|-------------------------------|-----------------|---------------------------------------|--------|
| Number of repository URLs in 856 fields of | Bibliographic database report | ISL catalogers | For each grant report after the first | 50 |

| | | | | |
|------------------|--|--|--|--|
| ISL MARC records | | | | |
|------------------|--|--|--|--|

Progress: Testing in mid-April was entirely successful, and training was completed for Secretary of State employees in late April. However, revisions in the depository to better accommodate migrated versions of deposited digital objects in the future have delayed the start of deposits of official documents for cataloging, as opposed to test data. ISL intends that this outcome be actualized by the end of the extended grant period.

ISL Technical Services staff members are guiding the CEP project in assuring future interoperability between metadata and MARC records for the same publishers of state documents in electronic and print formats. An authority file for metadata will be developed and implemented in the fall of 2005.

Narrative

Open Archives Initiative (OAI) Compliance

As of mid-September 2005, the Illinois CEP OAI server is running and has been added to the UIUC public registry of OAI metadata providers. Please see <http://gita.grainger.uiuc.edu/registry/searchform.asp> for the registry. The "OAI server" is actually a routine CGI script running under a web server, not a separate program or protocol. To find Illinois CEP files using OAI, go to the above URL and search for "lsjackso." It's also possible to search for "PEP", but that returns many irrelevant hits. Even more direct, go to <http://gita.grainger.uiuc.edu/registry/details.asp?id=1369>, our individual listing.

On the registry site, the entire State of Illinois Web presence is divided into sets, which equate to individual CEP spiders. These are named with the spider's name, which is an abbreviation of the website's name.

The "ListIdentifiers" links are named using the spider name plus an accession number. OAI is primarily for communication between metadata databases, so the name isn't particularly human-readable. But, it is unique, which is the main consideration. Clicking on "ListRecords" will return multiple records. Choosing a website with a *small* number of records is advised for demonstration purposes. Some sites have scores of thousands of records, which will all list in one huge "ListRecords HTML" report. Again, this function was intended for machine-readability, not human access.

This server is providing access to only the contents of the most recent harvest of the Illinois Web sites. Even so, it's one of the largest OAI datasets. Attempting to expose everything in the archives would probably crash it. The Illinois CEP files crashed UIUC Grainger Library's Java version of an OAI server last year before all the then current websites were exposed. This URL will be posted to various lists of OAI providers.

To operate an OAI server after installing CEP, other states or organizations only need to go into their spider configuration files and set OAI to "yes". Until they do that, they're running an OAI server with no outgoing metadata. Once "yes" is selected, metadata starts accumulating with each spider that gets accepted into CVS for permanent retention.

Thanks to Ms. Yiyi Zeng, who programmed the CEP part of this, using Hussein Suleman's public PERL CGI script. Tom Habing, and others, at UIUC's Grainger Engineering Library assisted in testing and specifications also have earned our appreciation.

Upgrade Installations

Although, not all the travel to install software upgrades was completed by September 30, 2005, the principal investigator plans to go to Utah and Alaska for the last grant-covered software upgrades during the coming, final year of the grant period. All CEP operations for the other States are concluded.

The Arizona software upgrade was done remotely over the Internet. Alaska wanted a software upgrade, which requires a visit, but there are no open airplane seats to Juneau left in October 2005. The upgrade was installed at North Carolina in September 2005. Utah and Montana have to sort out local issues as reported elsewhere in this document.

Access Tools

Alaska still provides the access to their PEP archive as described in the last report. That is, using the Samba software that came with their Redhat Enterprise Linux operating system, they have created a READ-ONLY shared volume consisting of just the files in the compressed CVSRepositories (Controlled Versioning Software) directory. Daniel Cornwall of the Alaska State Library has also created a new user on the Linux server, called "searcher", so the searching activity can be distinguished from regular management activities in the log files. Library staff can use Window's native search function to search the repository by file name or containing text without having direct access to the PEP/CEP server. The repository can also be browsed. If staff identifies an HTML or text file that fits their criteria, they can look at it unaided. If they find a binary (Word, PDF, etc) file that looks promising, they need to request a "CVS checkout" of the entire spider's Web site so the specific file(s) they are seeking can be emailed to them. The date of the CVS repository to check is determined by the date and version information within the file. This allows the PEP/CEP server to be accessible from the network within the Library, but not in the wider Internet.

