Digital PRESERVATION
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

Focus Group Report:

University of North Texas - Denton - August 2005

Prepared by:

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

Inga Hsieh
Research Assistant, Web-at-Risk Project
University of North Texas
ikh0003@unt.edu

February 14, 2006

The following people contributed to this report.

University of North Texas Cathy Hartman

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 August 23, 2005 at the University of North Texas in Denton, Texas. The one and one-half hour discussion was facilitated by the Assessment Analyst for the Web-at-Risk project.

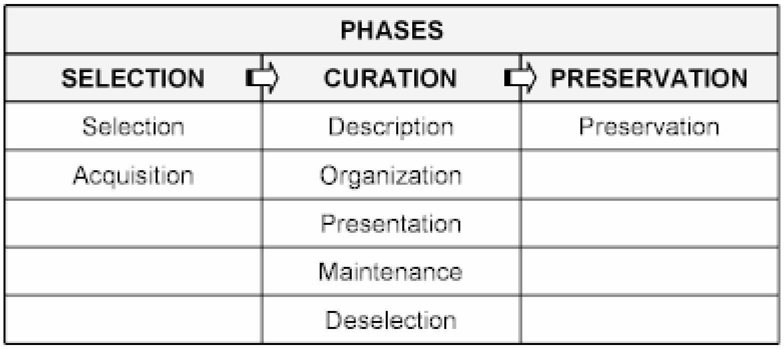
The report includes the following three sections: (a) the methodology used to conduct the focus groups and analyze the data, (b) the detailed results of the analysis organized into phases of the collection development process, and (c) a discussion of the key findings.

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

A total of seven people participated in the group discussion. Participants were librarians from the range of libraries within the University of North Texas, including the main library, media library, science and technology library, engineering library, and virtual library. (See Appendix B.) Three were department or library heads and all had some collection development responsibilities either as primary departmental liaisons or with electronic resource acquisition for multiple departments. Three of the seven participants indicated they had previously created archival copies of web sites or subject lists and a few had used the Internet Archive. Four had no experience with web archiving.

2.3 Data Collection

The discussion was recorded and subsequently transcribed. Additionally, 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 a questionnaire (Appendix C) that identified demographic characteristics and captured their thoughts regarding:

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

2.4 Data Analysis

Collection development provided the overall framework (Appendix A) 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 created.

These categories were used to analyze the content of the first focus group. Additional categories were added as necessary. This process was repeated for each of the four focus groups that followed.

Two analysts categorized the transcripts and notes from each focus group. Discrepancies between the analysts were discussed and resolved.

3 Findings

3.1 Policy

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

- Presentation of archived materials will require commitment and planning
- A few participants thought it would be a good idea to create a plan for the archive, which might include policies and practices for:
  - Metadata specification
  - Requirements for depositors
  - Content preservation criteria
    - Example: Specification of migration parameters (when, what, how…)
- For e-journals, it may be important to identify the data format or record construction of the content to ensure that the content can be accessed by different interfaces. This problem might emerge if the journal changes publishers.
- Agreements for depositors to an institutional repository or web archive would specify roles, responsibilities, services, and tools

<u>Organizational Support</u>

- Access and presentation of archived materials will require a major commitment for infrastructure and ongoing organizational support
- Some journal publishers are allowing faculty to publish a copy of their articles on their institutional websites. One participant saw this as a benefit both to the university and to a

global community of users (researchers). However, faculty members have not been generally supportive of the idea.

