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

Content Identification, Selection, and Acquisition Path

Focus Group Report:

New York University - New York City - September 2005

March 24, 2006

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

The following people contributed to this document.

New York University          Michael Nash
Contents

1 Introduction............................................................................................................. 3

2 Methodology........................................................................................................... 3
   2.1 Framework................................................................................................... 3
   2.2 Participants .................................................................................................. 3
   2.3 Data Collection ............................................................................................ 4
   2.4 Data Analysis............................................................................................... 4

3 Findings.................................................................................................................. 4
   3.1 Policy ........................................................................................................... 4
   3.2 Selection ...................................................................................................... 6
   3.3 Acquisition.................................................................................................... 8
   3.4 Description................................................................................................... 9
   3.5 Organization .............................................................................................. 11
   3.6 Presentation............................................................................................... 11
   3.7 Maintenance .............................................................................................. 11
   3.8 Deselection ............................................................................................... 11
   3.9 Preservation ............................................................................................... 12

4 Discussion ............................................................................................................ 13
   4.1 Dealing with Change.................................................................................. 13
   4.2 What to Preserve ....................................................................................... 14
   4.3 Needs & Issues.......................................................................................... 15
   4.4 Need for Collaboration............................................................................... 16

Appendix A. Collection Development for Web Archives................................................. 17

Appendix B. Participants............................................................................................ 18

Appendix C. Participant Questionnaire......................................................................... 19

Note: Spelling correction made in Appendix B on May 16, 2006.
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 September 23, 2005 at New York University in New York City. 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>
<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></td>
<td>Presentation</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Maintenance</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Deselection</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.2 Participants

There were eight participants in the group discussion. Six participants were from the NYU Bobst Library and two were from the Tamiment Library. (See Appendix B.) One participant held a managerial position and most had collection development responsibilities. The participants were generally familiar with LOCKSS and some directly participated in the NYU LOCKSS effort. While none of the participants indicated they had built archives of web sites, two participants are currently involved in the Web-at-Risk project and have some familiarity with web archiving. 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, one note-taker attended the focus group and created a record of the discussion as well 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

- There is a generation of wired students (especially freshman and sophomores) who expect everything to be accessible online.
- From a web archive perspective, how will curators define what they want to collect: As a crawl process? In terms of information a crawl reveals about a site(s)? As a specific crawl or crawls of a site? As the site itself?
- In order to appreciate a collection's value and legitimacy, its context and purpose are important. Within a particular context, even poor quality materials may be deliberately collected.
  - Example: Sample of scholarship in an important and growing area for which a comprehensive collection is desirable.
- Including web-published materials and web sites in collections is an extension of existing collection practices. The job is the same: selecting “materials that are appropriate for our users.”
- When web-based reference or source materials for a given discipline are not in the library’s possession or catalog, librarians are still cognizant of the materials and can direct users to them as needed.
  - Example: Historical census data on the University of Virginia website.
- Priorities for collections of web materials need to be driven by the curriculum at the institution and the corresponding assignments faculty give to students.
When determining a subject area’s focus and range for a web archive, collection plans could specify choices similar to those made for any collection.

- Should the collection have materials representative of the range of topics in an area or materials limited to one or more topics in the area? What’s more important to preserve?
- Quote: “When you’re close to a subject, like ‘progressive social movements’, you realize how much variety there is and if we’re collecting 100 websites under this project, do we collect 50 about terrorism or do we collect a representative sample of the variety of the whole?”
- Quote: “If I were making an archive I’d put all those association websites in and the data websites, but I might pick up a blog here and a rant there and put them in. I assume that even if nobody’s going to use it today, somebody might want to use it in the future.”

Policies for web archives might include:
- What format types will be supported at any given time?
- What format types will be preserved?
- If format migration will be done for various format types.
- If emulation will be done to preserve access to materials.
- How websites in the archive that become “broken” will be preserved outside of the archive:
  - Copied to microfilm
  - Printed on acid-free paper
  - Deaccessed in accord with a policy
  - Some other method

Organizational Support
- An ongoing, top-down institutional commitment to a continuing relationship among ITS and curators is required to successfully build and preserve web archives that are meaningful to the university community. This includes administrative procedures in support of the commitment.

Institutional Repository
- One participant wondered what faculty materials would be included in an institutional repository at NYU:
  - Numeric data in support of research
  - Digital images referenced in monographs
- Another participant is working on a white paper to address certain questions relative to an institutional repository at NYU:
  - Quote: “What can we promise to preserve given what standards of creation?”