So far, the experience has been that searching the entire repository for a file name takes five minutes or less. Searching the entire CVS archive for a "containing text" match does not seem to be feasible. Searching a single spider's archive for "containing text" does work and seems to take between three and ten minutes, depending on the size of the archive. The staff users only have READ-ONLY access to the files. There hasn't been much call for this access--only twice has someone used the function.

Daniel would like to link the CVS archive to Google Desktop, which would cut the search time. However, state IT regulations currently prevent that change. There is an understandable concern that, if Google Desktop reports its index back to Google in some way, deployment over state networks would risk leakage of confidential data.

Arizona's Richard Pearce-Moses reports plans to catalog each agency's collection in their OPAC. They would very much like to see the html metadata records for the documents in the archives kept online for searching. However, they are open to some

other means to let the public know what has been captured. Arizona wants a better way to access to the documents in the PEP archives.

Illinois has taken all harvested files out of compression in an effort to set up a user-friendly search function. The amount of data stored makes this effort difficult. The data is nearing the six-terabyte server capacity and the open-source SWISH-E search software was not designed to handle such large volume. Therefore, Jackson is experimenting with dividing the data so it can be searched by different instances of the software and then applying a federated search. Illinois uses the most recent PEP harvest to populate its state government search engine, <http://findit.lis.uiuc.edu/cgi-bin/search.cgi>, but this engine does not search earlier harvests at this time.

Montana has decided that the resources to sustain the PEP system are not available at this time. Therefore, access tools are not currently needed.

Utah library and archives requested the hardware requirements for the PEP system with the intention of sustaining the program after the grant period. However, Utah has been undergoing a statewide reorganization and consolidation of Information Technology. A House Bill effective July 1, 2005 has created a new Department of Technology Services (DTS). Services that were once controlled at the library have been forfeited to DTS, so much is unclear at this point.

Before the change, the library bought a new search engine (or components) from Autonomy. Ray Matthews was managing the crawling, fetching, and interface development of it until mid-May. Since May, the project has been managed by the ITS division (now DTS), and it just went live in early October (<http://search.utah.gov>).

The original plan was that, at some point, a server with the CEP files would be crawled to create an index to it. Theoretically, providing access via the search engine would be an easier solution than programmatically creating a white box user interface for accessing the data. At the beginning, Ray conceptually created four Autonomy Idol Servers because each idol server had a size limit of 500,000 documents. The state domains were on #1; city and county domains on #2; intranets were on #3; and #4 was to be reserved for the web archives. When Ray left the project, only two had been configured. DTS could still configure another idol server, a fetch and database called Web Archives, and then crawl and index the data once the safety net files are on Utah's own server. The problem is that to give such a search any utility, a parametric search (in Autonomy lingo) by domain or agency, and date would have to be created. If this type of metadata is available, such a parametric search may be possible.

Wisconsin's goal for CEP participation has been to use what has been collected to cover any gaps in the materials they've sent to the OCLC Digital Archives. They are forming their plan to reach that goal.

Upgrades

PEP software has been installed and upgraded with bug fixes, statistical report functions and a facility to identify previously unseen URL links in Arizona, Illinois and North Carolina. It has been sent to all participating states.

The reporting package, in addition to reporting statistics concerning the latest harvest, retains all prior versions of the report in files with date-times in their names. Website managerial information on all the spidered websites of the state's group is included. The numbered links for each website/agency drills down to the details, re: each spidering of that agency/website. The report lists (1) all the host computers mentioned in any spider configuration, and (2) a printout based on the file configured to command the harvest spiders to ignore specified host servers. For each host each spider encounters, it subtracts off known hosts, then subtracts off "IgnoreTheseHosts" hosts, then reports a list of the remainder -- hosts that may have escaped the manager's awareness. These hosts can then be investigated, and either added to the IgnoreTheseHosts.txt, or have new spiders created for them. They will then appear on the next month's report.