- Support has to come from the top of the organization and then be built into the funding fiber in a sustainable fashion, i.e., web archiving needs be a mainstream activity of the library/institution NOT a short-term project
  - Quote: “Because what good is all of this if next year there is a budget problem or somebody changes positions and the new person in that position says it’s not worthwhile to keep this particular project going.”
- Participants discussed ways to market the concept of a web archive as an institutional repository within the university and felt this effort was essential to getting organizational support and funding.
  - Market to university administration:
    - Market the concept that the repository/archive contained the ‘collective productivity’ of the university research community
    - Create presentations that sell the value of the repository and present these to administration
    - Target university unique content that sells well to administration, for example, the past year’s efforts at ‘branding the university’. This collection would consist of a number of electronic files that would of value to preserve.
  - Market to administration and departments via the library’s liaisons
  - Market to alumni by including their ‘stuff’ in the archive, for example, alumni with a certain specialty that are publishing (via the web)
  - Market to the university’s own communication group within Public Relations. They are already concerned with capturing and keeping information. Propose a collaborative effort.
  - Market to the faculty as a site for their research which could be of some value during the faculty review process or a patent application process.
- One participant brought ideas from an institutional repository conference:
  - Include formal research done collaboratively between a student and a faculty member. This effort is intended to engender student pride, which will hopefully translate into donations from them when they become alumni.
  - Mandate faculty deposits of their published research. If mandated, faculty will do this.
    - Quote: “It just has to be a requirement. You publish a paper; you put a copy of it in the institutional repository. If you just do that widespread, then it becomes the basis for success.”
- Another participant echoed the idea that support from university administration for the idea of hosting faculty research papers helps promote success.
  - Quote: “If he gets behind it, it’ll be a done deal.”

Institutional Repository

Note: It is likely that an institutional repository and a web archive would be built on a common infrastructure and would operate under a common set of policies and practices. This might explain why participants did not generally distinguish the two concepts.

- There was general agreement that it is difficult to get faculty buy-in for an institutional repository.
- The repository would prevent material loss in some cases:
  - Example: When a faculty member with responsibility for a website or the content in a website leaves the institution
  - Example: When funding is cut to a department or program resulting in discontinuation of support for a website published by that department or program
- The materials and content of web resources on subject guides are candidates for the repository.

- Including a duplicate copy of the faculty's published articles in the institutional repository provides an additional preservation 'safety net' while at the same time increasing the exposure and availability of the research. (Note: Not allowed by some publishers)

Financial Challenges

- Resources for cataloging, people, and hardware
- The archive effort will require infrastructure investment.
- The infrastructure to support user interaction (i.e., presentation) with archive materials is a "huge" challenge.
- For cataloging, money for staff resources is a bigger problem than lack of expertise
  - Quote: “We don’t have enough dollars to complete the retrospective conversion we started 20 years ago.”
    - As budget cuts were made, the retrospective conversion project sank to a lower priority.
- Sustainability of funding for cataloging web archives was seen as a major problem. Even if grant funding was obtained and catalogers hired, what happens at the end of the funding? The experience of the retrospective cataloging effort may extend to cataloging web resources, i.e., it may sink to a low priority in the face of budget cuts.
- Participants could not identify any library funds that were sacrosanct. This fundamentally poses a sustainability problem for a web archive or institutional repository.
- There was an acknowledgment that the technology used in the archive would change over time and there would be a future need for new technology and the funding to pay for it.
- In a collaborative archiving effort in which the university hosts the repository for entities outside the university, different business models could be developed. One variable feature in the pricing/service levels might be various levels of descriptive cataloging provided by the university:
  - Cataloging becomes a service offering and the depositors fund the required university resources.
  - Basic service example: Depositor supplies standard metadata for the web sites / materials they deposit in the archive.
  - Custom service example: Depositor contracts with the university for the level of descriptive cataloging they require. University produces metadata records for the depositor's materials at the contracted rate.

Technical Challenges

- Non-standardized content from publisher to publisher is a challenge.
- Skill sets and technology to support preservation activities are needed.
  - For resource format migration
  - The necessary technology platforms to render all resource formats must be available.
- The infrastructure challenge is “huge” for the university. There is a need for:
  - Expertise
  - Equipment
  - A “willingness to do it.”

Roles & Responsibilities

- For archive collaborations with or among small organizations, it was asserted that someone within each organization must be charge of the preservation effort. The analogy was made to websites of small town chambers of commerce. These websites are often out-of-date by as much as a year or more because the person who was responsible for building the web site no longer has this responsibility or has left the organization. In short, responsibility for preservation must always be someone's responsibility or the preservation effort will not succeed.

