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

Content Identification, Selection, and Acquisition Path

Focus Group Report:
California Digital Library - Oakland CA - August 31, 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

March 21, 2006

The following people contributed to this report.

California Digital Library

Tracy Seneca & Patricia Cruse
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Introduction</td>
<td>3</td>
</tr>
<tr>
<td>2 Methodology</td>
<td>3</td>
</tr>
<tr>
<td>2.1 Framework</td>
<td>3</td>
</tr>
<tr>
<td>2.2 Participants</td>
<td>3</td>
</tr>
<tr>
<td>2.3 Data Collection</td>
<td>4</td>
</tr>
<tr>
<td>2.4 Data Analysis</td>
<td>4</td>
</tr>
<tr>
<td>3 Findings</td>
<td>4</td>
</tr>
<tr>
<td>3.1 Policy</td>
<td>4</td>
</tr>
<tr>
<td>3.2 Selection</td>
<td>7</td>
</tr>
<tr>
<td>3.3 Acquisition</td>
<td>9</td>
</tr>
<tr>
<td>3.4 Description</td>
<td>10</td>
</tr>
<tr>
<td>3.5 Presentation</td>
<td>11</td>
</tr>
<tr>
<td>3.6 Deselection</td>
<td>11</td>
</tr>
<tr>
<td>3.7 Preservation</td>
<td>11</td>
</tr>
<tr>
<td>4 Discussion</td>
<td>12</td>
</tr>
<tr>
<td>4.1 Dealing with Change</td>
<td>12</td>
</tr>
<tr>
<td>4.2 What to Preserve</td>
<td>15</td>
</tr>
<tr>
<td>4.3 Needs &amp; Issues</td>
<td>15</td>
</tr>
<tr>
<td>4.4 Need for Collaboration</td>
<td>17</td>
</tr>
<tr>
<td>Appendix A. Collection Development for Web Archives</td>
<td>18</td>
</tr>
<tr>
<td>Appendix B. Participants</td>
<td>19</td>
</tr>
<tr>
<td>Appendix C. Participant Questionnaire</td>
<td>20</td>
</tr>
</tbody>
</table>
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 31, 2005 at the California Digital Library in Oakland, CA. 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

<table>
<thead>
<tr>
<th>PHASES</th>
<th>SELECTION</th>
<th>CURATION</th>
<th>PRESERVATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection</td>
<td>Description</td>
<td>Preservation</td>
<td></td>
</tr>
<tr>
<td>Acquisition</td>
<td>Organization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Presentation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maintenance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deselection</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.2 Participants
A total of eleven people participated in the group discussion. Six participants were from the UC Berkeley campus, one was from UC Davis, one was from the UC Office of the President, and three were from Stanford University. (See Appendix B.) Participants were librarians or curators from a variety of campus libraries and research centers. Three participants were in managerial positions and all had some collection development responsibilities. Two participants indicated they had built archives of web sites and three participants were currently either involved in or
investigating preservation projects. The remaining six participants had little or no experience with
web archiving, except perhaps as users of archives and repositories.

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

Collection Policies, Practices, & Plans

- To address user expectations, planning efforts need to:
  - Identify what materials are not available electronically
  - Identify strategies to address user expectations that cannot be met
  - Identify barriers imposed by search engines
- It is important to identify the users a collection is serving. Research institutions have a
  variety of users including research scientists, state agencies, undergraduates, and the
  general public. Each of these types of users has different information needs.
- “Web-born” has no relevance to end users, who generally expect everything to be:
  - Discoverable via Google
  - Available electronically via the web
- One existing archive in a “niche market” decided to archive only one material format but
discovered a need to establish a practice of also archiving the software to read each
version of the format.
  - Quote: “We’ll take them [web materials] as PDFs. We archive the software to read
    the PDFs. But we just decided that we’re saving PDFs and that’s all we’re saving.”
- The granularity and scope of a collection is influenced by the discipline.
  - Quote: “In law they get these massive compilations and they save the whole thing.”
Some participants hope that grant-funded studies, such as the current NDIIPP project, will result in a preservation method “in whatever format it takes so that each individual library doesn’t need to invent its own system for doing it [preserving web materials].”