Under each numbered website listed, one can view plots of the website sizes over time in bytes and in number of files, as observed by CEP. Further, it provides a numbered list for each spidering of that website. If Java is installed, the treemap visualization can be displayed. Note the file types pie chart and accompanying table of numerical values. Sometimes the automated pie draws the labels of very tiny slices too close together, in which case the manager can take these numbers and into Excel. Treemaps are very good ways to learn of duplicated regions within websites due to imperfect copy/move operations, DNS aliases, etc. The eye will rapidly note large similar chunks in these graphic representations. Drilling down via mouse clicks to an individual file, one more click takes you back to the visualization of the whole website.

Concerning hosts, links encountered, broken links reports, and such, the new reports are all specific to a single spidering of a website, and are only produced by the CEP 1.0 software. The new "Web Links Statistics" section is based on features of wget software. The link Richard requested is labeled "Other website names not previously seen." Another link of interest to the webmaster is the "broken links report" which lists (a) a link to exactly the link target which failed spidering, (b) the HTTP error code the server returned when the spider asked for said link, and (c) a link to the exact page which asked for the broken link to facilitate link repair.

CEP Archives Operations Guide CEP Software Installation Guide, the new Archives Operations Guide, contains material on adding a disk to the CEP host, and creating symbolic links to redistribute the data storage. CEP can produce "alerting service" reports. Also "SmartCleanup.pl" can erase the bulk spider data, leaving only the log files in each spidering directory. This can save much disk use, unless avoided because of concerns about retrieval from CVS.

Alaska will receive the upgrade when Jackson can arrange a flight to Juno.

Arizona has found advantages to the recent software upgrade.

- It has given the ability to spot domains that haven't previously been identified, which is invaluable.
- The ability to spot processes is a basic Unix command, but it's nice that it's been packaged. The way Jackson put it together improved workflow.
- Of course, statistics are always useful, and the new statistics function will be most helpful.

Montana found that the state agency webmasters with whom this information had been shared earlier this year were interested in the data.

North Carolina didn't have the PEP software before upgrades because their system was run from GSLIS, so they can't make comparisons. Now that they have now run five harvest spiders on their own in North Carolina, they are happy with webmaster reports, even though the graphics are not working yet. .

Utah has also been pleased with the statistical reports that have been created since their software upgrade.

Wisconsin has not used the upgrades.

Electronic Documents Initiative

The Illinois Electronic Documents Initiative (EDI) created a state depository, which provides permanent public access to publications of the State of Illinois in electronic form. The depository was designed to assign unchanging URLs for access from MARC records.

The Electronic Documents Initiative (EDI) depository design allows ingest of electronic documents either by upload or spider harvest, in the case of a single-file document. The spider-based document acquisition mechanism was hard-coded to only accept single-file materials from the Internet. Single-file materials, plus invisible stuff like their bigrams and trigrams that are used for duplicates checks, are named/stored one way while the multi-file materials are named/stored another way. Jackson is in the process of making those changes needed to allow for the reality that all documents are potentially multi-file documents, eventually.

The original design function was, when a document underwent a format change or other correction, it would result in (1) single-file documents being recast as multiple file documents, and then (2) an alternative link would be added, labeled as to explain (e.g., "The original application that created this document requires an operating system no longer available..."). However, to do that would "break" the first document's persistent URL. Instead, the code is being changed so that all documents are treated as multi-file documents—even if the multiple is one. This will let EDI deal with format obsolescence without invalidating any URLs.

If a document is identified as rights-restricted, it will not be reachable online via EDI, but its Blue Card, its metadata, displays. The Blue Card notifies the user to "see your library", instead of providing a hyperlink to the document. The system will automatically review the metadata of deposited documents quarterly and report the URLs of documents identified as having expired rights restrictions. The provided URLs allow the person who receives this report and holds the necessary password to click directly in and edit the rights-related metadata to activate hyperlinks to newly unrestricted documents.

EDI is equipped to automatically send an email alert to the system manager if any document is changed in any way that would alter its checksum. Therefore, any tampering with files would be discovered almost immediately, and backups could replace damaged or corrupted files.

