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

Content Identification, Selection, and Acquisition Path

Focus Group Report:

ALA - Chicago - June 2005

Prepared by:

Kathleen R. Murray
Assessment Analyst, Web-at-Risk Project
University of North Texas
krmurray@unt.edu

February 14, 2006

The following people contributed to this report.

University of North Texas	Arlene Weible & Inga Hsieh
California Digital Library	Tracy Seneca

Contents

1 Introduction

The Web-at-Risk project is one of eight digital preservation projects funded in 2004 by the Library of Congress. The project is a 3-year collaborative effort of the California Digital Library, the University of North Texas, and New York University. The project will develop a Web Archiving Service that enables curators to build, store, and manage collections of web-published materials in distributed repositories located at the three project partner sites. The project will also produce tools and guidelines to assist curators and other information professionals with collection development for web archives.

In support of this effort five focus groups were held in 2005. The purpose of the focus groups was to elicit the needs and issues librarians, curators, and end-users have in relation to web archives. This document summarizes the discussion held on June 26, 2005 at the Metropolitan Library System in Chicago, Illinois. The one and one-half hour discussion was facilitated by the Assessment Analyst for the Web-at-Risk project.

The report includes the following three sections: (a) the methodology used to conduct the focus groups and analyze the data, (b) the detailed results of the analysis organized into phases of the collection development process, and (c) a discussion of the key findings.

2 Methodology

2.1 Framework

Collection development for web archives includes three major phases: selection, curation, and preservation. By breaking down collection development into a series of activities within each phase, the functional view shown in Table 1 emerges. Librarians will recognize the activities as those commonly employed in collection planning. (Appendix A provides a brief explanation of the activities in each phase as they apply to collection development for web archives.)

Table 1. Collection Development Framework for Web Archives

PHASES		
SELECTION ⇨	CURATION ⇨	PRESERVATION
Selection	Description	Preservation
Acquisition	Organization	
	Presentation	
	Maintenance	
	Deselection	

2.2 Participants

Participants in the Chicago focus group were volunteers from two sources: (a) the general membership of the Law and Political Science Section (LPSS) of the Association of College and Research Libraries (ACRL) who subscribe to the LPSS discussion list and (b) member organizations of the Chicago Metropolitan Library System which were identified by library system staff members. The participants came from academic libraries, including medium-sized universities and large research universities, as well as non-profit organizations. (See Appendix B.)

A total of eight people participated in the group discussion. Two participants held library management positions, one was an archivist, and the remaining five were librarians. Two of the eight participants indicated they had some prior experience creating web archives.

2.3 Data Collection

Two note-takers attended the focus group and created a record of the discussion as well as a summary of the key points that emerged. Participants completed the questionnaire (Appendix C) that identified demographic characteristics and captured their thoughts regarding:

- User needs addressed by a web archive
- Critical areas their organization needs to address to successfully implement a web archive
- Hurdles their organization faces in creating a web archive

2.4 Data Analysis

The collection development framework (Appendix A) provided the overall framework for analyzing the focus group discussion. Based on a discussion in May of 2005 with curators involved with the Web-at-Risk project, an initial categorization of concerns and issues within each collection development phase was developed. Ideas and concerns that emerged during the focus group were sorted into these categories and new categories were added as appropriate.

3 Findings

This was a lively discussion with all participants eager to contribute. Some participants were more experienced with archives but all were grappling with preservation needs and organizational issues. Much of the discussion could broadly be classified as policy issues and concerns. Additionally, a good deal of time was spent on web archive material selection, acquisition, and description issues and concerns. Unfortunately, time did not allow for in depth discussion of the issues involved in the other phases of collection development: organization, presentation, deselection, maintenance, and preservation.

