



PRESERVING ELECTRONIC PUBLICATIONS (PEP)

Interim Report: Oct. 1, 2001-March 31, 2002

PROGRAM PURPOSE: *Preserving Electronic Publications* is a collaborative proposal between agencies and states to develop a national model for monitoring changes made to electronically published state government documents and records to ensure permanent public access. The value of these partnerships is two-fold:

- 1.) By reaching definitions of electronic record and electronically published document that are held in common across state lines, the PEP partners could pave the way for inter-state interoperable searching.
- 2.) By comparing and contrasting methods and assumptions, the PEP partners could define the most cost-effective methods for the highest quality results.

INTENDED OUTCOME: The project out-come will be a model for other states as they design programs for permanent public access to government records and publications on the Web.

Partners Involved:

The Illinois State Library, The State Library of Ohio, Illinois State Archives, and Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign.

Program Influencers:

IMLS; the public; PEP partners; state legislatures and executive officers; other state libraries and archives, state and federal GILS.

Indicators:

Efficiencies in change monitoring methodologies being utilized in Illinois and Ohio; improved quality of metadata and the improved webmaster training in Illinois and Ohio.

Data Sources:

Log file statistics showing number of cached pages and number of lines of metadata; user feedback concerning the system functionality; capture and retention of system usage and load statistics.

1. Arriving at definitions of "electronic record" and "electronically published document"

The PEP partners agreed to adopt the definition of the term "electronic publication" this is being used by the Joint Electronic Records Repository Initiative (JERRI) in Ohio. The definition was formulated by JERRI, a cooperative effort between the State Library of Ohio (SLO), the Ohio Historical Society, the Ohio Department of Administrative Services and the Ohio Supercomputer Center, who are addressing long-term preservation of electronic records. The definition of "electronic publication" being used is the Internet version of any items listed below. This list follows the model of the SLO's current paper format requirements for acquiring State of Ohio publications created for the general public. (The list is not exclusive):

- Annual Reports
- Bulletins/Circulars
- Databases (a.k.a. directories)
- Forms
- Handbooks/Manuals
- Laws & Rules
- Maps/Charts
- Multi-media files (sound, video, interactive)
- Newsletters
- Pamphlets/Brochures/Fact Sheets (a.k.a. General Publications)
- Press Releases
- Technical/Research/ Statistical Reports

The PEP partners discovered that the distinction between a record and a publication becomes blurred in the electronic environment. Traditionally, libraries have preserved publications, information that is produced for dissemination, and archives have retained records, which are primary source government materials for which a commission or board has determined a value and a retention period. However, now records are posted on the Internet for public access, which constitutes electronic publication. Therefore, the partners agreed that the overall PEP initiative should focus on preserving electronic publications, whether or not they are records.

In a related matter, in March 22 legislation was introduced in the Illinois General Assembly to add the definition of "published material" into the State Library Act [15 ILCS 320]. The legislation proposed defines "published material" as publications in print and electronic formats duplicated by any means, including material downloaded from a publicly accessible electronic network. This legislation, which has been approved by the state legislature and will be sent to the Governor to sign into law, will impact permanent public access to state Web-based publication.

2. Creating a project Web page and project promotion

The Illinois State Library is developing a Web site for the project. Projected components of the Web site include the following:

- Space for downloading files, starting with files of rules for automated subject classification from Access Innovations under the Washington State Library's 2000 IMLS National Leadership Grant;

- Reports from Larry Jackson of U of I GSLIS;
- Final report from Kelly Meiers state GILS subject tree building;
- Reports to IMLS;
- IMLS logo with a link to <http://www.ims.gov>
- Links to the Web sites as follows:

Find-It! Illinois <http://finditillinois.org/>

Find-It! Illinois grant site http://www.cyberdriveillinois.com/library/isl/lat_ims_grant.html

Find-It! Illinois What's New <http://finditillinois.org/new.html>

LAT Division Web page <http://www.cyberdriveillinois.com/library/isl/division.html#LibraryAutomationandTechnology>

Larry Jackson of the University of Illinois at Urbana-Champaign has been posting regular reports to a password protected Web site. During March, Mr. Jackson attended a presentation by Joe Futrelle of NCSA on his [COCOA and Open Archives In-a-Box](#) toolkits for building Open Archives Initiative-compliant metadata servers and harvesters. He also attended the IMLS [WebWise2002 conference](#) in Baltimore on March 20-22, and will make a presentation on the PEP initiative and system architecture at the GILS conference in Arizona on April 24-26.