Lessons learned

Another server died during the final spidering and DVD preparation at UIUC for North Carolina. Backup copies were moved onto the Illinois CEP server for the preparation of deliverable materials--DVDs, log files and such. Jackson installed the CEP software and their data in late September 2005. The hardware problems of last year had prompted additional backup provisions, which prevented data loss from this server failure.

The five copies of Illinois CEP archives all resided in Champaign. They were in four different rooms, spanning three different buildings, but some disasters are big. Since the havoc caused by hurricanes in the Gulf, these arrangements have been reconsidered. Therefore, rewrite-able DVDs will be used to send a backup copy to Springfield each month. In addition, the "write once" type (DVD-R) will be used to annually send a copy of the whole inventory, including software for retention, to the State Library.

Deliverables

At the conclusion of the second year of the grant and the end of the project phase involving the participating institutions in other state, each participant received a set of DVDs. The files on the DVDs included that state's captured electronic publications in CVS compression, the associated metadata and a copy of the PEP/CEP software with current upgrades. A more detailed explanation of content sent to the participating states is attached to the end of this report.

Sustainability

As Illinois develops further enhancements and bug fixes for the PEP software, packets will be prepared for sending to CEP participating states and for download from the Web site.

Alaska

At the current pace, the current drive will be full next fall, even though the material is compressed with CVS. Cornwell recognizes the need to begin planning to meet the requirement for additional disk space.

Arizona

Pearce-Moses sees a need for a 'best practices' meeting. Kristin Martin mentioned possibly holding a meeting in North Carolina the end of March 2006. Pearce-Moses suggests that the CEP states' representatives could use that forum to discuss things like how to configure spiders for problem pages or how often to start a new archive. Arizona continues to feel a need to know a lot more about Unix system administration and the details of CEP.

However, that doesn't mean Pearce-Moses is not a PEP/CEP fan. At a recent ARL forum on government documents he spoke some about the CEP and OCLC/UIUC projects, and found a lot of interest. So much interest that there may be another round of institutions willing to buy machines and use resources to learn how to implement and configure them.

CEP is now a program in Arizona, not a project, and Pearce-Moses is encouraging other states to adopt it.

Illinois also considers the PEP system of capturing state e-published documents to be a program rather than a temporary project. Along with EDI, it is now established in the ISL budget. Ever expanding disk space needs and server/network administration costs need to be addressed annually.

Montana reports learning a great deal by their participation in the CEP grant project concerning issues of capturing and reconstructing websites. Previously, education on these topics came mainly from email, listserv discussions, and in-person conversations at national library meetings. Due to not having enough staff dedicated to the technical aspects of implementing the PEP system and unresolved collection development policy governing websites as publications, the staff has for the moment decided to concentrate on a strict definition of electronic state publication. Hence, they have adopted the OCLC Digital Archive product for permanent preservation of state publications that have a print equivalent. Still they remain very interested in revisiting the idea of harvesting entire websites with the PEP system at a future date.

Funding will continue to be the main challenge for the Montana State Library to carry out their mandate of providing access and preservation of state publications. Handling emerging file formats, developing publishing standards, and eliciting greater state agency cooperation in our depository program are perennial challenges.

North Carolina State Library is presently talking with the North Carolina Archives to discuss future of electronic archiving in that state. The Archives is involved in a pilot project using Archivelt in ARC format from Internet Archive. The project files are on the Way Back Machine and sent to the state. The institutions are comparing CEP costs with Archivelt. ARC format rights are privately owned, but the code has been published, which eases some concerns about proprietary format. North Carolina has 228 GB in CVS compression, or about 18% of current capacity. A collection policy is being considered to maximize the value of available capacity.

Utah is in a transitional period with the creation of a new Department of Technology Services (See Access Tools above.) Whereas Elizabeth in the State Archives, Department of Administrative Services indicated they have the freedom to program and develop components, the staff of the State Library, Department of Community and Culture has been specifically told they cannot program or manage servers. The personnel and services available to the library for projects like CEP are still in question. The new rate structures to buy back these services from DTS will not be in place until May or July 2006, so definite plans are not possible at this point.

