

Preserving Electronic Publications is a collaborative proposal between agencies and between 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.

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

Activities for one year:

1. Arriving at definitions of "electronic record" and "electronically published document".
Working together will increase the likelihood that the product of this activity will be useful nationwide.
2. Creating a project Web page and project promotion.
Throughout the project, dissemination of descriptions of activities, challenges and findings will be key.
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.
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.
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.
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.
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. Such technical labor will be supplied by a mix of full-time staff and student labor at GSLIS at UIUC.
8. 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. Data collected will be available at the end of the project.
9. 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.