3.1 Policy

<u>Collection Policies, Practices, & Plans</u>

- Most collection policies include E-journals but most do not include web sites.
- It is problematic to extend existing collection policies to materials that are not owned by the library.
- Web-published materials are a ‘new’ class of materials for collections.
- Collaboration among libraries is needed to build web archives.
  - A model for this is needed.
  - Cataloging is a way to share information about holdings among organizations.
- Frequency of acquisition, (i.e., how often to reacquire and preserve) needs to be addressed in collection plans and the implications of decisions in this regard should be understood and specified.
  - For example, one participant relayed a story of a student trying to locate an item in a newspaper that they had previously referenced. Reference librarians looked in every edition in the newspaper archive and could not locate the item. It was eventually found in the afternoon addition of the newspaper. The archive only preserved the morning editions.
- The scope of the collection needs to be specified in terms of material types and web site functionality.

    - For example: How will forms, data-driven web sites, and RSS feeds be addressed in the collection?
- Contracts between archives and depositors could specify the metadata that would be supplied with deposited materials for the archive.
- Since authenticity of web-born documents is always questionable, policies should specify how authenticity is addressed and what the known limits are. Possibilities identified by participants included:
    - A policy that materials cannot be removed from the archive, with a complementary policy that updated or corrected versions of the materials could be added but could not replace existing versions.
    - A policy that hard copies of web materials would be created and archived in accord with traditional preservation practices.

Organizational Support

- IT organizations need to approach web archiving in collaboration with librarians and archivists. A helpful attitude would be one of creating 'new options' for a 'new challenge'.
- A coordinated preservation effort is needed between the IT department, the library or archive department, and the overall organization (e.g., the university or non-profit organization).

Institutional Repository

- Participants from large research libraries view web resources as one class of materials for institutional repositories.
- Organizations in general need to address the preservation of their own 'born digital' materials.
- Collections of organizational publications and web sites need policies that address
    - Metadata creation
    - Material or resource change evaluation and identification to establish automatic triggers for reacquisition

Financial Challenges

- Coordination of effort across the organization is driven by the resource-intensive nature of preservation activities. There was a general consensus among participants that libraries cannot 'go it alone'.
- Funding for preservation is required for:
    - Infrastructure
    - Staff

Technical Challenges

- Multiple versions of materials may require multiple hardware platforms, software platforms, or applications. Policies need to address which versions, platforms, and applications the organization or institution intends to support.

Roles & Responsibilities

- Preservation activities generally involve IT organizations yet IT practices do not always satisfy preservation requirements.
    - As on participant explained: "IT needs to understand that archiving is not the same as a backup and that preservation goes beyond the three months that backup copies are retained."
    - Another participant amplified: "A systems mentality of backing up data is not the same as preserving it."
- Librarians and archivists have expertise in preservation and curation. IT does not have this expertise.

- One participant related the experience of IT updating all the organizational web pages with current logos and dates, without regard to preserving the originally published formats.

3.2 Selection

Identification of Source Materials

- The key questions are "Should WE save this?" and "Is SOMEONE ELSE already saving it?" Participants struggle with these questions and seldom fine easy and ready answers.
- Participants see a need for a nationally coordinated effort that would include a directory of archived materials.
- Materials on the web that might be 'lost' or 'disappear' are candidates for selection.
  - For example web resources listed in subject guides.
- Historical and legal researchers need all materials referenced in their research and generally all versions of source materials to conduct their research. Participants serving this clientele need to know what organization(s) is already collecting the materials their researchers need.
- Archives are seen as an important mechanism to support the legitimacy of scholarly research by enabling replicability and ensuring validity.
  - Example: Researcher uses a database on a website and publishes a paper based on this data. The database web site is subsequently updated to reflect a correction in the original database or perhaps to reflect new data (i.e., new data replacing older data). The only way the researcher's results can be replicated is with access to the database version originally used in the published research.

Lost Materials

- Currently, participants successfully use the Internet Archive (http://www.archive.org) to retrieve some lost resources.
- Both universities, in their experience with NGO publications, and non-profit organizations identified similar experiences with:
  - Web-based materials disappearing and
  - A lack of organizational commitment to preservation.