Organizational Support
- There are no resources identified for cataloging web-published materials.
- One participant was frustrated with the number of meetings and the number of ‘hoops’ libraries have gone through with institution-wide preservation projects that have not resulted in a common preservation system and approach available to librarians and curators. In the absence of an institutional approach, this person thought the number of local preservation efforts happening within the libraries, including paper archives, was a valid option.

Institutional Repository
- The scope or policy of an institutional repository does not always embrace all the important creative work from campus research centers. When the centers cease to exist, perhaps at the end of their funding or mission, there may be no ‘person’ or organization responsible for researchers’ working papers or the “legacy” collections of the center.
- Example: CDL e-Scholarship repository guidelines don’t address this class of information content. As a result, papers from these research centers have been lost.

Financial Challenges
- Librarians identify print publications they would like in digital form but government grants for digitization projects have been cut back and publishers may not be willing to undertake the effort unless there is a national market (versus a local market) or unless the undertaking will be profitable.
  - Quote: “You see a lot of useful things out there but this money crunch, resources to make it happen . . . it’s very difficult.”
- It’s expensive to create an archive of web-published materials:
  - Expensive to select
  - Expensive to harvest
  - Expensive to create metadata
- Grant-funded preservation projects require a lot of organizational effort but address relatively small preservation efforts within the organization and take a relatively long period of time to complete. In short, there is a history with grant-funded preservation projects of little return relative both to the amount of effort invested and the magnitude of the preservation problem.
- Libraries face dual resource challenges: An existing staff shortage and a growing demand for staff to collect and archive web resources.
  - Quote: “There’s more stuff we would like to add [to our archive]. We do not have the resources to be able to do that.”
  - Quote: “It’s far more labor intensive than people realize to archive web stuff in my experience.”
  - Evaluating web resources (one at a time) for weeding will be resource intensive.

Technical Challenges
- Quote: “We probably do not have the technical skills to be able to do that [add more materials to our archive]. Those are harder to find.”
- Some servers are hostile to harvesting efforts and pose problems for spiders seeking to harvest their content for an archive.
- After a period of time, the data may be in a format that nobody can read
- After a period of time, no one may be able to interpret the data that has been saved in a meaningful manner or context.
Example: Even if all the numbers in a database are saved and can be read, how will anyone know what they mean?

Quote: “How does a future user know the difference between the New York Times and The Onion?”

Print materials allowed for a delay in addressing preservation because once it is published, it lingers. Web-published materials disappear quickly and “when it’s gone, it’s gone.”

Preservation of web materials currently involves a great number of individual decisions regarding selection, archiving, and deselection. Technologies and tools to streamline these activities aren’t yet available.

Roles & Responsibilities

There is a sense of urgency regarding preservation of web materials, a sense that it is critical to do something now, to learn by doing, because materials are being lost.

Quote: “Just do something starting now, not wait for four years for something... the magic package... because by then ... things will be gone!”

Example: Pandora (web archive of the National Library of Australia) is on its fourth or fifth spider software because they’ve encountered problems with each one as they’ve gone along.

Because of space and budget constraints, increasing pressure is exerted on libraries to justify their government information preservation efforts when a private vendor is making the information available. The issue for librarians is that they cannot count on how long private vendors will continue to make the information available.

Librarians have a role in advocating for preservation as well as in the selection of materials to be preserved.

Agreements between content providers and publishers might specify that providers include certain levels and types of metadata and that publishers preserve the content.

Maybe states should pass laws to mandate preservation of certain web-published materials.

Libraries are already adapting in response to the changes around them and need to continue to do so

Undergraduates need access to print collections. Recent experiments at CSU in Monterrey showed that undergraduates were lost without actual print materials in the library.

Libraries need metadata specialists for describing and cataloging web materials however librarians can help determine appropriate metadata for these special collections.

Librarians are in a “transition period”. They continue to apply their training and skills to print and other physical materials while facing an increasing need to apply their training and expertise, in combination with a new technical skill set, to web materials.

Librarians spend a good deal of time selecting web materials. These tasks are in addition to their existing work.

• Quote: “We’re still having to do all the traditional things with print materials and then, on top of that, we’re having to deal with web materials.”

• Quote: “It just seems like with this material we’re having to make so many individual decisions.”