3. Change monitoring at the State Library of Ohio

SLO will document experiences in a methodology that applies the selection and cataloging skills of librarians to the challenge of identification and preparation of electronic publications for permanent public access.

- SLO staff working on this project Jim Buchman, Director of Public Services; Diane Fink, Head of Fiscal Services, and the SLO PEP Team; Kathy Hughes, cataloger; Nicole Merriman, GIS librarian; Gretchen Persohn, Head of Research Services; and Jeff Heard, Head of Cataloging.
- During November and early December, the team reviewed position descriptions for metadata catalogers, decided on project deliverables, and developed a request for proposal to hire a consultant for this project.
- The RFP was issued December 10, 2001. Responses to the proposal were due December 31, 2001. The RFP issued included the following:

Purpose: The State Library is soliciting for services of one consultant as a special project manager for the PEP (Preserving Electronic Publications) Project.

Background: For the past eighteen months, the State Library of Ohio (SLO) has been investigating ways to identify, harvest and store, Web publications that are "born digital" by state of Ohio agencies. SLO staff has

partnered closely with staff from other state agencies and state libraries who are also concerned about this issue.

Currently, the SLO is a beta tester of the OCLC Digital Vault project. It is anticipated that this digital vault, once it is put into production, will provide the server space required to house the SLO archive. In addition, the SLO is partnering with the Illinois State Library, which is developing its own digital archive. It is through this latter partnership, which is referred to as the PEP (Preserving Electronic Resources) Project, that the SLO is able to fund a project consultant.

Scope of Work: The PEP Project Consultant will work closely with State Library of Ohio staff. This individual will be responsible for the coordination and successful completion of the State Library of Ohio's PEP Project deliverables, as defined in this RFP. This person will assume responsibility for the planning, development, and implementation of sensible and effective policies and procedures for cataloging electronic resources.

Proposal Requirements:

- 1) Qualifications: Contractors must clearly explain in detail their qualifications and library experiences.
Required: ALA-accredited degree in library or information science or equivalent combination of degrees and experience. Knowledge of current cataloging standards and practices. Knowledge of MARC, and experience with Dublin Core metadata standards. Experience with OCLC and CORC. Experience with cataloging electronic resources. Excellent written and oral communication skills. Experience in the creation and implementation of training materials.
- 2) Federal Tax Identification Number and/or social security number: Submit one completed and signed IRS form W-9 with the proposal.

Assumptions:

1. Communication: The consultant will coordinate all activities through the Head of Public Services for the State Library of Ohio. This coordination will include at least one face-to-face meeting per week.
2. Hours: It is expected that the consultant will work forty (40) hours per week. Consultants are expected to use "down time," (i.e. time not spent developing training materials) to identify and add records for electronic resources to the SLO catalog.
3. Conditions: It is mutually understood and agreed that the consultant is an independent contractor and will receive no fringe benefits, and that there will be no money withheld from compensation for income taxes or retirement for any purpose. The consultant will carry insurance, which will absolve the State Library of Ohio from any responsibility in case of accident.

4. It is understood that the consultant may have to learn specific software applications used at the State Library of Ohio, such as the Innovative Interfaces, Inc. system. Training will be provided by SLO staff; however, it is expected that the consultant will become proficient quickly and be able to work independently.

Deliverables: By the end of the project (September 30, 2002), the State Library expects the following deliverables:

- 1) Development of a complete training package targeted toward Web masters and content creators to provide detailed instructions on adding metadata to born digital Web publications.
- 2) Development of documentation that defines an electronic publication and specific selection criteria used to determine which publications are candidates for storage in a digital archive.
- 3) Implementation of #1 at the State Library of Ohio as it relates to creation of SLO Web publications. (The SLO will be used as the test site for the development of the training package.)
- 4) Creation of a core set of catalog records for inclusion in the digital archive.
- 5) Development of a practical cataloging workflow to be utilized by SLO.
- 6) Assistance with completion of the outcome-based evaluation.