Privacy

- Students frequently record class sessions for their private use. The privacy of other students as well as instructors or others in attendance needs to be addressed.
- Recordings can be made in many settings. Should any recording be considered 'public"? What permissions or releases are needed from the individuals on the recordings? What is included under the umbrella of 'public'?
- In regard to comments captured in various media at public forums by politicians, one participant adopted a stance of "if it's published, it's fair game".

3.3 Acquisition

Authenticity of Materials

- Digital copies are questionably authentic. This results in authenticity often, perhaps always, being in question.
- It's hard to 'prove' who is the original author or creator of web materials. It's also easy to modify most web-published materials.
- Quality control for the information in databases is often lacking.

    - One example from a non-profit organization was regarding the official membership database. The database contains errors and there is no quality review of the data. Printed membership lists from the past were of a higher quality.
- One person stated their archive policy did not allow items to be removed from the archive. This gave some measure of reliability that materials in the archive were not modified subsequent to taking up residence there.

Frequency

- Most participants thought that change in web source materials would need to be evaluated by the web harvesting system versus curator or human evaluation.
    - For example, when archiving institutional or organizational web sites, it is highly unlikely that employees (e.g., web managers or faculty) will send a change notice to the archive manager or archivist.
- Changes in web resources can trigger re-acquisition. However, not all changes for all materials need to be captured.
    - For example, the organization's main web site could be captured at certain intervals but each instantiation of a policy document might require reacquisition.
- How do you deal with RSS feeds?

Source Material Versions & Formats

- Web content frequently changes. This is characteristic of web-based resources identified for or by researchers, of organizational web sites, and of the nature of some types of resources, for example, RSS feeds.
- Mistakes in source materials often generate corrected copies of the materials. This is true for databases, public records, and newspapers.

3.4 Description

Level of Description

- Some participants thought that descriptions were only needed at the 'collection' level, similar to the collection descriptions in finding aids. Armed with this description, end users could navigate the resources themselves.

Original Cataloging

- Some participants thought there is of necessity (due to resource constraints in the library or archive) a dependency on the provider/creator/owner of web resources for descriptive information of the resources.
    - For example, newspaper publishers might supply metadata with their resources, in accord with a 'non-mediated' cataloging model.
    - This could be addressed as a submission requirement in contracts.
- Other participants asserted that 'end users' as information providers do not supply or create metadata. Neither do many publishers of web-born materials.
- Entries in the '776' field of the catalog record are used in at least one library to indicate if an electronic counterpart exists for a print resource.

Breadth of Cataloging

- Time stamps reflecting when web resources are harvested are needed for each archived version of a web resource.
- How do you timestamp RSS feeds?
- Cataloging is very problematic:
    - There would be a great deal of repetition across multiple versions of a single resource
    - There is no way to create one "record" of a resource and know it is stable

- How do you catalog RSS feeds?

3.5 Organization

User Expectations

- One participant thought users primarily need to browse not search the archive

Subject or Departmental Lists

- All academic libraries have selectors assigned by subject area. These selectors create resource lists that include web resources.

3.6 Presentation

Dark Archives

- One participant reported that at present their organization is creating a dark archive out of expediency and necessity. This is being done so that critical information in support of research (e.g., a database) is not lost.

3.7 Preservation

Stewardship

- One participant stated: “If we don’t own it, we don’t save it.” While this was a statement in compliance with copyright requirements, the participant also thought long-term responsibility for storage rested with material owners/producers.
- Preservation activities generally involve IT organizations yet IT practices do not always satisfy preservation requirements.
  - As on participant explained: “IT needs to understand that archiving is not the same as a backup and that preservation goes beyond the three months that backup copies are retained.”
  - Another participant amplified: “A systems mentality of backing up data is not the same as preserving it.”

4 Discussion

4.1 Dealing with Change

Building & Preserving Collections

A participant from a law library recounted that in the past resources for their collection were primarily received via a deposit model, wherein publishers provided resources and notified the library of changes to the materials in their collection. In an analogous manner, archives traditionally received materials or collections from donors. In short, libraries and archives were generally on the receiving end of materials and resources for their collections. Their job was to organize these materials and collections for access and to preserve them as needed for posterity.