Financial Challenges
- Time and money are the limiting factors when it comes to collecting and describing web-published materials.
- Monitoring the source and adjusting the capture frequency appropriately will probably require professional judgment. This will require resources.
- The ability to capture increasingly large numbers of websites will create resource challenges for creating metadata.
  - Quote: “Are you still wanting to create metadata for every individual site or for collections of sites. [This] may be an issue down the road.”
- Cataloging will present a significant resource challenge. It is already a challenge with the print materials.
  - Quote: “It’s already a problem for us.”
- Identification of the resources that will be of interest to future researchers is very important, but extremely labor intensive and hard to push for.
Roles & Responsibilities

- Libraries have a role in making scholarly web publications by their institution's scholars accessible and in providing assurance to users that the materials are scholarly and valid.
- Building collections of web sites confronts curators with new and unexpected technical responsibilities (e.g., specifying crawl parameters or reviewing crawl results).
  - Quote: "It's hard to separate the technical from the curatorial. I thought it would be a simple thing; I'd just have to think of nice sites but it's a lot more complicated."
- Librarians can assume the curatorial responsibilities but need technical expertise and support from outside their organization for new and 'cutting edge' technical endeavors.
  - Quote: "What has made LOCKSS feasible for all the people here at NYU involved with it is precisely that we had one person and only one person doing the technical work and that's his specialty. And I think that's why we were able to get a certain amount of buy-in from all the different subject specialists. Because it was made clear from the start that that [the technology] was something that nobody had to touch."
- The library IT group only supports the library's desktop computers and the library catalog.
- Some participants do not think the library IT group has the staff to support web archives.
  - Quote: "I don't think they [IT] have the staff to deal with it."
- The scope of the technical infrastructure and support (e.g., storage, network infrastructure and back-ups) needed for web archives requires the involvement of the campus IT organization. This involvement needs to be ongoing and requires a commitment to collaboration between curators and IT staff.
  - Quote: Librarians need to know "... how the technologies affect the curatorial issues ... and then [need to have] regular technical support because without that you can't really do the curatorial side. It's hard to get that regular technical support at this point."

3.2 Selection

Identification of Source Materials

- Scholars in philosophy and classics want to access textual materials for different types of primary sources (e.g., books or glossaries) and the scholarship written about them.
- Some organizational web sites that primarily focus on current social activities also archive relevant historical materials in a wide variety of formats.
- One participant observed that published articles in the fields of labor and the left are increasingly citing websites. It is not clear yet if it will important for scholarship in this area to capture these websites.
- The range of an institution's scholarly products that reflect new and emerging publishing methods and vetting mechanisms
  - Student and faculty produced working papers and resources.
- Scholars in sociology generally use finding aids to determine if field trips to physical archives appear worthwhile. Web-accessible archives of sociological finding aids describing materials in sufficient detail would increase the probability that a researcher's field trip would be worth their effort.
- Government documents are published in two parallel web sources and both are important to capture: (a) web sites with original publications from government agencies and non-government organizations and (b) web sites of reformatted government publications available from private publishers.
- Other university websites with content related to subject areas
- Context is critical for both selection and presentation
  - For selection, it is important to remember that an individual item selected for a collection may not be of great quality in and of itself, but may represent a certain aspect of the collection.
    - Example: The French Literature collection may contain a trashy novel because it is representative of popular literature for its time.
• Example: Scholarly work that is not well done may be the only work available for a certain area of study.

• For certain disciplines, research libraries don't offer adequate coverage and subject selectors rely on the web to access publications from research institutes and associations.
  • Example: Requests for information about a region or a town cannot usually be satisfied with libraries’ print materials. Librarians find and rely on quality information they discover on the Web.
  • Example: Biographies are one example of information on the Web that is applicable to a range of information needs.
  • Quote: "Often I can't use what's in print in research libraries and I track down source material that's out on the web. A lot of this stuff is open access stuff. It's not licensed. It's not under any kind of bibliographic control. It's important and it's at risk."

• Cultural Studies is a discipline which relies on non-traditional source material.

• Sites produced by less stable organizations (i.e. ones that have the potential to suddenly lose their funding) are important to archive.