Specific goals will be agreed upon monthly between the consultant and the Head of Public Services. In addition the State Library of Ohio expects:

- 1) Reports: Monthly progress reports that provide detailed information needed to document accomplishments and progress toward meeting deliverables. Specifics addressed in progress reports should be aligned with outcome-based evaluation criteria, as applicable.
- 2) Materials: All curriculum and training materials developed by the consultant will become the property of the State Library of Ohio.
- 3) Data: All raw data and summary data collected during this project are to be delivered to the State Library of Ohio upon conclusion of the project.
- 4) Final Report: The consultant will submit a final report summarizing the work completed and the status of the project. Suggestions for follow-up and future issues to consider relating to born digital publications should be included.

- One proposal was received. The PEP team reviewed the proposal; an interview was conducted with the candidate by the PEP team; and the PEP team recommended hiring the consultant, James Rubottom.

- An Agreement for Services was entered into between the State Library of Ohio and Mr. Rubottom on January 23, 2002. The consultant will be expected to complete project deliverables and goals by the end of the project as articulated in the RFP.
- As per the agreement, weekly meetings will be held between Mr. Rubottom and Mr. Buchman, with alternative weekly meeting between Mr. Rubottom and the PEP team. The first order of business will be the “development of a complete training package targeted toward Web masters and content creators to provide detailed instructions on adding metadata to born digital Web publications.” SLO will look at the work already done by the State Library of Illinois, as well as other states, as we develop our package. Mr. Rubottom began project activities in February 2002, and his initial efforts included drafting an implementation plan for the project for review by SLO staff and developing presentations on explaining the elements of metadata.

4. Reviewing electronic documents/records by the Illinois State Archives; developing a model for an electronic records system

The Illinois State Archives has discovered that the guidelines for records-management in print-based materials do not translate exactly to the electronic format, and that an analysis needs to be made to set up a system founded on best practices. The State Archives has finalized its job description for the electronic records consultant. The RFP for the position is as follows:

Under the provisions of the Illinois State Library IMLS "Preserving Electronic Publications" (PEP) grant, before September 1, 2002 the contractor will perform a general assessment of records management procedures and needs centered on the creation, description, preservation and access of electronic records in the Illinois State Archives and agencies falling under the State Records Act and Local Records Act. Duties in and around Springfield and Chicago will include but not be limited to:

- 1) Review current state laws and administrative rules concerning electronic record-keeping and make recommendations for changes or additions.
- 2) In consultation with Archives' records management staff, review and evaluate in writing ten selected state or local agency record retention schedules already existing for electronic records. Also, identify records created and/or maintained in electronic form from selected Illinois state and local agencies and give guidance in preparing retention schedules for these records, including guidance for appraising records in electronic mail formats.
- 3) Review options and make recommendations for accessioning, data conversion, storage, preservation, and access for electronic records appraised as archival.

- 4) Recommend what would be necessary for the Illinois State Archives to coordinate electronic records data conversion and /or other tasks and procedures necessary to ensure preservation and access to electronic records accessioned into the Illinois State Archives.
- 5) Develop programs or methods to inform state and local agencies of their obligation to maintain electronic records and to identify what would be necessary for the Archives to provide agencies technical assistance regarding electronic records creation, maintenance, preservation, and access issues.
- 6) Assess the current status of PKI digital signature technology in Illinois and apprise the Archives of its responsibilities in working with this technology and preserving these signatures.
- 7) Meet with Information Services managers from five selected local government agencies in Illinois and evaluate their concerns and needs in relationship to existing programs in the Archives' Local Records Unit.
- 8) Meet with members of the State Records Commission, Local Records Commission and the Illinois State Archives Board to learn their concerns about electronic record keeping.

5. Purchasing a server

A multi-processor PC running the Linux operating system and containing up to 1 Terabyte (1,000,000,000,000 bytes) of IDE RAID disk space is planned. This machine will be located on the UIUC campus, and will have access to the very-high-bandwidth Internet connections of the University of Illinois. Equipment and operating system configuration, maintenance, security updates, and network maintenance will be provided by the staff of the Information Systems Research Laboratory of the UIUC Graduate School for Library and Information Science. The equipment will be purchased between months 7 and 12 of the project.