- For collaborative efforts between the university library and surrounding community organizations, (e.g., chambers of commerce and public libraries), large university libraries can bring to bear:
  - Their culture of preservation, with its perspective regarding the historical value of information
    - Institutions should take an active role in educating the community about why it is important to preserve their web content, what their responsibilities are, and where to go for help.
  - Their expertise and leadership in preservation
    - Providing standards and guidelines
    - Identifying and supporting hardware and software solutions to help communities create their own archives
- Should the cataloger of materials in a web archive be an archivist?
  - At UNT, the archivist is also responsible for implementing state-mandated records retention guidelines. The general sentiment was that it was a bit vague whether web sites were included under these guidelines. The implication being that if web sites were included, then the archivist would be archiving some university web sites by mandate and the recruiting effort might consider specific qualifications for candidates in this area.
  - One participant thought that archiving the intellectual content of the web was a very distinct activity from archiving physical objects.

3.2 Selection

<u>Identification of Source Materials</u>

- Unique content and collections not available elsewhere
- Web sites containing materials of historical value are important to archive.
  - Example: Web sites that include scripts from television programs that no longer air are important historical reference and research materials for Film, Radio, & TV studies
- Web resources used in courses
  - Student resource materials in web sites included in discipline-specific subject lists created by librarians
  - Student resource materials in bibliographies created by faculty for their courses
  - For some disciplines, web resources are critical and a good deal of time is spent identifying quality resources of long-term value for teaching (e.g., Women's Studies). It is important to retain archived copies of these materials.
- Materials created at the university
  - E-journal articles
  - Finding aids (e.g., for rare books)
  - Microforms
- Electronic resources (abstracts & indexes, e-journals, e-books)
  - Electronic resources are critical materials to be archived. These include e-journals, e-books, databases, and data sets. The latter two are of particular importance to scientific disciplines.
- Library-generated web resources
  - Subject guides
  - Class resource pages
  - Electronic resource guides
  - Distributed learning materials:
    - Ordering physical books and articles
    - Automated Interlibrary loan process
    - Online tutorials
  - Online catalog

- Participants identified selection criteria for web archive materials:
  - Supports collection goals
  - Consistent with course content
  - Quality and source of content should be considered
  - Supportive of faculty scholarship and research
  - Supportive of student learning and research
  - Attractively packaged; Ease of use
- Frequency of use as selection criteria for inclusion in the archive was generally panned by participants. A few reasons cited include:
  - Usage stats can be misleading; Usage does not equal value
  - 'Scary' for an academic library to use usage as a sole criteria
  - Some materials may not be used much but have historic or permanent value
  - If a resource is underused, it may be because it is not 'marketed' (i.e., students do not know it exists).
- The degree of dependency on materials affects the assessment of the risk should they disappear from the Web
  - For some disciplines, specific web resources are critical (e.g., for Film, Radio, and Television the Internet Movie Database - IMDB - and TV Tome) and their loss would be great. These are good candidates for archives.
  - For some disciplines (e.g., English), web resources are included in subject guides but do not play as dominant a role as web resources do for other disciplines (e.g., Women's Studies). The more dominant the role played by web resources in the overall collection for a given discipline, the more detrimental the loss of those resources will be to the research and scholarship of the discipline if the resources are not archived.
  - For some disciplines, current web resources are valuable but past versions of those web sites are not (e.g., Hospitality Management). Archives are not as critical for these resources.
- Materials from small publishing houses or society publishing houses are "preserved" by sending CD-ROM with previous year's content to subscribing institutions. There is no true ownership of preservation of these materials.
- Materials produced by communities are at risk
- Materials from academic institutions with no Institutional Repository or archive are at risk
- Web sites of long-term or historical value that are included in subject lists, used in teaching, and vulnerable to disappearing are important to archive.
  - Example: Many of the web sites in Women's Studies are created by other universities and these tend to be lost or disappear more often.