• Selection is very important, but difficult because it is hard to identify today the materials that will be important in the future.
  • Example: If the web had been around in the 60’s, 70’s, & 80’s it would have provided a view of what feminism was about over this period of time. The challenge would be to know at the time that the materials would be of interest in the future.

• Key events that emerge amid lots of media attention and their related ‘grassroots’ information sources are important to capture for future researchers.
  • Example: Bush/Gore vote count in Florida

• Selection of web sites for some areas of study is “labor-intensive but an important effort”.
  • Examples: Terrorism, Neo-Nazis in Europe, Women’s groups in the Middle East, Death squads in Latin America
  • Quote: “Selection is a must because we can’t do [capture and preserve] it all. And no one else but the curators and the scholars can help push libraries and archives to do what’s necessary.”

• Criteria for prioritizing materials for selection are:
  • Use or demand
  • Consistency with the historical collection areas of the institution
  • Materials that will augment the current collection
  • One participant mentioned that the ideas produced in FRBR report from IFLA (i.e. work – expression – manifestation – item) might inform thinking about web archives

Unit of Selection

• The unit of selection (e.g., image or page or website) depends on the discipline and the purpose of a collection.

• For certain disciplines or types of research (e.g., anthropology and history), source material context is critically important. Building archival collections comprised of ‘parts’ of web sites (e.g., selected images or videos) would be a disservice to future scholarship and research.
  • Quote: "I think with web sites and with political performance video and that sort of thing . . . and audio interviews . . . I think that there needs to be context in order to begin to understand [the content]."
  • Example: Sites that are thematically structured, such as "the excavation site for a particular dig." "It really wouldn't make sense to select particular parts because it's [the website is] designed to be an integral [whole]."

• For other research fields (e.g., statistical research in sociology), the original web-context of the source materials in an archive is not always critical and might present discovery challenges for users.
  • Quote: "I know that most of the things that I take off the web for people are things that were difficult to find on a website and they're not really interested in the context
for the statistics we were looking for. . . . They just want the original documents, the original information. And you know this is stuff that five years from now, if it's caught continually, it will make excellent longitudinal studies."

Lost Materials

- University publications: Web-born working papers and other resources often disappear after a few years.
- Participants had experienced returning to the Web to access information they previously used only to discover that it is no longer there.
  - Quote: “Sometimes somebody will ask a question and I’ll want to refer them to a website, but I don’t because I’m scared that the information might not be there. Sometimes I’ll go find the information [online] and after that I’ll go try to find it in a book and say ‘use this book’.”
  - Quote: “There are those sites that just disappear into thin air, that are here today and gone tomorrow”

Capturing Content from External Links

- One participant mused that if a user encountered a link within an archive to an external source that was not included in or accessible from the archive, the user would react the same way they do now when they discover that the library doesn’t own a book that is referenced by another book in the library’s collection: “They come screaming to the subject librarian and say ‘why don’t you own this?’”
- It might be feasible to extend the practices used for print materials to web sites.
  - A book or journal may be in the collection but not all of the referenced materials within the book or journal will be in the library’s collection. Only those deemed important for the collection will be there.
  - Could there be an interlibrary loan service for web materials not included in the web archive?
  - One participant pointed out that a web archive could explain the extent that a website(s) was archived.
  - It would be valuable to identify the URL for web sites referenced in materials not included in the archive. This would enable users to use a web browser to go to sites external to the archive.

3.3 Acquisition

Authenticity of Materials

- Libraries have a role in providing assurance to users that the institution’s scholarly web publications in their collections are scholarly and valid.
- Librarians, as selectors, can only go so far in judging the validity and authenticity of web sites. Users have to bring a level of discernment to bear as well.
  - Quote: “It behooves the person who’s looking at it to see: Where are they [the authors or publishers] coming from? What’s their bias? Can I trust this or not?”
- To some extent the academic level of end users (e.g., freshmen or seniors) is related to how they measure and value the authenticity of online materials (i.e., the higher the academic level the more discerning the user).
- Because of technical issues involved in archiving dynamic websites, questions emerge regarding what constitutes a scholarly or authentic representation of the site.
  - All changes are not likely to be captured.
  - Some material formats may not be captured
Frequency

- Participants agreed that change in source materials is highly variable and depends on the content.
  - Example of constant change: A web page containing a business FAQ shared by several institutions, who can each add questions and customize which questions are presented to their institution.
  - Example of little change: A web site containing images of ancient archeological sites. New images may be added but the images themselves do not change.
  - Example of predictable change: Web sites containing datasets of vital statistics or educational reports that change annually.
- Regardless of the frequency with which source is being re-harvested, some content will likely be lost.
- Automation can help identify changes in content for re-acquisition but it cannot entirely replace professional judgment regarding what content changes should trigger re-acquisition.
  - Example: Frequency would need to increase around political campaigns. This type of need is something that would require a professional resource to identify.