6. Designing a permanent public access depository based on automated harvesting of electronically published state documents.

Elements of the system will be largely share-ware to cut costs and allow for distribution to other states when fully developed. The prototype will initially concentrate on Web sites employing open document formats (specifically, HTML and XML). As funds and time permit, proprietary formats (e.g., Microsoft Word, Adobe PDF) will be incorporated. Red Hat Linux with Apache Web server and CVS for file retention will be used as the host operating system for the archive facilities. Information indexing and search capabilities will be constructed using open source software such as iSearch and MySQL. Scripting software will be developed locally to link the production process (e.g., the spiders, parsers, indexers, archival software, retrieval engines, and Web servers). These scripts will also produce summary statistics concerning the quantity of documents or changes they process.

Mr. Jackson's analysis of State of Illinois Web sites as of January 2002 found the following:

- 5,617,823,744 Bytes of data
- 5,475 directories
- 84,155 documents, including:

| | | | |
|--------|---------|--------|------|
| 23,261 | HTML | 48 | XML |
| 639 | ASP | 9,284 | GIF |
| 168 | CFM | 5,457 | JPEG |
| 684 | S files | 26,366 | PDF |
| | | 1,659 | DOC |
| | | 42 | WPD |

 - 82,652 META tags

Through February, Larry Jackson saved 24,092 Web pages amounting to 1,732,730,880 bytes of information. 9,140 of those pages are HTML files.

7. Implementing a pilot project or prototype system for automated gathering of electronically published state documents

This system will key on a metadata element placed in the HTML code by the state agency that creates the document, and Web crawler facilities based on WGet (or equivalent), which will examine state Web sites daily to detect changes. Continual improvement or customization of the parsers and software tools is expected to be necessary to support a diverse community of Web authors in order to generate documents with metadata. Such technical labor will be supplied by a mix of full-time staff and student labor at GSLIS at UIUC.

Testing the interface and retrieval functionality

The Illinois State Library will enlist librarians across the state and students in interface design classes at the University of Illinois to provide user feedback concerning the system functionality. This process will occur in during the period of April 1-Sept. 31. Data collected will be available at the end of the project.

Capture and Retention of System Usage and Load Statistics

Data will be collected and recorded for study and analysis at the Graduate School of Library and Information Science at the University of Illinois with preliminary results available at the end of the project.

2002 IMLS National Leadership Grant

During the reporting period, the Illinois State Library submitted a National Leadership Grant application, entitle PEP², to IMLS to build on the activities of the current grant with the Washington State Library and the Arizona State Libraries and Archives. The goals of the PEP2 project is to develop a national model of replicable procedures and programs necessary to monitor and archive changes to publications of state government that are posted on the Internet in order to provide the public, libraries, archives and state government permanent access to this information. The objectives are as follows:

1. To create enough embedded metadata to demonstrate the correlation between metadata and precision retrieval;
2. To establish recommendations for best practices for permanent public access to state electronic publications based on the experience of four different states;
3. To establish a model for archives to preserve electronic records and make the tools to recreate the system available on the Internet; and
4. To demonstrate interstate interoperability across electronic archives to facilitate the development of the state GILS group into a “Find-It! USA” consortium

PEP Financial Report: October 1, 2001 - March 31, 2002

| | Amount Budgeted | Amount Requested | Amount Remaining |
|---|------------------------|-------------------------|-------------------------|
| Temporary Staff Hired for Project | | | |
| Alyce Scott | 29,250 | 14,750 | 14,750 |
| Bonnie Mathies | 29,250 | 14,750 | 14,750 |
| SLO, Librarian II | 38,000 | 8,800 | 29,200 |
| Consultant Fee | | | |
| Illinois State Archives | 40,000 | 0 | 40,000 |
| Travel | | | |
| IMLS Activities | 4,000 | 800 | 3,200 |
| Instate Travel | 5,525 | 0 | 5,525 |
| Material, Software & Programming | 100,000 | 0 | 100,000 |
| TOTAL | \$246,025 | \$ 38,850 | \$207,175 |