<u>Lost Materials</u>

- Materials included in subject lists often disappear
  - Regarding web sites listed in a health information bibliography, one participant stated: "I know that I've clicked on some links and (the information is) not there anymore . . . something that might have been quality at one time is either no longer there or it's moved someplace and we don't know how to get to it."
- Web resource materials produced at academic institutions are both volatile and vulnerable
  - Quote: ". . . the problem with academic institutions is that whenever money gets tight or something happens or maybe a faculty member moves away, well then that information (on a website) just disappears."
- Some ERIC resources were lost during transition
- A lot of staff time can be spent both identifying and keeping current links for web resources. When the materials disappear the loss can be quite significant to both research and scholarship and in terms of unrecoverable staff time.

- One participant related a problem in working with smaller public libraries: the libraries did not demonstrate a value for historical information regarding their own libraries or the activities conducted (e.g., annual web sites of book fairs)

3.3 Acquisition

Authenticity of Materials

- One person provided an example of quality assessment based on provider
  - Example: " . . . if it someplace like Cornell University or something like that, it is pretty much a sure bet that this is quality information."
- Because of the immaturity of the Internet, it is difficult to predict if authenticity would be threatened if a particular material format was not available.
  - Quote: "Is this (resource) going to be valuable to us 50 years from now to see the differences between the two (formats of the resource)?"
  - Quote: "I worry about a large website that we've got (archived). Is that (the website) someday going to be something that a historian and archivist in the future is going to want to look at in exactly as it actually appeared back then. I don't know the answer to that."

Frequency

- One participant noted that it is difficult to know what versions to capture and retain because it is seldom known when technology will change.

Source Material Versions & Formats

- The general consensus was that all versions and formats of materials should be retained.
  - Example: While students are often uninterested in materials if they have to use old technology to view them (for example a Laser Disc player vs. a DVD player), researchers may require older versions and formats of resource materials as well as the ability to render them.
  - In particular, researchers may be interested in different material formats because the uniqueness of each format or the differences among the formats are of research interest.
    - Example: Laser Disc vs DVD vs VHS
- Problems of retaining all versions and formats include:
  - Space
  - Organization for resource discovery

3.4 Description

Level of Description

- Students might well want detailed descriptions at the object level. However, all web users are used to much less granularity
  - Example: The usual practice of web search engines is to return very brief high-level descriptions.
- The level of description for a website might depend on the site itself.
  - For example, the amazon.com web site is searchable and the company has already created records for individual items. There would be no point in duplicating that effort and a simple descriptive record would suffice for the web site.
- One participant thought that library catalogs are created by librarians for librarians. Users may need something less sophisticated to accomplish their locating and evaluating tasks.
  - Example: The need to locate any information on fashion in the 1920's.

  - The counterpoint expressed by another participant is that there are a variety of users and some require the ability to locate materials using quite specific values.
    - Example: The need to locate an article written in the 1920's about fashion in the 1920's.

Original Cataloging

- There was a bit of discussion regarding the requirements for a web cataloger.
  - Should the cataloger be a web content specialist by profession and only secondarily a cataloger?
- Would be optimal to harvest metadata along with materials (for example, Amazon's item descriptions)
- Content creator might do some of the cataloging
- There was general sentiment that cataloging would pose many challenges.
  - Examples: ". . . it'll be harder than cataloging serials" . . . "It's like trying to catalog an art book."

Breadth of Cataloging

- Separate catalog records will be needed for each version of captured materials.

Standards & Guidelines

- What standard (e.g., Dublin Core) should be used for metadata elements?
- There was general agreement that standards and guidelines could be developed that identified both a set of metadata elements and the format for values of the elements that would be applicable to the variety of resources in an archive.
- There was also a general sentiment that establishing guidelines presented certain challenges:
  - What constitutes the title of a web page:
    - The page title in the URL?
    - The title included in the displayed content?
    - The title listed in the embedded metadata?
  - How should metadata included in a page or an object be handled?
  - What format should be used for date values?
    - Day-month-year (e.g., dd/mm/yy)
    - Month-day-year (e.g., mm/dd/yy)
  - Whose calendar will be used? Is there a universal calendar?

