DigitalPRESERVATION
NATIONAL DIGITAL INFORMATION INFRASTRUCTURE AND PRESERVATION PROGRAM

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

Content Identification, Selection, and Acquisition Path

Content Provider Interviews:

Summary Report

April 18, 2006

Prepared by:

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

The following people contributed to this document.

New York University	Michael Nash
University of California - Davis	Linda Kennedy & Juri Stratford
University of North Texas	Inga Hsieh

Note: This report replaces the April 10, 2006 report. Table 2 was revised.

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. The content will be collected largely from US federal and state government agencies, but will also include political policy documents, campaign literature, and information concerning political movements and labor unions.

One focus of the project is to produce tools and guidelines to assist curators and other information professionals with collection development for web archives. In support of this effort, interviews with potential end users of web archives and providers of archive content were conducted in 2005. The purpose of the interviews was to elicit the needs and issues end-users and content providers have in relation to web archives.

This document summarizes the results of the interviews with content providers. Section 2 identifies the interview methodology. Section 3 describes the results and Section 4 discusses the major findings.

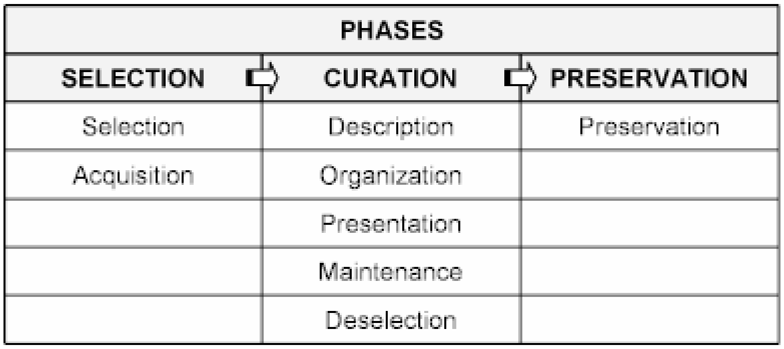
2 Methodology

2.1 Framework

In the second phase of the Web-at-Risk project, curators will build collections of web sites that share common topics, themes, or events. One outcome of the project is to identify activities, considerations, and issues curators might need to address in their collection development plans and policies. While librarians and curators are familiar with collection development activities for traditional print resources, this project wanted to identify any unique challenges inherent in building web collections.

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. (Appendix A provides a brief explanation of the activities in each phase as they apply to collection development for web archives.) It was expected that web content providers would offer important insights and requirements that could inform these activities.

Table 1. Collection Development Framework for Web Archives



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

2.2 Participants

Each of the project partners identified 2-3 content providers to interview. In all, seven interviews were conducted: three with union organizations and four with state government agencies. Table 2 identifies the participating organizations and Appendix B lists the individual participants.

Table 2. Participant Organizations

Project Partner	#	Organization Interviewed
NYU	3	Labor Unions 1. United Federation of Teachers 2. Transport Workers Union of America 3. American Federation of State, County, & Municipal Employees District Council 37 - New York City
CDL - UC Davis	2	State Government Agencies 1. CA Spatial Information Library & CA Environmental Resources Evaluation System* 2. California Legislative Data Center
UNT	2	State Government Agencies 1. Office of the Texas Secretary of State 2. Texas Building & Procurement Commission

* Interview was with the UC researcher who maintains these two agency web sites.

2.3 Data Collection & Analysis

The interviews were conducted by project team members, who used the interview questionnaire in Appendix C to guide the discussion. The six topics listed in Table 3 were discussed. Each topic provided information related to one or more of the web collection development activities.

Table 3. Discussion Topics

1.	Web-published Materials
2.	Digital Archives
3.	Access to Materials
4.	Authenticity of Archived Materials
5.	Intellectual Property of Archived Materials
6.	Agreements with Archive Providers

Interviewers summarized the discussions and identified the key points that emerged. The summaries were provided to the project's Assessment Analyst, who further analyzed the content and identified the themes and issues reported in section 3 of this report.