Transferring the functions of organization and preservation to web collections often shifts the responsibility for identifying the source materials for collections from external sources (e.g., publishers and collectors) to internal sources (i.e., librarians and curators). This added responsibility often involves discovering web-published materials for their collections as well as tracking updates and changes to the materials.

Roles & Responsibilities

In addition to the added responsibility for discovering source materials for their collections, librarians and curators are confronting the issues surrounding their responsibilities for the organization and preservation of web materials, and their updates, in their collections. Fulfilling these responsibilities often involves working more closely with Information Technology departments within their organizations and institutions.

While librarians and archivists have expertise in preservation and curation, IT personnel generally do not. At times, a clash in cultures ensues as librarians bring to bear their experience in information organization and their expectations for material preservation on an IT organization that may not understand or value either. Likewise, IT practices (e.g., backups and updates) are often either unsatisfactory for or in conflict with preservation requirements. Surmounting these differences to implement successful web archives is a major challenge for many organizations.

4.2 What to Preserve

Identifying what to preserve was a major issue for participants. In addition to what to preserve now, future issues were identified regarding what material formats and what elements from a website should be preserved. For example regarding web sites, is it critical to preserve only the "information content'', or is it essential to preserve the website's layout, interactivity, and functionality?

Participants raised these important questions and understood that answers are not clear cut at this point in time. Participants did target the following candidates for preservation.

- Federal government agency information
  - One participant mentioned that much information disappears at changes in administrations
- Publications citied in research
  - Publications of public policy groups
  - Non-government organizations (NGO's)
  - Political parties
- Campus research center publications
  - Newsletters
  - Working papers
- Organizational publications & resources
  - University web sites
    - Main web site
    - Other web sties (organizations, departments, etc.)
  - Organizational web sites
  - Organizational membership lists

4.3 Needs & Issues

At the end of the focus group discussion, participants completed the brief questionnaire in Appendix C. The questionnaire elicited information regarding the critical user needs that an archive of web materials would meet in each participant's environment. Additionally, the questionnaire allowed participants to record the critical areas their organization needed to address and the biggest hurdles they faced in building an archive of web-based materials. In general participants' written responses echoed and provided a summary of the discussion itself. The responses are summarized below.

User Needs

In terms of what user needs participants thought archives of web collections would address, the following three were identified:

1. Access to materials for research and reference
    a. Historical records
    b. Born-digital materials from non-traditional publishers
    c. Scholarly materials from the institution's researchers and research centers

2. Access to an institutional or organizational repository
    a. For preservation of the historical record of the organization
    b. This is especially needed for born-digital materials, which often disappear from organizational web sites.

3. Provision of value-added services, specifically:
    a. Context for the resources (e.g., author, publisher, creation date, etc.)
    b. Organization (e.g., subject lists)
    c. Aggregation from multiple and disparate sources (e.g., newspapers from around the world)

Critical Areas to Address

Participants were asked to identify two critical areas their organizations needed to address in order to successfully implement a web archive. The areas are listed below in order of criticality.

1. Organizational support (Management, faculty, IT organization)

2. Technology (Infrastructure and upgrades)

3. Policies related to web materials

4. Copyright issues

5. Consortial efforts

6. Resources

Biggest Hurdles

Participants identified several organizational hurdles they need to overcome prior to creating web archives. While no one hurdle clearly stood out as the greatest across the participants, two were identified a more often than others: organizational commitment and staff resources. Both of these hurdles were identified in the context of institutional repositories, which was an area addressed by several of the participants in the group.

1. Organizational understanding of the need for preserving born digital materials and organizational commitment of resources to preservation efforts

2. Staff resources to educate faculty and other contributors regarding the value of an institutional repository and staff resources to assist faculty in performing preservation activities

4.4 Need for Collaboration

Participants generally agreed that medium and small libraries are unable to allocate resources to preserving collections of web-based materials. The general sentiment was that there is a need for collaboration and sharing of this resource-intensive preservation task.