• Quote: “We are all selectors and we’re used to doing this thing here [with print materials] and we’re trying to translate that to here [with web materials] and pick up a whole other type of responsibility that requires all kinds of different skills. Some of us can do it and some of us can not and really we should be looking at developing digital librarian positions in our libraries, somebody who can come in with the skill and devote the time to it.”
3.2 Selection

Identification of Source Materials

- Government agencies sometimes retain web-published materials for as short as a few months and generally for no longer than a few years.
- Web-published state agency materials older than two or three years are usually not retained because the agency assumes no one is interested in dated materials. The general population served by state government agencies usually wants relatively current information. However, academic researchers doing comparative analyses need publications over long periods of time.
- Collaborative reports issued by two or more federal government agencies, state government agencies, and university research centers are often difficult to identify. Discovery and acquisition of these reports may be via traditional methods, for example from the GPO, but it is often very difficult and may be accomplished by ad hoc efforts such as communication lists mentioning them. This is often because no one organization is identified as responsible for retaining and widely distributing the reports.
- Some state government publications are totally converting to web-published content yet the publishing agency is not ensuring that preservation efforts begin with the first all-electronic publication.
- Example: California Statistical Abstracts - The California state library’s ability to archive these abstracts lags behind the state agency’s web-born publication dates.
  - Quote: “I called the California State Library... I said, what are you guys doing? Are you guys archiving this, this digital version? And they said, well... they’re planning to, but they haven’t got to it yet...”
- Example: - A decision was made to contract out preservation of the printed version of the California state budget. One problem was that the publisher was unaware that there were different versions to be preserved.
  - Quote: “I remember asking the publishers, “This is wonderful! Thank you! Are you doing [printing] all the versions?” and they came back and said “What versions?”
- One effort among librarians and California state agencies has been to establish a list of ‘essential titles’ to be preserved. Acknowledging that not every version of every publication can be preserved, the list identifies those publications that ought to be preserved. Currently, in the absence of legislation mandating preservation of web-published state government information, this good effort has no real clout behind it.
- Information requests for “lost materials” can be a source of identifying materials for preservation (See next section)
- “Selection is, I think, underrated but valuable.” This quote sums up what one participant saw as the value subject librarians bring to selecting web materials to archive. Not all versions of many materials either will be or ought to be archived. Subject expertise is critical to the selection process.
• Quote: “I don’t want to be the guy with a pamphlet collection that’s worth absolutely nothing in 20 years but there are certain web sites that produce certain things that I will take a slice of at a certain date and I will version.”
• However, in many disciplines, it is very difficult to decide which version(s) of materials/publications to select for preservation. The files can be ‘massive’ and some disciplines need all versions of them.
  • Examples: Articulated maps, social science data sets, laws
• Selection in a web environment is labor intensive, more so than buying books. More work has to be done up-front. Because it is so labor-intensive, it would be helpful to automate the selection process where feasible.
  • Example: Having both a mechanism within software applications and a policy for individuals to routinely save their works-in-progress to an archive would reduce the selection work.
• Librarians may have to give up some control over selection of web resources compared to print material selection.
  • Quote: “So, we know what we’re getting [with print materials and microfiche] but when we’re just going out and harvesting websites and harvesting digital information we are going to be less familiar with the individual pieces of information than we were when we were able to actually handle each document. Losing that control doesn’t necessarily have to be a bad thing. It’s just really difficult to make that transition.”
• Dwelling on answering the question of what to preserve can bog down preservation efforts because no one knows what will be needed tomorrow and consequently what should be preserved today.

Lost Materials
• Ad hoc requests are quite common from users seeking information they previously obtained from web sites but can no longer locate or access (e.g., engineers creating research proposals or sociology students doing law-related research). Sometimes the information has disappeared completely and sometimes it’s been moved to a new location. Examples include:
  • Federal agency web sites
    • USGS
    • NOAA
    • FEMA
    • US Department of Interior
  • State agency web sites
    • State water resources reports/publications
    • State public health agency publications
    • Quote: “Nobody would ever know that it [an agency publication] was ever there unless you know that it used to be there [on their web site].”
  • Regional offices of Federal agencies
    • Sacramento Region of the Army Corps of Engineers
  • Regional offices of Federal agencies publishing materials in collaboration with:
    • Counterpart state agencies
    • University research centers
• State agencies that do not retain their own publications often come to academic libraries when they need their own, older, publications. Libraries currently use old print materials to answer information requests from these agencies (e.g., “Can you tell us when we changed names?” or “Can you tell us what our major accomplishments were 30 years ago?”)
• At the change of administrations at the state level, information from the out-going administration is lost.
  • Quote: “I remember when we were looking for Pete Wilson’s [CA governor, 1991-1999] speeches the day after the new governor came in and his [Wilson’s] page with
all those materials was gone. All of those press releases, his announcements, some of his speeches and stuff -- gone."