3 Findings

3.1 Materials

The subject of interest in each interview was archival of agency or organizational websites and the materials contained in them. In general, website content related to the mission of government agencies or topics of interest to union members. Depending upon the target audience, content was sometimes password-protected or database-generated and consisted of a wide range of source materials. Most websites included:

- Publications related to meetings (e.g., agendas for public meetings or hearings)
- Government publications (e.g., guidelines, legislative bills, or administrative codes)
- Information created for the website (e.g., organizational history or linked resources of interest)
- Databases (e.g., content of publications or news articles)
- Brochures (e.g., public information or topical information)
- Image files (e.g., a wide range of materials from GIF images to detailed map images)
- Technical and research reports (e.g., results of agency or organizational hearings or studies)

Labor Unions

With the exception of searchable news article databases on two union websites, none of the union websites generated web pages from a database. Two unions reported that their websites included video and audio content. Additionally, two union websites included periodical publications and news content. One union website linked to a web log. Websites for national unions linked to their local affiliates and conversely, the local union's website provided links to related national unions.

Agencies

Databases comprise various amounts of the materials for all of the government agencies. In one case, databases are the principal component of the agency's web content. In another case, an agency's major weekly publication is created from database content submitted online by various other agencies.

Agencies vary significantly in the number of their web pages that are programmatically generated using database content. The range is from zero percent to between 45% and 75%. Three of the four agencies used a forms-based interface to collect information from visitors. Additionally, between 25% and 50% of the web pages presented by these three agencies are customized using personal information about the visitor. Only one agency reported that their websites included video or audio content.

3.2 Digital Archives

Unions

None of the unions had previous experience with digital archives. Regarding the inclusion of databases in a web archive, one union would include their news item databases while two were not certain if they would include their databases. All could target the frequency with which they would want their web sites harvested, specifically, every 2-3 days, at least once per week, and once per month. The two main triggers for updates cited by content providers were collective bargaining activity (contract negotiations) and political events. Other triggers mentioned included elections, meetings, and conferences.

Agencies

The participants have a range of experience with both digital and web archives. One agency follows its state archival agency's guidance; another has been archiving its website since its inception 13 years ago. A third agency has a contract with a university to archive its weekly publication for historical access; another agency took a snapshot of the databases on their website 10 years ago but has not subsequently implemented an archival strategy.

Three of the four agencies would include databases in an archive of their websites. The fourth agency currently provides HTML and PDF formatted content to its archive vendor. Most of the databases change daily yet not everyone thought they would need to be archived on a daily basis. For one agency, an annual snapshot would suffice for archival purposes; for another agency, it is important to capture the database weekly; for a third agency, daily capture of their databases is important.

Some online materials are not addressed by state retention guidelines. In some cases, paper format remains the 'official version' of a publication and this version is retained in state archives. Web-published versions of publications that change over time are retained for varying periods, for example, six months in one case and 13 years or since initial web-publication in another case.

There was some discussion regarding the triggers for re-harvesting materials. Again, a range of triggers were identified. Triggers tend to be associated with current operations. For example, if information for a previous legislative session must be changed via manual intervention, that might trigger reacquisition of the information. Likewise, each time a new or modified administrative policy or rule impacts information on an agency website, re-harvesting might be triggered. It seems that the frequency of harvests is specific to an agency and is related to both the mission of the agency and the agency's posture in regard to archiving their web sites.

3.3 Access to Materials

Current Access

There was common commitment among both the agencies and the unions to make the information on their web sites publicly available. For privacy reasons, one union website requires password access to somewhat less than 25% of its web content (i.e., staff lists and seniority lists). One agency requires password access to about 45% of its content (i.e., vendor and customer data and applications).

