The Web at Risk: A Distributed Approach to Preserving our Nation’s Political and Cultural Heritage

Suzanne Samuel
California Digital Library
DLF Spring Forum 2005
Project Goal:

Develop web archiving tools that will be used by libraries to capture, curate, and preserve collections of web-based government and political information.
Four overlapping paths of activity:

- Content Identification and Selection (CISA)
- Content Ingest, Retention and Transfer (CIRT)
- Content Harvest and Analysis (CHA)
- Partnership Building (PB)
Content Identification, Selection, & Acquisition

- Identification and analysis of web archiving strategies and issues
- Toolkit for curators to assess user needs and plan collections
- Framework to conduct and analyze sample crawls
- Case studies, best practices, and collection development guidelines
- Investigation into extensibility of tools and collection strategies
Main Menu

Pre-Submission

DPR Submission Materials

Validate a single METS file

Validate multiple METS files

Submit Objects

Submit a single object

Submit multiple objects

Search

Find an object

Browse

Inventory Groups

See the objects that have been deposited for a single inventory group.

- Sandbox 1
- Sandbox 2
Content Ingest, Retention, and Transfer

- Modification and enhancement of CDL’s Digital Preservation Repository to provide ingest and retention functions
- Specification and implementation of a data model for Web Archive Digital Objects (WADOs)
- Evaluation and implementation of storage hardware and software
- Development and testing of remote replication strategies
- Integration with software deliverables from other components of the project
Content Harvest and Analysis

- Tools to conduct and support curators' web crawling:
  - Crawler
  - Analyzer
  - Curator User Interface (CUI)
  - Export/Import Handler
Partnership Building

• Develop model agreements (data retention, data exchange, data capture, and codification of partnerships [MOUs])
• Execute agreements for this project
• Evaluate mechanisms to sustain work beyond grant and extend to other libraries
• Assess costs of sustaining a distributed approach to web archiving
More info:

www.cdlib.org/inside/projects/preservation/webatrisk