Source Material Versions & Formats

- Need to save all formats of an item as each contains different information
  - Image example: TIFF versus JPEG formats
  - Digitized manuscript example: “Having a shot of the text in different sizes will definitely be very helpful”
- Some formats are acceptable where others are not
  - Example: Law students require PDF formats
  - Quote: “Html is more susceptible to someone changing it, whereas PDF is more of a snapshot they can trust.”
  - File extensions do not always accurately reflect the actual format of a file.

3.4 Description

Level of Description

- In certain disciplines (e.g., popular culture), applying metadata to individual objects (e.g., ads within publications) would help discovery and increase the utility of an archive.
  - Example: Images of Kraft dinners or of Ronald McDonald in advertisements from inception to present.
  - Quote: “Freshman and sophomores, in the field of popular culture, want a snapshot of a particular period and will want indexing to advertisements. It's like every ad that isn't indexed in some way is less useful to them.” “It’s a broader brush to the way scholars may look for things.”
- The level of acquisition may be different from the level of description needed.
  - Example: May be useful to acquire an entire website but describe specific web page or object level descriptions.
- It is important to “provide some higher-level topical access, even if it is only derived only from the title as opposed to the actual content.”

Original Cataloging

- Participants generally agreed that there would need to be a good deal of original cataloging of the materials in a web archive to make it useful to the range of archive users.
- Consideration must be given to the trade-offs between the value and usefulness of metadata and the amount of time and effort required to create it. New approaches that apply technology and include users might provide “indicators of usefulness”.
The Web-at-Risk: Focus Group Report - NYU - New York City - September 2005

Kathleen R. Murray 10 of 19 March 24, 2006

- The more certain sites in an archive are used, the more useful they may be to users.
- Users could add their own metadata to web sites or materials in an archive.
- Quote: “If you’re creating an interesting archive, the more people use it, the more valuable it becomes; parts of a website become increasingly valuable as other parts become increasingly devalued.”
- Quote: “. . . the wiki aspect to metadata . . .”
- Examples: Google’s relevance ranking; Amazon’s user reviews
- Evaluative metadata might be automatically generated based on usage of materials or websites within an archive. However, one participant cautioned that was an inherent risk in such data and that it ought not to replace scholar’s individual judgments.
- Online citation indexes, such as Arts and Sciences Citation Index and Social Sciences Citation Index, are examples of applying usage (i.e., frequency of citations) metrics to publications.
- Quote: “We use general citation reports but, in the end, a lot of that is useless and the good articles are articles that are good regardless of where they appeared or who wrote them. And I think it’s always up to the scholar [to assess the value of source materials].”
- Quote: “We re-catalog websites already a lot of times.”

Breadth of Cataloging

- Metadata requirements vary widely depending on the users: Undergraduates may need individual objects within a web site or journal described while some scholars may only need finding aids.
- Human-generated metadata for archived websites will quickly get beyond the resource capabilities of a library as the number of websites harvested increases. Automated metadata generation will be needed and might minimally include:
  - Topics
  - Number of pages
- Librarians cannot create exhaustive metadata for source materials and cannot anticipate all future resource discovery needs. Researchers need to use their own searching skills to find the source materials they need within collections.
- With print or electronic resources that include a table of contents, (e.g., books), there is no ability to include the table of contents within the MARC record. Cataloging web sites, that will likely not include lists of their contents, would present a big resource problem.