A few web sites allow visitors to register for services, for example, a union website offers an 'alert' service that emails registered visitors selective web content based upon their preferences. One agency offers all visitors free weekly notices consisting of the table of contents of their primary weekly publication. This agency also offers other state agencies fee-based value-added services that include customized email notifications triggered by user-specified parameters and a search service that provides access to previous issues of its publication.

Archive Access

There was some discussion regarding the authentication mechanisms that would be required of an archiving organization to harvest the web materials and of end users to access the archived materials. Regarding access by archiving organizations, many of those interviewed were not exactly sure but indicated that some authentication would likely be negotiated between the content provider and the archiving organization. Others would only require authentication to harvest their databases, general web site pages would be freely available for harvesting. A few agencies are aware that their material is already harvested on a routine basis. One agency has a formal archival agreement and provides FTP access to its files. In the case of harvesting some databases, the impact on the server is a concern (e.g., slowing down its speed or crashing it).

Two of the unions thought they would control access in a similar manner as they now do with their paper archives, while the third union was not sure what access control they desired. Two of the agencies anticipate no end user access restrictions to their archived materials. However, a few agencies would need to have some type of authentication in place for access to certain content within their archives.

Harvesting Linked Material

Unions and some agencies share a desire to harvest specific linked content based on relationships between union organizations and government agencies. For example, national unions need the content from their local affiliates harvested and local unions need the content from their national union harvested.

In a similar fashion, state agencies need referenced content from federal agencies harvested (e.g., electronic forms or publications from GSA or EPA).

For one agency that serves as an information clearinghouse or one-stop shop for content linked to diverse information sources and publishers, the agency estimates they would require web site content from external links to be harvested to at least two levels or hops. Another agency indicated that harvesting material external to their web server is dependent upon the function of the archive: Is the archive a mirror site or is the archive a repository? If the archive is a mirror site, then external links would need to be operational.

Forms

None of the union web sites rely on forms and back-end databases to generate web pages. For agencies that rely on this ability, arrangements would have to be made between the content provider and the archiving institution. In general, a web-crawl to harvest databases is not feasible.

3.4 Authenticity of Archived Materials

Exclusion of Content

In general, there was concern among the content providers regarding removal of some web content from the archived version. However the concern varied in degree. For one agency, if notice of what was removed were provided that was satisfactory. For another agency, if only superfluous content (e.g., logos) was removed but the text content was wholly intact, that would suffice. Union organizations were concerned but were either unsure of the extent of the issue or thought material evaluation by archival authorities was common practice and had little issue with that practice for digital archives.

The content providers were asked if the authenticity of source materials would be compromised if users were given notification of specific content changes, such as removal of certain objects contained in the original web page, deactivation of email links, or deactivation of hyperlinks to external servers. In a few cases, content providers were adamant that this practice is not acceptable at all or in the majority of cases and would constitute a violation of content. Other content providers thought that certain proprietary pages ought not to be archived and a few observed that email links would likely become invalid over time.

It appears that unions want important linked content included in the archive and careful discernment given to what is excluded from the archive. Some agencies are mandated to create “official” versions of their information. How they satisfy this requirement is also mandated and influences their acceptance of alterations to content. It appears that some type of statement or disclaimer of an agency archive as “official” or as “unofficial” might be needed by some agencies.

Addition of Metadata

In general, content providers have no issues or concerns regarding the additional of metadata to web materials. Two providers would want to approve the metadata that would be added. One provider thought the addition of metadata might improve access to their archived materials.

3.5 Intellectual Property of Archived Materials

Unions

Union content providers generally thought they had intellectual property rights to the materials published on their web sites. The persons interviewed would need legal consultation to determine explicit copyright issues. Most thought that with legal permission it would be OK to reformat materials as necessary for continued access. Regarding ceding control of their intellectual property rights to archiving institutions, either they would not be willing to do so or needed to consult with their lawyers prior to rendering an opinion.