Wisconsin has never viewed hosting a PEP system as a practical possibility. They simply don't have the personnel or other resources necessary. However, Sally Drew reported that working with GSLIS has been great. Not only did Jackson and his graduate assistants operate the system for Wisconsin, but they also helped identify Web sites previously unknown to the State Library. Sally hopes to get past Web-published documents from the data gathered by PEP software to deposit or use the PEP software as a harvester with the OCLC Digital Archives as storage.

Challenges

Bureaucracy is a wonderful system for maintenance of government programs because it is difficult to turn or stop a huge ship of state. However, the ship of state also has

momentum that impedes establishment of new initiatives. The challenges of sustainable funding; hiring and retaining skilled staff; and collaboration with other state agencies all can be attributed to a lack of understanding the value of preserving electronic state publications. Only highly valued programs can be assured survival in particularly difficult fiscal times. The preservation of electronic publications is particularly vulnerable because even a temporary interruption in the activities creates an unrecoverable void in the record.

The risks are high because some inherent challenges to electronic or digital preservation will always exist and can never be fully resolved. Obsolescence of formats and operating systems, as well as ever-emerging creation software will continue to be threats to preservation and issues to address for the foreseeable future. Preservationists can only retain digital objects for blocks of time while formulating plans for retention during the block of time to follow. A permanent solution for preserving access to content will likely never be realized. Therefore, continuous support by governing bodies will be required.

In the final year of the extended grant period ahead, ISL will concentrate on outreach. The Outreach Coordinator will be contacting agencies and training publications staff to deposit electronic copies of state publications in the EDI depository. There are no illusions that these contacts will be universally welcomed. The Illinois Document Depository Program has worked constantly since 1967 to collect print publications from state agencies with limited success. There is no reason to believe that digital deposits will be easier to inspire. However, we have a strategy to include agency heads and publication administrators who have not historically been involved. An important component of this strategy is an emphasis on the advantages of electronic publications deposit for the creating agency. Additionally, ISL Technical Services staff members plan to extract the most important electronic publications from the PEP/CEP archives and deposit them in EDI. By implementation of this plan, ISL hopes to lay the foundation for a successful program.

Recent initiatives by the Library of Congress and the Institute for Museum and Library Services to support digital preservation are positive signs that state libraries and archives are not on their own to create and maintain such programs. Beyond seed money to begin digital preservation initiatives, these federal institutions may be able to raise awareness among state and federal authorities as to the importance of supporting such programs and activities over the long-term.

ATTACHMENT: Documentation shipped with final DVDs to participating states' institutions

CEP grant conclusion - DVD shipment contents
4 October 2005

With the conclusion of the Capturing Electronic Publications IMLS National Leadership Grant, a set of DVDs containing the captured data and metadata for your state websites is being shipped to you. This document describes the contents of those DVDs.

The DVDs are divided into two groups:

(1) The larger group ("the CVS DVDs") is labeled with the State abbreviation, the letters "CVS", and a serial number. They contain the bulk data, archived in CVS form, but then possibly subdivided if the CVS directory was too large to fit on a single DVD volume. The CVS DVDs are intended to be read on a UNIX computer, and may not retrieve all files if read on a Microsoft Windows or Macintosh computer.

(2) The smaller group ("the software DVDs") is labeled with the State abbreviation, the words "CEP software", the CEP software version number ("1.0") and a serial number. The software DVDs can be read on UNIX, Microsoft Windows, or Macintosh computers. They contain the CEP software (current and previous versions), documentation for CEP (current and previous versions) and for other technology components (e.g., HTML, style sheets) which are used within CEP, certain CEP statistical reports, and copies of the CEP log files which have proven most useful.

Note that DVDs are often said to be a media which will be stable for many years. However, we have encountered some DVD failures after only two years. Accordingly, you are advised to consider these DVDs as means of transference of the files from the UIUC site to your own. Please install these materials promptly. If an installation fails, it may be possible to obtain a replacement for the affected files from copies held at UIUC. UIUC will keep a copy of your materials until such time as the retention facilities become needed for some other project.