Standards & Guidelines

- Use familiar tools from the print world as models for electronic resources
  - If users typically use online catalogs and journal indexes for books and journals in print, then it would be best to provide similar tools for electronic formats of these materials.
- Some emerging disciplines use source materials differently from how catalogers are used to describing them, for example, an advertisement within a trade publication.
  - Example: Cultural Studies is a discipline that often seeks source materials from a wide variety of approaches and at various levels of granularity
- Archives generally have a one-to-one relationship with finding aids. It seems that a web archive might have one-to-many relationship with finding aids.
  - Example: The web sites in an archive might have one finding aid listing web page titles and another finding aid listing a specific type of content included in web pages.
- Resources in a shared web archive might be cataloged in accord with guidelines customized for library-specific sets of users. This could result in a centralized archive with distributed catalogs uniquely describing the archive's resources.
  - The practice of submitting additional subject headings to the LC for a book to make it more useful to a library’s particular users is an example of this in current cataloging practice.
3.5 Organization

User Expectations
- Electronic finding aids are confusing and frustrating to end users who expect to access the materials described.
  - Quote: "They're under-whelmed with citations and finding aids. It's like: What good is this thing?"

Subject or Departmental Lists
- There was general agreement that the same criteria for inclusion in subject guides applies to print materials and websites.
  - Example: Stable sites that are updated: Association web sites and university web sites.

3.6 Presentation

Usability - Accessibility
- End-users may require natural language searching in addition to indexed search terms.

Look and Feel
- Researchers in some disciplines need to experience the original context of the source materials in order to understand them.
  - Example: Videos produced in difficult political and sociological situations may have poor quality. The poor quality may not be understood without the context of where the videos came from and under what conditions they were produced.
- Researchers in other disciplines are best served by working directly with extracted portions of original source materials.

3.7 Maintenance

Technology Roles & Responsibilities
- For cutting edge technologies, librarians will always need technical support. It is not feasible for librarians to become technical experts.

Access
- Over time some archived websites will become ‘brittle’ or ‘broken’. This may happen because the content format can no longer be rendered.
  - Archives need to be able to identify these websites.
  - Archive policy may guide which material formats will be emulated or migrated. Those formats outside the guidelines will no longer be rendered.
  - Quote: "Is there a reporting function that will [let the archive] know by itself that 65% of the websites are already broken?"

3.8 Deselection

Frequency of Use
- Usage reports range in value depending on their structure and specificity.
  - If simply the number of hits a site received is included, this is of little value.
  - If the frequency of use for each question within an FAQ database was known, then this could serve as a measure of how effective users’ needs were being met.
• Could be of value in designing user interfaces to web sites (e.g., more heavily used items could be top-level links).
• It would be useful to measure the amount of time that people are spending at the site.
• Usage of an archive is not important as an indicator of its value. Archives are created because they contain materials deemed important by the archiving institution.
• However, “archives do deaccession materials.”

Evaluation
• Participants thought buying additional servers for storage might be a more economical solution than weeding a web archive.
• Determining the value to researchers of a special collection is done in part through curators’ interaction with the researchers who visit and use the materials in the collection. One participant wondered if this feedback mechanism would ultimately vanish as researchers move away from a reliance on physical objects in libraries and archives to a reliance on web-accessible archives and collections.

3.9 Preservation

Methods
• File format changes over time present preservation challenges
  • Quote: “A number of years ago Real supported mpeg or video and then suddenly you upgraded Real and it can’t play mpeg4 files any more.”
  • Quote: “There’s no guarantee that in five years we’ll be able to look at jpegs at all.”
• File extensions do not always reflect file formats. File content needs to be validated for format type. If the format of the file cannot be verified, steps cannot be taken to preserve it.
• PDF files may contain JPEG formatted content or “broken” objects.
• Images on websites are often “derivative assets” as opposed to preservation quality masters. In the digital library community there is an effort to preserve master copies but it is recognized that the “derivative assets” are “throwaways” that cannot be preserved. What are the implications of this for web site preservation?
  • Quote: “When you say you’re going to preserve the websites, what is it about the website that you’re going to preserve? Can you really promise to preserve the content of the website . . . for such complex objects that we’re creating in the digital library team?”

Stewardship
• The organization or individual responsible for producing web-born information ought to take responsibility for archiving the original materials. However, this is not generally the case.
• The business information arena offers this model of stewardship: Different libraries have agreed to create, maintain, grow, and make web-accessible collections pertaining to a certain business information area. These collections include materials they digitize as well as web-accessible information.
  • Quote: "Here in New York, New York Public, Columbia, and NYU work quite a lot together and so, for example, New York Public has agreed to continue to pursue trade publications in whatever format but mostly online. They will continue to provide access to that [collection] while we focus on certain other things."
4 Discussion

4.1 Dealing with Change

Building & Preserving Collections

Material Selection & Description

“In the library field we already have experience gathering and making resources available. If we transition from print, where we have our catalog, into providing access to website[s], it’s not really all that far off from what we already do.”