Agencies

Two of the agencies provide specific, comprehensive bodies of information content that is considered within the public domain and there are no intellectual property issues of concern to them. The other agencies provide access to materials that are somewhat, but not solely, in the public domain. Neither of these providers is willing to cede their intellectual property rights to archiving organizations.

3.6 Agreements with Archive Providers

Unions

The union content providers agree that the following should be specified in agreements:

- Content reformatting and migration terms and conditions
- Selective removal of archived objects over time
- Specification of the materials to be archived
- Copyright and intellectual property for the materials
- Support and maintenance

Two of the union content providers thought data (content) delivery specification (e.g., protocols and verification procedures) did not need to be addressed. However one provider did want it addressed in agreements.

The union content providers did not think minimal metadata expectations for data needed to be specified in agreements. In general, they agreed that archival agreements and practices for web materials should be consistent with existing archival agreements for paper materials.

Agencies

All of the agencies agree that the following should be explicitly identified in agreements:

- Specification of the materials to be archived

Three of the agencies agree that the following should be specified in agreements:

- Selective removal of archived objects over time (e.g., document/data retention guidelines)
- Content delivery specifications (e.g., protocols and verification procedures)
- Support and maintenance
- Copyright and intellectual property for the materials

There was some variance regarding specifying terms and conditions for the following in agreements:

- Content reformatting and migration
    - There are no agreements addressing this now with organizations who harvest their content so not certain how important this is.
    - Migration of material formats will be necessary over time.
    - Clarity in this regard is important for the content provider and the archive provider.
- Metadata
    - If providing datasets, two agencies thought metadata describing dataset structure and meaning should be specified.

    - One agency was quite positively disposed to archive providers using metadata to enable public access to their web site materials. It was not clear if they thought this should be specified in agreements.
    - One agency expected an archive to notify them if metadata were created by the archive. This agency also expected archived pages to include the time and date they were archived in the page header.
- Copyright and intellectual property for the materials
    - Materials in two agencies are not copyrighted
    - Others want these specified.

4 Discussion

4.1 Perspectives on Web Archives

What is a web archive? Different views of web archives emerged in the discussions. In large part content providers' views were influenced by their experiences with archives and their roles as providers of web-published content.

Web Archive as Extension of Print Archive

The content providers from unions have a long-standing history with an archival organization for their print materials and clearly want to extend that relationship to include their web-published materials. It is a matter of concern to them that agreements with archive providers are explicit in terms of what is to be archived and how the materials are to be maintained. They also think it is important that they retain intellectual property rights to the material. Some are adamant that their web sites not be archived without their express permission.

Web Archive as Crawl Result

Several agencies are aware that their sites are already being crawled by the Internet Archive as well as by Google, Yahoo, and other commercial entities. The content providers put no additional effort into packaging their content nor do they establish agreements with these organizations. This view of an archive is not without some concerns on the part of providers, particularly if the archive provider is reformatting and repackaging harvested materials. However, in the absence of formal archival agreements, content providers recognize they have little control. Content providers do expect 'good harvest behavior' on the part of crawlers, for example, respecting robots.txt and not impacting server performance.

Web Archive as Part of a Risk Management Strategy

A few agencies could envision an archive as a safe back-up site for disaster recovery or as a mirror site providing alternate access to their web sites. In either case, the archive would be expected to provide equivalent operational access and functionality. This view of an archive is more in line with an operational computer systems management strategy. Funding challenges have prompted systems and project managers to identify opportunities for effective risk management options beyond their own organizations.

Web Archive as Trusted Repository

Most content providers share a view of a web archive as a safe repository for specific web-published content of historical value that is beyond the purview of providers' own retention mandates or beyond their resource ability to retain. In this view, web archives function as repositories for posterity and enable research into historical records for analysis of change over time.