- When government agencies reorganize, web content is often lost.
  - Quote: “All it takes is some major government reorganization of agencies and it’s [the web site with agency publications] is gone.”

- When university research centers cease to exist, research papers and other materials are at risk of being lost in three ways:
  1. They might disappear as a center.
  2. Their data might not be interpretable over time.
  3. Their data might be in a format no one can read.
  - Quote: “Their papers just disappeared or are now on a CD ROM and that’s it with no other kind of access. We have dozens of these centers that are spawned and they’re all generating research, and depending on their grant, they may not be around forever.”
  - Quote: “The center we had, the Center for Working Families, all of their stuff had been on the web but then they lost their money or ran out and that was the end. So it [the center and its website] just disappeared. And so, I have this little CD [of their website].”

- Association conference papers that are not published in future journals or books may continue to be cited in indices but are not available through the association.
  - Example: ASA - the American Sociological Association

- Data sets published on the web that are routinely updated without preservation of and access to previous versions render comparisons of the data over time impossible.
  - Quote: “That’s the way the web will be. That’s true of a lot of data sets.”

### 3.3 Acquisition

#### Authenticity of Materials

- For maps and GIS data regarding environmental or natural resources and agricultural reports, it’s important for users to:
  - Know they are viewing an archived report
  - Identify the version control date
  - Be assured of the authenticity of the report

- Some authentication mechanism is needed for official legal documents that are published online by commercial vendors. With printed materials, publishers include an imprint vouching for the authenticity of the materials (e.g. “Reprinted from official documents produced by…”). “How do you build confidence into an online system?”

#### Frequency

- In legal research, critical source materials change daily, (i.e., “regulations are being promulgated, court decisions are being decided, and laws are being enacted”) and researchers need access to historical versions.
  - Example: “A lot of times you go to an agency and you can get their current manual of procedures but what it [the manual] was last week or last year or ten years ago, you don’t know. If you’re trying to do some kind of law-related project, that’s often very critical.”

- One participant envisioned a method for the acquisition of all versions of user-created files (e.g., documents) that involved integrating a background ‘Save in Repository’ action as part of the typical ‘Save’ functionality in application software (i.e., word processors, spreadsheets, etc.). If mandated by the state legislature or university administration, this proposed functionality for works-in-progress would ensure all versions were captured with little effort on the part of their creators. Libraries could offer a value-added service that essentially tracked and provided access to the various versions of files.
Source Material Versions & Formats

- In one research center with a niche market, a decision was made to archive only the PDF formats of research papers.
- Changes to print materials were traditionally indicated by an edition statement, on the web this usually does not happen.
- Dates on web-published materials are not reliable as indicators of change because content creators often neglect to modify the date to reflect the changed content.
- Some web-published materials include only the most current data and have no version control per se. One participant described an online database that is modified as updates become available. The publisher enters a notation regarding each modification. In this case, capturing each unique version would be a challenge.

3.4 Description

Level of Description

- In legal research, the date of materials, (i.e., when was the information in force or “this is the information as of what date”) is critical.
  - Example: “One of the commonest questions we get is: ‘What were the regulations of the such and such board in 1983?’ So, any kind of web archive for legal materials would have to be very careful [to get the dates right]."
- For maps and GIS data regarding environmental or natural resources and agricultural reports, “version control” is critical.
  - Quote: “In GPO terms I guess a version control so that you know when the information was gathered and when it was updated and a method for making that very, very clear to the user.”

Original Cataloging

- Data poses a challenge to metadata creation. Researchers and their collaborators are in the best positions, possibly the only knowledgeable positions, to create metadata for their data.
  - Quote: “…you have to know about a variety of things that you just can’t necessarily glean from … definitely not just from the raw data and you may or may not be able to glean it from the … whatever documentation they give you.”
- Providing tools for researchers to create metadata as they collect data would be beneficial.