Selection of web-based materials is typically an extension of existing collection practices and plans. For some disciplines, materials that used to be published in print are now web-published. Formats have changed but selection has not been significantly impacted (e.g., government documents). For other disciplines, organizational web sites offer a wider diversity of genres and formats than their print material counterparts. Blogs are one example of a new format type of interest to some disciplines. For less-established disciplines (e.g., cultural studies), there is not an abundance of available print material but the web has enabled an extensive amount of material to be published digitally. In this case, selecting materials in support of the curriculum is much more labor-intensive than it used to be.

If web archives are centralized repositories offering shared access to many institutions, it might be possible to create institution-specific catalogs for virtual collections. Archived materials might be locally cataloged in accord with guidelines customized for sets of users within an institution, for example, freshman and sophomore undergraduates. This practice could result in a centralized archive with distributed catalogs that uniquely describe the archive’s resources.

Resource Discovery

There are practices or examples in the print world that can be extended to describing resources in a web archive. In terms of resource discovery, creating new tools for electronic resources modeled after familiar tools from the print world would be helpful. For example, if users typically use online catalogs and journal indexes for books and journals in print, then it would be best to provide similar tools for electronic formats of these materials.

Preservation Challenges

It is anticipated that digital material formats for today’s web content will become obsolete over time and, in some cases, it may not be possible in the future to render some archived content. While format migration may mitigate this problem, there are current preservation practices from the print world that may be applied to digital materials. These might include creating print versions of the digital materials and preserving them on microfilm or acid-free paper. Analogous digital preservation techniques to these print preservation techniques may also emerge.

Archives

Three current practices with traditional archives were discussed in relation to web archives: finding aids, material context, and deaccession. Each of these practices illustrates limitations and possibilities for web archives.

Archives generally have a one-to-one relationship with finding aids. The current level of detail and access in finding aids is of limited or no value to many users who need richer content descriptions and who expect online access to the materials in some form. It seems that a web archive might
have a one-to-many relationship with finding aids and, where feasible, link from descriptions to
digital formats of the materials. For example, the web sites in an archive might have one finding
aid listing page titles and providing hyperlinked URLs while another finding aid would list specific
types of content contained in the pages (e.g., advertisements) and provide hyperlinked URLs to
that content.

Archives of physical materials may or may not be maintained in a manner that provides a larger
context for groups of the materials. In some fields, the absence of context might be unfortunate
and limit research but in other fields the loss is irrelevant for research purposes. It appears that
the research or scholarly need for capturing and rendering the original web context of source
materials, versus providing web-published materials out of context, will vary with research
requirements and be somewhat related to the field of research.

Best practices for deaccession are well-established for archives. It may be that hybrid practices
that merge deaccession from the archive world and weeding from the library world will offer
alternatives for managing long-term storage and preservation of materials in web archives.

**Roles & Responsibilities**

*The job is the same: selecting “materials that are appropriate for our users.”*

While collecting and preserving web sites and web-published materials present unique
challenges and, in most cases, increased technical and curatorial resource requirements, many
of the activities involved from the curatorial perspective are an extension of current roles and
responsibilities.

Creating archives or institutional repositories enables academic libraries to extend their curatorial
role in two directions: (a) making their institution's scholarly web publications accessible and (b)
providing assurance to users that these materials are scholarly and valid.

However, the IT staff within the library will not be able to handle the technical support
requirements for web archives and institutional repositories. For librarians to meet their curatorial
responsibilities for web-published resources in support of their users, collaboration and support
from the campus IT organization is required. Additionally, the university administration must be
committed to the effort.

**4.2 What to Preserve**

- Government documents
  - Web sites with original publications from government agencies and non-government
    organization
  - Web sites of reformatted government publications available from private publishers.

- Discipline-specific materials in support of university curriculum
  - Classics: Collections of ancient scientific images
  - Philosophy: Collections of pre-published materials
  - Philosophy and Classics: Textual materials for different types of primary sources
    (e.g., books or glossaries) and the scholarship written about them
  - Sociology: Statistical publications in education and vital statistics
  - Cultural and political studies: Websites of current social organizations and activities,
    including historical materials
  - Excavation sites for particular digs

- University publications
• Web-born working papers and other resources from faculty and students
• The range of an institution's scholarly products that reflect new and emerging publishing methods and vetting mechanisms

• Research institutes, associations, and organizations
  • Information about regions or towns
  • Foreign, regional and local information
  • Biographies

• Websites of other universities
  • Content related to curriculum subject areas

• Key events
  • Key events that emerge amid lots of media attention and related ‘grassroots’ sources
  • Example: Bush/Gore vote count in Florida

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 responses echoed and provided a summary of the discussion itself. These results are listed below.