Note that date-time values are encoded within CEP in "YYYYMMDDhhmm" format (year, month, date, hour, minute), or rarely "YYYYMMDDhhmmss" format (adding seconds on the end). Leading zeroes are always used, on a per-field basis. Hours are using a 24-hour clock. For example, September 30, 2005 at 1:23 PM is encoded 200509301323. Under this encoding, alphabetical order and chronological order are the same.

===== THE CVS DVDs:

The variable, and often very large sizes of archived websites produces a situation where; (1) multi-volume DVD backups are necessary, (2) some single websites will be backed up across multiple DVD volumes, and (3) it will be desirable to pack DVD volumes somewhat efficiently to reduce the cost and quantity of backup media. Accordingly, the table of contents and the volume label of each DVD will be critical to supporting retrieval of the contents of the CEP archive from the persistent media (DVDs). Each CVS DVD will include a copy of its own table of contents file, "DirectorySizes.txt", that details its contents in terms of the CEP slang-names of the websites included.

The CVS DVDs contain full or partial copies of the UNIX CVS depository areas used to store all the archived web materials. Although not a compression format, the net effect of using CVS to store web files is a very significant reduction in the amount of disk space consumed. Unfortunately, the CVS copy is not browseable via web browsers or file system browsers (e.g., Microsoft Windows Explorer). However, such materials can be retrieved from the CVS copy, as described in the CEP Archives Operations Guide. If you wish to do that, ensure you have a project UNIX computer with sufficient available disk space.

It will be necessary for users of the persistent media to view these volumes as conceptually sub-parts of a single, possibly very large, hierarchical structure. As such, it may be necessary for users to re-assemble the portion of the hierarchical structure they are interested in on a workstation's disk space before they are able to examine the materials. For example, if an agency's archive is distributed over two DVD volumes, the user must copy (UNIX "cp -r") the materials of interest from both the volumes, and must put them together on a sufficiently large scratch disk. At that point, a hierarachy duplicating that originally exhibited by the CVS archive of the agency's website on the CEP computer will exist. That done, CVS commands for the extraction of any individual spider download of any individual agency can be issued.

For example, using a portion of the table of contents listing for the latest production of DVDs for the Illinois CVS data:

```
4396523520 /data1/tmp/Iso-0052/DoTran/DoTran
4396523520 total
4379848704 /data1/tmp/Iso-0053/DoTran/DoTran
4379848704 total
57344 /data1/tmp/Iso-0054/Audio/Audio
102400 /data1/tmp/Iso-0054/Audio/CVSROOT
590196736 /data1/tmp/Iso-0054/DoTran/DoTran
458752 /data1/tmp/Iso-0054/HouseGOP3/CVSROOT
54587392 /data1/tmp/Iso-0054/HouseGOP3/HouseGOP3
22892544 /data1/tmp/Iso-0054/NatHistSurv3/CVSROOT
3670364160 /data1/tmp/Iso-0054/NatHistSurv3/NatHistSurv3
4338659328 total
```

The first field is the size of the item on the DVD, in bytes. Totals are printed for each DVD volume. Note that the CVS archive of "DoTran" (Department of Transportation) spans volumes Iso-0052 through Iso-0054. Note that there are also other materials on volume Iso-0054.

A typical UNIX command sequence for the recovery of these files might be:

```
mkdir BigPlace
```

```
# Insert disk Iso-0052
mount /dev/cdrom
cp -r /dev/cdrom/DoTran/ BigPlace
umount /dev/cdrom
# Remove disk Iso-0052
```

```
# Insert disk Iso-0053
mount /dev/cdrom
cp -r /dev/cdrom/DoTran/ BigPlace
```

```
umount /dev/cdrom
# Remove disk Iso-0053
```

```
# Insert disk Iso-0054
mount /dev/cdrom
cp -r /dev/cdrom/DoTran/ BigPlace
# At this point, the CVS area for DoTran has been completely restored, and CVS commands
can be supported.
# If the other websites on disk Iso-0054 are also desired to be restored, continue as...
cp -r /dev/cdrom/Audio/ BigPlace
cp -r /dev/cdrom/HouseGOP3/ BigPlace
cp -r /dev/cdrom/NatHistSurv3/ BigPlace
```

```
umount /dev/cdrom
# Remove disk Iso-0054
```

===== THE SOFTWARE DVDs:

"crontab.txt" is the excerpts from the UNIX cron table used at UIUC in connection with the monthly archival work for your State. You may wish to set up similar automated schedules. Refer to the UNIX documents "man cron" and "man crontab" concerning use of this feature.