Standards & Guidelines

- Librarians can apply their information management skills to the creation of metadata. In this regard, one library has created its own subject headings, a thesaurus, and their own authorities (for building names, software objects, authors, etc.).
- One library is using modified Dublin Core with some qualifiers and enhancements.
- One participant thought that while Dublin Core is pliable enough to use as a metadata format for many applications it may not be sufficient for special collections.
- Standards are “broad guidelines” and “infinitely elastic”.
  - Quote: “I’m not sure the standards are actually going to save us because the standards are infinitely elastic”
- Both publishers and funding sources could drive the acquisition and retention of research data and its metadata by making these required activities for both publishing and funding.
  - Journal publishers could have a policy stating they “will not publish a journal article unless your [the researcher’s] data is in an institutional repository with [a specific type of] metadata describing it.”
  - Grant-funding agencies could require specific levels of metadata application for data collected under the auspices of the grant.
3.5 Presentation

Look and Feel
- In legal research, the authenticity of archived materials is critical. The archive must be able to provide and present “some assurance that what you’re looking at is official.”
- For maps and GIS data regarding environmental or natural resources and agricultural reports, it’s important for users to:
  - Know they are viewing an archived report
  - Identify the version control date
  - Be assured of the authenticity of the report
- In a research center that decided to archive only PDF versions of research papers, it became necessary to also archive the specific viewer associated with the PDF file as some files could only be viewed in beta versions of the viewer that never got published.

3.6 Deselection

Frequency of Use
- Some items in an archive might be kept for political reasons (author is a senior faculty) but are rarely if ever used.
- One participant thought it might be useful to extend the Google practice of placing high demand items near the top of search results to their server environment by placing high demand materials on the server with the best performance and conversely place low demand materials “on the back-burner, on our slow server”.
- Usage statistics are readily available in a server environment; however there was sensitivity to applying usage statistics to weeding a web archive.
  - “You can just go on usage.”
  - “That’s what we do for storage - usage; hasn’t circulated in 10 years.”
  - “Breaks my heart [to deselect based on usage].”
  - “No method is perfect.”

Evaluation
- One participant commented that most of the effort in digital collections was currently on selection of materials. This person wondered if and when the materials in digital collections were being evaluated in terms of the library’s need for them.
- Storage is relatively cheap and that has contributed to a lack of focus on weeding digital collections.
- Some archives are considered ‘permanent’ and there is an expectation that they will never be evaluated for deselection of materials.

3.7 Preservation

Methods
- Some librarians and libraries are engaged in preservation efforts:
  - PDF files of articles
  - Paper copies of government publications with entries in the library catalog
- One participant speculated that web archives could extend the social bookmarking/personal archiving concept of Furl (www.furl.net) to create a service that allows users to select and archive entire web sites. In this manner, a social network or “society” would choose what is important and what continues to be important to preserve.
Stewardship

- State agencies often do not retain their materials beyond a short-term time period. These same agencies will come to librarians seeking archived copies of their own publications.
  - Quote: “You can’t count on those agencies to keep their own material. They’re looking for old information about themselves. So, they don’t even keep their own material, things that they’ve published.”

- State libraries are in a logical position to preserve state publications but in the case of California, the state library is understaffed and under funded.

- When consideration is given to the users of state government published information, the clarity of stewardship disappears:
  - Example: A state agency might be publishing materials for current consumption by the general public or for short-term reference by their own and other agencies. Should they assume stewardship for preservation beyond these timeframes?
  - Example: A research institution is primarily serving the needs of its researchers, who in turn often need access to long-term, historical government content. Should the university assume stewardship for long-term preservation?
  - Quote: “It’s the issue [of] is the agency [National Resources Conservation Service] really committed to archiving those things [soil survey reports and maps] or do we need to archive them as well.”

- Publishers’ commitment to preservation, or the lack of it, has different implications for print materials versus web materials.
  - Example: If a library has a print copy, then the publisher’s commitment to preservation is of little consequence because the library can preserve the print copy and make it accessible to end users. However, in the context of a library providing access to web resources but not having them in their possession, problems with preservation and future access emerge.
    - Who will take responsibility for preservation and access? Will the publisher? Should the library? Is some other organization already assuming those responsibilities? These are critical questions regarding preservation stewardship for web resources that frame the dilemma with which librarians are struggling.
  