3.5 Organization

Note: No comments from this group applied specifically to the organization of resources in a web archive. However, some comments in other categories might have implications for this collection development activity. For example:

- Problems of retaining all versions and formats include:
  - Organization for resource discovery

3.6 Presentation

Intellectual Property Issues

- Legal barriers are potentially affecting longevity of websites such as TV Tome as these types of sites seem to come and go
- One participant relayed that two websites of collections were abandoned because it became too difficult and expensive to obtain copyright permissions for the photos and articles

- Another participant recounted that she had a large number of newspaper articles documenting the history of her campus that she would like to digitize and archive. She was dissuaded by advice that she would have a difficult time getting releases from newspapers.

Usability – Accessibility

- ADA requirements for materials in the archive would have to be applied to all of the versions. This would increase, for example, the level of transcription. By extension there will be an increased need for resources to ensure compliance.

Dark Archives

- Sometimes archives receive collections that have stipulations regarding future access, for example, not accessible for 50 years. These types of collections might need some type of dark archive.

Look and Feel

- Presentation of web-published information depends somewhat on the materials themselves.
    - Participants agreed that preserving the content of journal articles would suffice. Coupled with an understanding of their structure, the articles could be presented within any suitable interface/context.
    - However, some participants thought other content types would need to be presented in their original web context (i.e. exactly as they appeared prior to capture). This was of particular importance for historical research within many disciplines

3.7 Maintenance

Note: No comments from this group applied specifically to the maintenance of resources in a web archive. However, some comments in other categories might have implications for this collection development activity.

3.8 Deselection

Frequency of Use

- Do not use frequency of use as a deselection factor.
    - Usage statistics can be misleading.
        - Example: Low usage could indicate a discovery problem. Clearly, if no one can find the content, it will not be used.
- It's 'scary' for an academic library to use usage as a sole criteria. Academic institutions have a responsibility to keep and make available quality resources that support the curriculum even if they are seldom used.
- Usage does not equate to the value of a resource.

3.9 Preservation

Methods

- In regard to archiving journals, this question was posed: Does the interface to the journal contents need to be preserved or just the content? Participants generally agreed that the content was what was important to preserve.
- The general consensus was that all versions and formats of materials should be retained.
    - Researchers may be interested in different material formats because the uniqueness of each format or the differences among the formats are of research interest.

- Examples: Laser Disc vs. DVD vs. VHS

<u>Stewardship</u>

- Participants recognized two drivers behind a commitment to preservation.
  - Funding: Larger vendors and societies (e.g., Elsevier or IEEE) with enough resources will generally take responsibility for creating preservation archives.
  - Value to scholarship & research:
    - JSTOR archives the most important (leading) journals in specific fields.
    - Project EUCLID at Cornell archives independent and society mathematical journals.
- There was general agreement that responsibility for preservation rests with content producers (e.g., originators or owners). Whether small or large, content producers should preserve materials themselves, participate in a collaborative preservation effort, or make arrangements for someone else to preserve materials for them.
- Large journal publishers/vendors should take responsibility for archiving e-journals.
  - Quote: “The big vendors can afford to archive (their publications). We pay for that.”
  - One participant thought assurance that materials are being archived was needed from large vendors (publishers) of e-journals (e.g., Elsevier).
- Currently, some small publishers and some society publishers, provide subscribers with a CD as an archival copy.
  - One participant suggested this is already an outdated archival access method and thought that some university needs to step up and assume responsibility to archive these publications.
  - Example: Some mathematical journals are not necessarily candidates for JSTOR or Project Euclid.
- For some publications of small publishers, print copies are the only archive that can be trusted.
- For archive collaborations with or among small organizations responsibility for preservation must always be someone's designated responsibility or the preservation effort will not succeed.

4 Discussion

4.1 Dealing with Change

<u>Building & Preserving Collections</u>

Different challenges emerge when building or preserving collections of web materials versus traditional print materials. As one participant put it, archiving the intellectual content of the web is a very distinct activity from archiving physical objects. Most physical or pint materials are in forms that endure for some predictable period of time. The urgency to address the preservation of web materials is that their longevity is unpredictable and materials are often lost. One participant attributed the loss of web-based materials to ignorance on the part of the material owners or creators of their preservation responsibility.