It is clear that the variety of content, organizational mandates and missions, as well as intellectual property concerns pose challenges for both content providers and archive providers. There appears to be no one-size-fits-all in terms of approaches and agreements between the parties. On the other hand, there is likely to be some one-size-fits-many approaches that can be identified. For example, providers of union content share much in common in their archival requirements as do some state agencies.

4.2 Content versus Packaging

What is being archived? The material content providers would like to archive ranges from databases, specific files related to publications, portions of web sites, external content linked to their web sites, and all materials on their web sites. Not only did concerns and issues emerge related to what is being archived but in regard to how the archived content will be provided to archive visitors.

Look and Feel

An issue emerged regarding replication of the 'look-and-feel' of content providers' web sites in archives. For one union content provider, reformatting and removing content over time posed concerns regarding the authenticity of the archived materials. For other content providers, their databases and datasets are the "meat" of their content and to varying extents all other content is superfluous. These content providers are not concerned with replicating their websites' 'look-and-feel'. One content provider hoped an archive provider might enhance access to their content through the addition of metadata, although making a profit on content repackaging raised concerns for another content provider. One content provider suggested that selective inclusion of materials in an archive, as well as omission of linked materials, might impact the usability of an archived web site.

Content Discovery

When datasets or final publication files are the content being deposited with an archive, issues regarding discovery of the content emerge. It seems that responsibility for providing a search interface to the archived content would fall to the archive provider.

Value-Added Services

Existing web sites generally provide search functionality within their site. Some provide email alerts based on registered visitors personal preferences. Some provide enhanced search and alert services for a fee. How would these services be offered by the archive provider? Would different services, reflecting the archival environment, be more appropriate?

4.3 Authenticity

Security and access are two major concerns of agencies: Security that ensures the integrity of the materials and access that fosters public accessibility to the materials.

> *"Our main concern is that the integrity of what is archived be maintained and kept as current as possible and that there is communication between archive and source group to ensure integrity. [We are also concerned] that no one would have access to the back side of the data and possibly change it. Integrity of access is a major concern to ensure that the average person as well as the scholar would have access."*
>
> -- Agency

What constitutes official information? One agency is concerned about other organizations repackaging harvested agency publications and misrepresenting them as "official" state sites. Another agency thought archived sites need to be identified in some way as "unofficial" sites.

In the case of archived copies of official materials, there is a concern that 'database back-ends' be secured so that the content cannot be altered without the express permission of the agency. However, at times archived versions of official materials need to be corrected. Mechanisms need to be in place for this. For example, there may need to be a notification service in place so that content providers can alert archive providers of changes.

4.4 Intellectual Property

> *"There are a number of policy issues and conceptual questions. The site is the 'public face' of the union and this means it is the responsibility of the elected officers of the union . . . Having something crawl into the sight . . . as a tool . . . will make folks very nervous."*
>
> -- Union

Materials accessible from government agency web sites and agency-sponsored web sites may or may not be in the public domain. Agencies that produce a product or publication or have responsibility for a number of publications in a defined area seem to regard their information as publicly available and not subject to intellectual property concerns. However, these agencies are concerned that archives either advertise content as 'unofficial' or at least not represent content as 'official'.

Agencies or agency-sponsored organizations that hold copyrights to some or all of their web content intend to retain those copyrights and archive providers would need permission to both harvest and modify web content. Likewise, non-governmental organizations generally have intellectual property rights for the materials they publish and any harvesting of content as well as deletions or modifications to web sites or materials on the part of an archive provider would require explicit permission from the organization.