- The Federal and State depository programs are an example of commitment to preservation on the part of the government as a publisher. The government transfers responsibility for preservation to the depository libraries. This transfer of the commitment to preservation from government to libraries began with print materials and there is an expectation that it extends to web-born materials.

- Regarding government documents, one participant thought private publishers’ value-added content was a good addition to a collection. But they thought there was a need to ensure that private vendors were not the only providers of government produced documents. Their reasoning was that government documents were produced with public funds and neither libraries nor the public should have to pay to access them. A public alternative to private vendors is still needed.

4 Discussion

4.1 Dealing with Change

“The things we’re talking about are basically the things we’ve always done with the print collection. But I think they’re just much harder with web-archived material.”
Building & Preserving Collections

Government Information

This group had a number of participants involved with federal and state government information at their libraries and institutions and so issues and concerns related to web-published information from state, regional, and federal agencies were discussed. With the continuing shift from print documents to web-based materials, responsibility for archiving government publications is often unclear or non-existent and many government publications are simply disappearing. Both cultural and historical information of value to a range of researchers is disappearing.

Librarians involved in state and federal depository programs have traditionally assumed a preservationist role for government publications. As a rule, print publications of government agencies are distinct entities. There is generally a first edition of a publication followed by mid-year or annual editions. Each printed edition can be reliably preserved as its own entity. Additionally, official legal documents published in print by commercial vendors carried an imprint attesting to their authenticity as official documents produced by some organization.

With web-born publications, an edition of a government publication cannot be relied upon to be a constant entity. Many participants have encountered instances of web-born publications that were altered and for which no indication of the alteration was evident either in a versioning scheme for the publication or in the creation/modification date. Also, there is no analog of the ‘authenticity imprint’ for online legal documents published by commercial vendors.

Commitment to Preservation

With print materials, libraries (not publishers) often assumed preservation responsibility. However, publishers’ commitment to preservation, or their lack of it, has different implications for print materials versus web materials. The preservation requirements posed by web information, in terms of both the expertise of staff and the quantity of resources, are very different from the requirements to preserve print materials. Libraries are often hard-pressed to undertake a role as preservationists of web-published materials. Critical questions regarding preservation stewardship for web resources frame the dilemma with which libraries are struggling: Who will take responsibility for preservation and long-term access? Will publishers? Can libraries?

Collection Development

While collection development activities for web materials conceptually parallel the activities for print materials, they are more labor-intensive and lack established practices and standard references. Material selection requires more up-front work and often involves individual review of materials. Criteria for weeding both digital collections (and by extension web archives) are not generally agreed upon. In fact, the concept of ‘weeding’ an archive poses a bit of cognitive disjuncture. Until an economic or technical driver emerges, such a dramatic increase in storage cost, libraries will continue to focus on building collections of digital materials rather than on deselecting materials from collections. It is anticipated that at some point maintenance and deselection of archived materials might become routine practice. It may be that digital materials will be moved to a virtual remote storage area, from which they will remain discoverable via their catalog records.

Digital Libraries

Attempts have been made both to convert existing physical libraries to virtual libraries and to create new libraries as virtual libraries. Some efforts have not yielded expected benefits or results. For example, Los Alamos National Laboratory had expected to decrease the number of
librarians needed when they converted to an all-digital library but discovered they actually required more librarians. And in a different vein, both California State University at Monterey Bay and UC Merced planned to create digital-only libraries but both discovered that undergraduates wanted access to print collections.

Roles & Responsibilities

Preservation

While preservation of materials has been a cultural responsibility assumed by libraries and librarians, many question if libraries can continue their role as preservationists in regard to web-published or electronically published materials. One librarian stated the concern this way:

“We’re kind of shifting. Something was published in paper and we had it. We took on the responsibility for preserving it and the issuing agency or publisher didn’t have to worry about it anymore. Whereas [now] if it’s being published electronically, they’re still not worrying about it. But the way in which we are being asked to take responsibility for preserving it is so different. I just don’t know if libraries have the resources to do it.”