In part this ignorance is related to the dearth of knowledge and expertise regarding preservation practices for web materials. Preservation practices for printed and physical objects are well-established within most organizations, whether coded into retention guidelines or collection management practices. Additionally, it is commonly known who has responsibility for preservation within the organization. However, for web materials neither preservation practices nor the designation of who is responsible for implementing them are established in most organizations. Web materials come and go quite easily and often no one assumes responsibility for preserving them.

Roles & Responsibilities

Preservation of web materials may challenge existing functional roles and responsibilities within an organization. For example the university archivist at UNT is responsible for implementing state-mandated records retention guidelines. The general sentiment of the group was that it was a bit vague whether web sites were included under these guidelines. The implication is that if web sites were included, then the university archivist would be responsible for preserving some university web sites. This led to a question regarding the current recruitment effort for a new university archivist: Should candidates' skills in the areas of web site preservation and archiving be considered? Previously, this would not have been a consideration.

4.2 What to Preserve

The following list of preservation candidates emerged in the focus group discussion. The list includes both classes of materials as well as specific examples.

- Discipline-specific web sites included in subject lists created by librarians and bibliographies created by faculty for classes
  - Some disciplines are more dependent upon web materials than other departments. This may be related to either the 'newness' of a discipline or the nature of its reference and resource materials.
  - Women's Studies
    - Program is new and has a small budget and is quite dependent on web resources, for example, free materials available from university web sites and other 'trusted' web sites (e.g., government websites)
  - Criminal Justice
  - Health Information
  - Film, Radio, & Television
    - Internet Movie Database (IMDB) and TV Tome
- Electronic Resources
  - E-journals (both licensed and unlicensed)
  - E-books
  - Databases
  - Data sets
- Library websites
  - Subject guides
  - Class resource pages
  - Electronic resource guides
  - Distributed learning information and guidance for students:
    - Ordering physical books and articles
    - Automated interlibrary loan process
    - Online tutorials
  - Online catalog
  - Finding aids
- University-created materials
  - Articles by faculty in e-journals
  - Student papers and publications
  - E-journals published by the university faculty
  - Locally published unique collections
  - Materials of long-term value to university
    - For example, the recent work on university branding
- Content from small publishing houses or society publishing houses
- Materials produced in the surrounding community

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. These results are listed below.

User Needs

In order of importance, the three user needs a web archive would address were:

1. Access to materials for research and reference
   a. Historical record of web-based information
   b. Unique and valuable electronic collections
   c. Materials created by the institution

2. Access to an institutional or organizational repository
   a. University and library web pages of long-term significance
   b. Repository of faculty published scholarship (e.g., E-journals and articles)
   c. To foster an identification with the university for faculty

3. Provision of value-added services
   a. Searchable content
   b. Friendly design
   c. Content that meets user-defined needs

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. (The two areas in item three were of roughly equal importance.)

1. Organizational support (university administration, faculty)

2. Technology (infrastructure, technical expertise)

3. (a) Policies related to web materials (what to archive, cataloging)
   (b) Resources (funding, staff)

Biggest Hurdles

Participants identified management commitment as the biggest hurdle their organization faced in creating a web archive. While a shortage of staff was mentioned by a few participants as another hurdle, the majority sentiment seemed to be characterized by the words of one participant who thought the biggest hurdle was: “Political will - we can find the expertise.”

It might be helpful to distinguish the concept of an institutional repository from the concept of a web archive. The institutional repository was described as embracing and preserving many classes of materials: university publications, university websites, faculty research publications, the libraries' web-published resources, and university e-journals. The web archive, which might be a collaborative venture with other organizations and institutions, might include: non-university web sites used in support of teaching and research, web sites of local community organizations and governments of historical value, and web sites of smaller academic or public libraries.