"httpd.conf" is the web server configuration file used at UIUC in connection with the server computer which hosted all monthly processing for your State. Presumably your new installation will have its own configuration, however checking through this file may be informative.

"IsoDuReport2005*.txt" - if provided, this is a UNIX "du" listing for the CVS repositories used by CEP for all your state websites. It tells you how much disk each CVS repository consumes, and may be useful for facilities planning.

"*CvsDvdToc2005*".txt - this file is the table of contents of all the provided CVS DVDs. Website archives are generally stored on the DVDs beginning with the largest archives on DVD volume 1. Archives selection for subsequent DVD volumes generally continues in decreasing order of size, except that unused space on a DVD will be filled using the archives of small websites. Website names here are the "slang names" used in CEP processing, not the formal name of the website or sponsoring agency.

"SpiderConfigurationFiles" directory - if provided, this directory contains the CEP configuration files for each website harvesting spider operated by UIUC for your State. Presumably your CEP operation will begin with this set, editing it as necessary to adjust configurations, add websites, or remove websites from active harvesting. The CEP spiders are each custom-configured to the website they serve. In the vast majority of cases, all that is customized is the name of the website and its homepage. However, in some cases, additional directives have been developed over the period of CEP processing to do such things as exclude directories or files which were problematic, or to deal with websites which were case insensitive and thereby engendered large amounts of redundant downloaded material. To continue CEP or other web archiving operation in the future, this information will be very valuable. Without it, spider definition efforts would have to start from scratch. File!

s here are named with the CEP "slang name" of the website, followed by the suffix ".xml", with the exception of the file "SpiderDefaults.xml" which provides global defaults for all spiders of your State.

"ApacheWebServerDocumentation" directory - this directory contains a copy of current Apache software foundation (www.apache.org) documentation on the operation of the Apache web server software, and is provided for reference by users of the future.

"DocumentationOnHTML" directory - this directory contains a copy of current World Wide Web consortium (www.w3c.org) documentation on HTML, and is provided for reference by users of the future.

"DocumentationOnStyleSheets" directory - this directory contains a copy of current World Wide Web consortium (www.w3c.org) documentation on HTML, and is provided for reference by users of the future.

"OldCepDocumentation" directory - if provided, this directory contains a copy of the first version of the CEP software. It is provided for reference only, as your State processing was being done using CEP version 1_0 at the conclusion of the CEP grant.

"CepArchivesOperationsGuide_1_0" directory provides the archives operations guide for the current version of the CEP software. After installation and for normal daily use, this document will be the primary reference. Begin reading this document with the file "index.html".

"CepSoftware_1_0_0" directory provides an UNIX tar file containing the current version of the CEP software, to be installed on a UNIX computer.

"CombinedSpiderAndCvsStatistics" directory, if provided, contains statistics files which reflect combination of information from MasterMetadataStatistics files with CvsDepositionLogs materials. These combinations were of use in the formation of the StatewideSpiderStatistics reports, and may not be otherwise very useful.

"CvsDepositionLogs" directory contains the CEP log files from the depositing of spider downloads into CVS repositories.

"MasterMetadataStatistics" directory contains the CEP files which post-spider processing writes, containing all the known, extracted, or inferred metadata on each file in the given download. These files also contain directory listings and tree-format directory displays. With these files, and the SpiderRunLogs files, one can answer most questions concerning whether or not a given file was harvested and archived.

"SpiderRunLogs" directory contains wget spider run logs for each spider which has been run by UIUC for your State. With these files, and the MasterMetadataStatistics files, one can answer most questions concerning whether or not a given file was harvested and archived.

"StatewideSpiderStatistics" directory contains statistical summary tables of all the websites of your State. Files here contain date-time of report generation within their names. Files reflect the definitions of State website spiders as of the time of report generation.