Perhaps the scope of libraries’ preservation efforts for electronic publications may need to be limited to smaller, more targeted projects in support of their users. As another participant stated: “I think, just on a day-to-day basis, maybe all we can do is these little things that are of immediate use to our particular users.”

Motivation

The preservation of web-published materials requires participation from both publishers and libraries. However, there are conflicting values and motives between commercial vendors and libraries or between corporate profit and altruistic preservation. One participant thought new technologies, developed in the private domain and transferred to the public domain, might offer economically viable preservation solutions and bridge the natural divide between publishers and preservationists.

Transition

Building and maintaining collections of web-published materials have confronted librarians with the challenge of applying their information organization expertise to a class of materials that behaves badly and over which librarians have almost no control. This challenge has emerged amid shortages of standards, staff, finances, and infrastructure and with an undercurrent of preservation urgency for web-born materials of historic and research importance that vanish from the web at alarming rates.

For the most part, librarians continue to work in the familiar world of print materials while increasingly accepting responsibilities in the ever-growing world of electronic resources. While interested in embracing the challenges inherent in this new responsibility, they often lack the technical expertise, the resources, or both to successfully meet the challenges. Some librarians acquire new technical skills, some try new approaches to traditional practices, and some seek to collaborate with more technically trained librarians. While approaches vary, a number of local preservation efforts are underway and many librarians are hopeful that grant-funded preservation projects will provide solutions in the near term to ease the transition into management of collections of web-born materials.
4.2 What to Preserve

- State government agency publications
  - Agency websites
  - All versions of publication
  - Both print and digital formats
  - California “essential titles” list
  - State water resources control boards
  - Regional water quality control boards
  - State public health agency publications
  - Statistical Abstracts
  - California budget
- State government website content from each administration
- Federal government agency web-based publications
  - USGS
  - NOAA
  - FEMA
  - US Department of Interior
  - Web-based publications of regional offices of Federal agencies
    - Sacramento Region of the Army Corps of Engineers
  - Web-based publications resulting from collaborative efforts by regional offices of Federal agencies, their counterpart state agencies, and university research centers
- Institutional materials that are not accepted into the Institutional Repository
  - Websites and working papers from university research centers after they cease to exist
  - Legacy collections that have no where else to go
  - A university created statistical database and web site containing data about Mexico
  - Web materials identified as ‘lost’ by academic researchers and library patrons
  - Association conference papers that are not preserved by the association or published and preserved in other publications.
- Soil survey reports and maps
- At-risk geospatial data
- Documents from the California shared cataloging project

Preserving in Print

In the absence of preservation solutions for web versions of publications, some librarians are printing copies of the web versions and preserving the printed copies. This solution is employed to preserve web-born state agency publications. Likewise, the Institute of Governmental Studies Library at UC Berkeley has undertaken a print-based preservation effort of policy reports from various think-tank organizations.

4.3 Needs & Issues

At the end of the focus group discussion, participants completed a brief questionnaire. (See 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 responses echoed and provided a summary of the discussion itself. These results are listed below.
User Needs

By far, participants indicated the most important user need a web archive would address was:

1. Persistent access to a wide-range of materials for research and reference:
   a. Government publications - critical, web-born, endangered, fugitive
   b. Unanticipated research needs
   c. Core electronic-only content

The second need in order of importance was:

2. Provision of value-added services, specifically:
   a. Access to disparate digital collections
      i. Consistent interface to variety of numerical databases/datasets
      ii. Interoperability with other repositories and indexes
   b. Discovery of born-digital content
   c. Historical access by subject
   d. Authentication & version control

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. Four main areas were identified and are listed below in order of criticality. Specific technical aspects of preservation were the most frequently cited areas. Not far behind these areas were the areas of resources and policies for web archives. A few participants identified organizational support as a critical area that their organization needed to address.