For the participants in this group, the sense was that an institutional repository might more readily garner funding and support from both university and library administrations than a web archive. However, it was postulated that a web archive could be implemented on the repository's infrastructure. Said another way, since university administrators often do not understand the need for the preservation of the web-published resources librarians see disappearing, then librarians ought to market the more readily understood concept of an institutional repository to administrators. Once the institutional repository is in place, the more altruistic web preservation work can be done using the repository's infrastructure.

4.4 Need for Collaboration

Ideas regarding collaboration emerged within this group in the context of the need to preserve at-risk web materials in three general areas:

1. Journal publications from small and society publishers
2. Web-published resources from smaller academic institutions and other trusted sources
3. Web-published resources from local government and civic organizations, including public libraries

It was generally agreed that large academic libraries have a preservation role to play in each of these areas. This is consistent with their expertise and leadership in preservation and their tradition of appreciation for the long-term value of intellectual and cultural history and artifacts.

Small Publishing Houses

In regard to journal publications from small and society publishers, the current practice of providing CD ROMs as backups could be replaced by one or two institutions assuming preservation responsibility for the materials for the publisher. In addition to preserving and providing access to the journals, universities could establish content and format standards so that diverse publishers' content could be managed and presented via standard interface tools without the need for costly customization. One model for this might be Cornell's Project Euclid for the distribution of peer-reviewed journals in mathematics and statistics [projecteuclid.org/].

Small Academic Institutions

Shared institutional repositories might offer a financially viable model for smaller academic institutions, for which archive and repository costs would be prohibitive in most cases. The repository could be hosted at either one of the participating smaller institutions or at a larger academic institution. One value for larger academic institutions in participating in such collaborations is the preservation of smaller institutions' valuable web-published resources, which are sometimes used by faculty in courses. Currently, this type of web resource is often lost and seldom preserved.

Local Government & Civic Organizations

This group spent a fair amount of time discussing collaborative efforts with community governments and organizations. They spoke of a preservation leadership role that their university could provide to these smaller organizations. The local governments and civic organizations in the smaller communities nearby the university host many websites for which there is commonly no preservation commitment. These websites may be out-of-date by as much as a year or more because the person who was responsible for building the web site no longer has this responsibility or has left the organization. The university could provide an archive service in collaboration with these communities.

The web archive could be a shared repository that included both the university's own publications of long-term significance as well as local web materials of historic and long-term value produced by the smaller organizations. As part of their leadership role, the university might develop guidelines and tools to facilitate smaller organizations in the deposit of their materials into the archive. This type of collaboration would promote good relations between the university and the communities it serves.

Service Models for Collaborative Web Archives

Leveraging the infrastructure investment in an institutional repository, a large academic university could expand its preservation scope to include one or more of the three classes of the at-risk web resources identified earlier: e-journals from small and society publishers, web resources from other academic institutions used in courses, and web resources of historic and long-term value from local government organizations.

To fund the expansion, the university might consider entering the archive service provider business. Smaller publishing houses, academic institutions, and community organizations might contract with the university for any of a range of preservation and archive services. Basic services might be offered at a nominal fee. Enhanced services would be offered at higher fee in accord with provisioning costs. The service model and fee structure could ensure that the archive was self-sustaining. In this manner, the university would both satisfy its own preservation requirements and provide leadership and support to enable smaller organizations to meet their preservation requirements.

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

Gayla Byerly
Reference Librarian
Liaison, English & Women's Studies

Danielle Cain
Electronic Resources Acquisitions Librarian
Co-liaison, Philosophy

Leora Kemp
Head, Virtual Library, Dallas Campus

Jo Monahan
Liaison, College of Education

Sue Parks
Head, Media Library
Liaison, Radio, Television & Film

Martha Tarlton
Head, Reference & Information Services
Liaison, Merchandising, Hospitality Management

Gay Woods
Head, Research Park Library
Liaison, Material Science, Engineering Technology, & Mathematics

Appendix C. Participant Questionnaire

1. I work in:

______ K-12 School
______ College or University
______ Federally Funded Institution
______ State Government Institution
______ Local Government Institution
______ Non-Profit Organization
______ Corporate 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!