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

Trade Unions

United Federation of Teachers
Tyrone Butler, Archivist & Records Manager
Tom Dickson, Assistant Archivist

Transport Workers Union of America
Dr. Robert Wechsler, Director - Education and Research

American Federation of State, County, & Municipal Employees
District Council 37 - New York City
David Paskin, Director of Research

California State Agencies or Agency-Sponsored Organizations

CaSIL (California Spatial Information Library)
CERES (California Environmental Resources Evaluation System)
Quinn Hart, CERES Technical Researcher, CaSIL Developer

California Legislative Data Center
Bill Behnk, Coordinator of the Legislative Information System
Linda Heatherly, Librarian in the Office of the Legislative Counsel
Three programmers

Texas State Agencies

Office of the Texas Secretary of State
Dan Procter, Director - Texas Register
Chair, Records Management Interagency Coordinating Council

Texas Building and Procurement Commission
Eva Dechene, Records Management Officer
Vice-Chair, Records Management Interagency Coordinating Council

Appendix C. Interview Questionnaire

Topic 1. Materials

1. Describe the central focus of the web materials you own or publish?

2. What types of materials are included?

Journals & Periodicals	
Books, Brochures	
Databases	
Newspapers	
Videos	
Audio files	
Image files	
Technical & Research Reports	
Proceedings of Meetings & Symposia	
Doctoral Dissertations & Master's Theses	
Government Records or Documents	
Unpublished Work & Publications of Limited Circulation	

3. How much of the data or materials you own or publish consists of:

	0%	≤ 25%	≤50%	≤75%	100%
Password protected files					
Encrypted files					
Forms for collecting data					
Pages programmatically generated with no database					
Pages programmatically generated using data from a database					
Pages customized using personal information about the visitor					

Topic 2. Digital Archives

4. Describe your experience with digital archive providers.

5. Would databases be included in the materials you might provide to a web archive?

6. At what frequency would the materials you might provide to a web archive change?

7. What do you think should trigger a recapture of this data or materials that change over time?

Topic 3.Access to Materials

8. What type(s) of access apply to the materials you own?

   a. Publicly available
   b. Access restrictions exist
   c. Fair use guidelines applicable

9. If access restrictions exist, how are they currently enforced and monitored?

10. What authentication mechanisms would you require of a web archive provider:

    a. To harvest your materials?
    b. For end user access to your materials?

Topic 4.Authenticity of Archived Materials

11. When websites are archived, hyperlinks between archived materials are preserved within the archive. How do you expect other links (e.g., email links and links outside of the archived web pages) to be handled?

12. If forms are used within your website and you do not provide the necessary data for completing requests issued via these forms, how do you expect the forms to be handled within the archive?

13. If an archive provider does not include parts of a web page or certain material formats in an archive consisting of materials you provided, would you consider the authenticity of the source material compromised?

14. Would it concern you if an archived web page was altered or tagged in some way to alert end users to a modification of the source material for the following reasons?

    a. Removal of an object that is linked into a web page
    b. Deactivation of an email or hyperlink because it links outside of the archive
    c. Are there other situations you can think of where modification of the source materials and alerting the end users to that modification is acceptable?

15. How would this type of alteration or tagging compromise the authenticity of the materials you provided?

16. What concerns would you have if an archived web page were altered to include additional metadata?

Topic 5.Intellectual Property of Archived Materials

17. How confident are you that your organization possesses control of the intellectual property included in all of the materials that you might make available to a web archive?

18. How do you address copyright permission for embedded content in web pages, for example, images or audio files?

19. What concerns do you have regarding copyright permission for reformatting and migration of materials in the web archive?

20. What intellectual property rights for the archived materials are you willing to cede to the archiving institution?

Topic 6.Agreements with Archive Providers

21. How do you think reformatting and migration issues should be addressed in agreements?

22. How should agreements address deselection (i.e., selective removal of archived objects over time)?

23. How should the following issues be addressed in agreements between content producers (information producers or publishers) and web archive providers?

    a. Specifications of the data (materials) to be archived
    b. Minimum metadata to be provided with the data (e.g., description of structure and meaning of data sets)
    c. Data delivery specifications: protocols, verification procedures
    d. Support and maintenance
    e. Copyright and intellectual property

Closing

24. From your perspective as an information provider, what is your primary concern about web archives?