1. Technology
   a. Preservation challenges (how to archive, capturing and storing files, metadata creation for a large number of files, preserving functionality embedded in websites, “fixing” versions and dates)
   b. Infrastructure
   c. Technical expertise

2. Resources
   a. Staff
   b. Ongoing funding

3. Policies related to web materials
   a. Identifying what to archive
   b. Metadata guidelines

4. Organizational support
   a. Creation of a central unit for preservation
   b. Within library
   c. Within larger institution

Biggest Hurdles

Participants identified resources, in terms of funding for staff and technology, as the biggest hurdle their libraries and centers faced in creating web archives. This was closely followed by policy and political hurdles, which for these participants meant gaining agreement in the organization regarding what to archive and what technology platform(s) to use. Lastly, participants identified staff shortages and ongoing maintenance of the web archive as hurdles faced by their organizations.
4.4 Need for Collaboration

State Government & Academic Libraries

When collections include government publications, then roles and responsibilities among the agencies and the university for the preservation of web-born state government publications need to be determined. Participants articulated their efforts to work with the state library to preserve web-born publications from state agencies. While uniquely positioned to undertake this effort, the state library is “severely understaffed” and has budget problems. Librarians also stated their frustration working with some state agencies in terms of identifying who is responsible for publication and preservation of web-born materials. There seems to be an opportunity at the state level for collaboration among state agencies, the state library, third-party publishers of state documents, and university libraries.

Local Collections & Networked Technology

One participant identified two components to preservation efforts for web-born or electronically published resources: a technical component and a social component. Some participants thought the national Digital Preservation Program was addressing the technical component through grants that funded the creation of technologies and standards to enable libraries to preserve web-published materials. It was suggested that libraries address the social component by identifying niche collections or items to preserve and by avoiding duplication of preservation efforts. Avoiding duplication of effort might be coordinated through networking among libraries. In sum, collaborations would let the national grant-funded projects develop the tools and technical standards, let individual libraries assume responsibilities for preservation of targeted or local web-published materials, and set up a coordinated network among libraries that identifies what each institution is preserving so that duplication of effort is avoided.

University System

There seems to be an opportunity for collaboration at both the preservation and staff levels within the university system. While several local preservation efforts are either underway or being contemplated within the university libraries, not all of the librarians and their respective libraries were involved in creating archives of digital materials. One participant stated, “We are waiting for help. We are waiting for a collaborative approach [within the university system].” Another participant hoped metadata specialists could be a shared resource among the libraries.
## Appendix A. Collection Development for Web Archives

<table>
<thead>
<tr>
<th>POLICY SETTING</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SELECTION</strong></td>
<td>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.</td>
</tr>
<tr>
<td>Selection</td>
<td>Web-published materials are acquired or ‘harvested’ using crawling tools, which either globally or selectively capture web-published materials.</td>
</tr>
<tr>
<td>Acquisition</td>
<td>Web-published materials are acquired or ‘harvested’ using crawling tools, which either globally or selectively capture web-published materials.</td>
</tr>
</tbody>
</table>

| **CURATION**  | 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. |
| Description   | 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. |
| Organization  | 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. |
| Presentation  | 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. |
| Maintenance   | 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. |
| Deselection   | Preservation challenges are numerous. They include persistent naming, format migration and/or emulation, inventory management, volatility, replication, re-validation, curator-operator error, and storage. |

<table>
<thead>
<tr>
<th><strong>PRESERVATION</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Preservation</td>
<td>Preservation challenges are numerous. They include persistent naming, format migration and/or emulation, inventory management, volatility, replication, re-validation, curator-operator error, and storage.</td>
</tr>
</tbody>
</table>
Appendix B. Participants

University of California - Office of the President

- Linda Vida - Director/Head Librarian
  Water Resources Center Archives

University of California - Berkeley

- Harrison Dekker - Data Services Librarian
  Doe/Moffitt Library - Social Science Data
- Susanna Hinojosa - Librarian
  Doe/Moffitt Library - State and Local Documents, Latin American Documents, Spanish & Portuguese
- Chuck James - Librarian & Information Services Manager
  Earthquake Engineering Research Center
- Debbie Jan - Head
  Public Health Library
- Beth Sibley - Librarian
  Doe/Moffitt Library - Political Science, Sociology, & Women's Studies
- Alice Youmans - Head of Reference
  Boalt Law Library

University of California - Davis

- Marcia Meister - Bibliographer
  Shields Library - Federal Government Information

Stanford University

- Elizabeth Cowell, Librarian
  State and Local Government Information
- Tracey Erwin - Geospatial Librarian
  Earth Sciences Library
- Will Wheeler - Curator
  Social & Behavioral Sciences
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!