Larger libraries may be in a better position to assist in preservation activities but are themselves resource-constrained and facing competing priorities for these resources. This environment obviates the need expressed by participants to eliminate duplication of effort in the preservation of web resources.

Simply put, if the materials needed by users are in archives at other institutions or organizations, then access to those materials is preferable to expending the staff effort and financial resources to preserve them locally. Participants foresee a requirement to leverage the culture of consortial relationships that characterizes the practice of librarianship into the preservation of at-risk web born materials.

Appendix A. Collection Development for Web Archives

POLICY SETTING	Policy factors influencing web archiving include political mandates, organizational mission, financial parameters, and technical capabilities.	
	SELECTION	
	Selection	Choice of web-published materials for archiving is impacted by the focus of the collection, unit of selection, web boundaries, copyright obligations, and authenticity of materials.
	Acquisition	Web-published materials are acquired or 'harvested' using crawling tools, which either globally or selectively capture web-published materials.
	CURATION	
	Description	Baseline metadata is machine-generated and gathered by a crawler at the time of data capture. Enriched metadata is generally specific to an organization and contains a mixture of human-generated metadata added subsequent to data capture as well as machine-generated metadata.
	Organization	Digital archives of web-published materials typically either retain the organizational structure of the materials as they existed on the web at the time of capture or modify the organizational structure to suit the archive's mission or constraints.
	Presentation	Presentation of web archive materials is related to how the content was captured and to post-harvest descriptive and organizational analysis. For example, archived materials might mirror the web at the time of their capture or might be categorized in accord with selection criteria, such as image files presented by subject.
	Maintenance	Several maintenance functions are critical to ensuring the successful use of materials in web archives: software and hardware training for archive support staff; hardware and software maintenance, performance optimization, backups, and upgrades; and duplicate detection.
	Deselection	Removal of materials from a web archive can be for several reasons: duplication, errors, legal or social considerations (e.g., offensive materials). Risks of removal and retention are weighed against policy and storage costs.
	PRESERVATION	
	Preservation	Preservation challenges are numerous. They include persistent naming, format migration and/or emulation, inventory management, volatility, replication, re-validation, curator-operator error, and storage.

Appendix B. Participants

Beth Arthur
Archivist
National Association of Realtors

Jackie Druery
Head, Donald E. Stokes Library
for Public & International Affairs and
The Ansley J. Coale Population Research Collection
Wallace Hall, Princeton University

Teri Embrey
Librarian
Pritzker Military Library

Eboni A. Francis
Resident Librarian
The Ohio State University
Food, Agricultural and Environmental Sciences Library

Deborah B. Gaspar, Ed.D.
Instruction and Collection Development Librarian
Gelman Library
The George Washington University

Dave Green
Associate University Librarian for Collections and Information Services
Ronald Williams Library - Northeastern Illinois University

Elisabeth Long
Co-Director, Digital Library Development Center
University of Chicago Library

Missy Roser
ICON Project Coordinator
Center for Research Libraries

Appendix C. Participant Questionnaire

1. I work in:

_____	K-12 School	_____	Local Government Institution
_____	College or University	_____	Non-Profit Organization
_____	Federally Funded Institution	_____	Corporate Institution
_____	State Government Institution	_____	Specify Other: _____

2. My current position is: _____

3. I have experience creating a web archive: _____ Yes _____ No

4. The two most important user needs that a web archive will address in my library or organization are:

   a. _____

   _____

   b. _____

   _____

5. Two critical areas my library or organization needs to address in order to successfully implement a web archive are:

   a. _____

   _____

   b. _____

   _____

6. As I think about the reality of creating a web archive, the biggest hurdle I see for my library or organization is:

   _____

   _____

7. Your comments are welcomed. Please use back of page if you need more space.

   _____

   _____

   _____

Thanks very much for your help!