User Needs

Participants indicated that by far the most important user need a web archive would address in their libraries was preservation and persistent access to digital or web-born materials for research and reference. A few participants indicated that an archive would provide value-added information management services to users and one participant mentioned the concept of an institutional repository for maintaining faculty projects. These three user needs are listed below in order of perceived importance:

1. Persistent access to digital or web-born materials for research and reference
   a. Contemporary social movements
   b. Information and materials characterized as fleeting, ephemeral, non-standard, not previously published, or not commercially viable
   c. Content consistent with collection parameters of the institution

2. Provision of value-added services
   a. Improved discovery
   b. Structured, well-organized information sources in subject concentrations

3. Access to an institutional or organizational repository
   a. Repository of faculty projects

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 three areas in item two were of roughly equal importance.)
1. Technology (infrastructure, standard discovery tools, technical expertise)

2. (a) Organizational support (university administration, faculty, IT)
(b) Resources (funding, staff, hardware)
(c) Policies related to web materials (what to archive, cataloging)

3. User needs (feedback mechanisms, ongoing discussion between librarians, faculty, & students)

**Biggest Hurdles**

Participants identified policy setting (establishing priorities and identifying what to archive) and staff responsibilities (integrating web archiving practices into collection development activities) as the two biggest hurdles their organization faced in creating a web archive. Technical support and collaboration between IT and curators were also mentioned as a hurdle.

**4.4 Need for Collaboration**

*"Why should everybody try to do the impossible?"*

**Libraries**

Collaborations and consortia have a central role in the culture of libraries. In this spirit, libraries at research institutions have discussed a collaborative archiving model in which different libraries assume responsibilities for establishing and maintaining collections in specific information areas and agree to make these collections accessible to other libraries. In New York, there is an example of this in the business arena: Different libraries have each agreed to create, maintain, grow, and make web-accessible collections pertaining to a certain area of business information.

LOCKSS is an example of collaboration among universities to archive and provide access to journals. Universities agree to capture and make available for replication certain journals that might otherwise be lost over time. “The whole LOCKSS concept seems to make sense in terms of not duplicating effort.”

**Libraries & Other Organizations**

Given the volume of web-published materials, collaborations among libraries and other organizations might be an effective way to provide increased coverage of a particular subject area. One example of this is the informal collaboration between one library/archive with a “community of creators”. The library collects and preserves materials created by this community thereby expanding its own collection. This collaboration existed for print materials and is being extended to web-published materials.
Appendix A. Collection Development for Web Archives

<table>
<thead>
<tr>
<th>POLICY SETTING</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SELECTION</strong></td>
<td>Policy factors influencing web archiving include political mandates, organizational mission, financial parameters, and technical capabilities.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SELECTION</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection</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>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>

<table>
<thead>
<tr>
<th>CURATION</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>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.</td>
</tr>
<tr>
<td>Organization</td>
<td>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.</td>
</tr>
<tr>
<td>Presentation</td>
<td>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.</td>
</tr>
<tr>
<td>Maintenance</td>
<td>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.</td>
</tr>
<tr>
<td>Deselection</td>
<td>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.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PRESERVATION</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

Angela Carreno
Collections Coordinator, Collections & Research Services
Social Science Bibliographer
Subject Specialist: Latin America

Donna Davey
Tamiment Librarian

Alicia Estes
Head, Business & Government Documents Center
Subject Specialist: Business, Hospitality, & Tourism

Paula Feid
Undergraduate Librarian, Undergraduate Services

Peter Filardo
Tamiment Archivist

W. Gerald Heverly
Reference Librarian, Humanities & Social Sciences Services
Subject Specialist: Philosophy & Classics

Jason Phillips
Reference Associate, Humanities & Social Sciences Services
Subject Specialist: Sociology, American Studies, Gender & Sexuality Studies

Jennifer Vinopal
Reference Librarian, Humanities & Social Sciences Services
Subject Specialist: French & Italian Language & Literature
Interim Manager, Digital Library Program
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!