Digital PRESERVATION
NATIONAL DIGITAL INFORMATION INFRASTRUCTURE AND PRESERVATION PROGRAM

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

Content Identification, Selection, and Acquisition Path

End User Interviews:

Summary Report

April 17, 2006

Prepared by:

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

The following people contributed to this document.

New York University	Michael Nash
University of California - Santa Barbara	Janet Martorana & Sherry DeDecker
University of North Texas	Inga Hsieh

Contents

1 Introduction

The Web-at-Risk project is one of eight digital preservation projects funded in 2004 by the Library of Congress. The project is a 3-year collaborative effort of the California Digital Library (CDL), the University of North Texas (UNT), and New York University (NYU). The project will develop a Web Archiving Service that enables curators to build, store, and manage collections of web-published materials. The content will be collected largely from US federal and state government agencies, but will also include political policy documents, campaign literature, and information concerning political movements and labor unions.

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

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

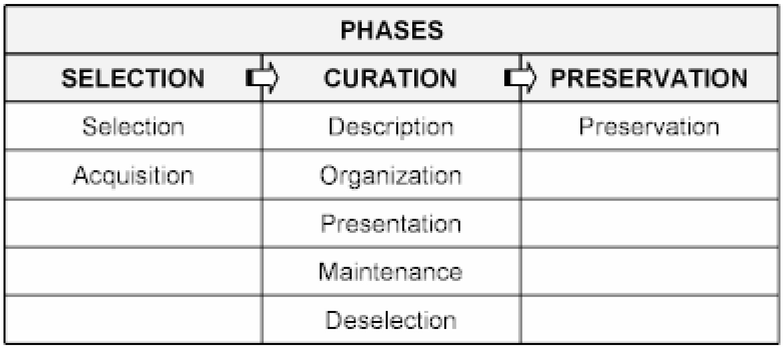
2 Methodology

2.1 Framework

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

Collection development for web archives includes three major phases: selection, curation, and preservation. By breaking down collection development into a series of activities within each phase, the functional view shown in Table 1 emerges. (Appendix A provides a brief explanation of the activities in each phase as they apply to collection development for web archives.) It was expected that potential users of web archives would offer important insights and requirements that could inform these activities.

Table 1. Collection Development Framework for Web Archives



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

2.2 Participants

Project team members at each of the project partner institutions interviewed researchers in the disciplines of history, political science, or law working at their respective institutions or archives. In all, seven interviews were conducted: four with historians, two with political scientists, and one with a professor of hospitality law and management. Table 2 identifies the participants' areas of research and Appendix B lists the individuals.

Table 2. Participants

Project Partner	Title & Discipline
NYU	Assistant Professor - History of American Business & Labor Associate Professor - African America & Labor History Professor - History of the Jewish American Left
CDL - UC Santa Barbara	Doctoral Candidate (ABD) - Political Science
UNT	Assistant Professor - 20th Century American History Assistant Professor - Political Science Professor - Hospitality Law & Management

With the exception of one person whose use of web-published resources was limited to archival finding aids for research, all of the participants used web-published materials in both their research and professional activities, although the extent of their usage varied widely. For some, web-published materials were more likely to be used in their teaching and for others, in their research or professional activities.

Some researchers make extensive use of specific web-published databases in their research (e.g., statistical databases from the US Department of Labor or the University of Virginia's US Census Browser). Some historians value web-published oral history interviews and transcripts (e.g., the American Memory and Thomas collections at the Library of Congress) as well as the archives of major newspaper's (e.g., the New York Times or the Chicago Tribune). Political scientists are more apt to use data sources related to specific research areas (e.g., state congressional committee membership lists or public opinion survey data about presidential candidates). These latter resources appear to be more vulnerable to being lost or replaced.

> *"State legislatures don't usually archive their own materials from the Web. They just replace last session's materials in favor of this session's. You can't get at committee assignments from 1999 to 2004."*

*****

> *"I had an instance where a paper I published on a web journal went out into the ether somewhere and I've never been able to find it again."*

Many had experiences with web-published materials being lost or with referenced hyperlinks failing to link to expected material locations. However for most the loss was not 'critical' to their teaching or research. The lost materials were characterized as causing "trouble" or as an "inconvenience". A few researchers are quite successful locating 'lost' materials in the Internet Archive [http://www.archive.org] and some ensure that critical or important web-published information is saved either in a personal 'archive' or in print format.

> *"If there's something on the Internet that's critically important to my research, I capture it. . . . One site, not affiliated with a university or an academic web site, updated data from the Kerry-Bush election -- all the latest poll numbers. Right around the day of the election, I saved as much of the data as possible in case the web site disappears."*

2.3 Data Collection & Analysis

The interviews were conducted by project team members, who used the interview questionnaire in Appendix C to guide the discussion. Five topics were discussed:

1. Selection of Materials for an Archive
2. Authenticity of Archived Materials
3. Interacting with Materials in an Archive
4. Searching an Archive
5. Preservation of Archived Materials

Each topic provided information related to one or more of the web collection development activities listed in Section 2.1.

Interviewers summarized the discussions and identified the key points that emerged. The summaries were provided to the project's Assessment Analyst, who further analyzed the content. Three questions (1, 3, and 15) asked participants to select values that best matched their opinions. For each of the three questions, weighted sums were calculated to rank responses. (See Appendix D for detailed responses to these questions.) The themes and issues that emerged in the discussions are reported in section 3 of this report.

3 Findings

3.1 Selection of Materials for an Archive

<u>Importance of Information Sources</u>

With the exception of web-published books, at least one of the participants rated each of the information sources in Table 3 as "high" in importance for their discipline, for either research or professional information. Across all disciplines, the most important web-published information sources were journals and periodicals, databases, government information, newspapers, and the proceedings from professional meetings.

> *"More digitization of newspapers would be a great thing for the profession moving forward. . . . That would be a big help to historians - not just the Times but regional powerhouses and the local newspapers too."*

An important information source identified by several participants, but not included in Table 3, was organizational web sites. These included the web sites of trade unions, in particular local union web sites, and discipline-specific organizational web sites, for example, the American Hotel and Lodging Association. These web sites contain valuable newsletters, articles, brochures, and links to other information sources. Additionally, it would be of value to historians if information sources from print archives were digitized and published on the Web. These source materials include manuscripts, posters, pamphlets, and photographs.

One participant researches committee memberships of state legislatures. This information is both printed quarterly by a commercial publisher and published on state legislative web sites. Often libraries do not retain back copies of the quarterly publication, which is not readily available from the publisher except at a cost. While the Wayback Machine has been an excellent source of data for this researcher, he has been frustrated by some state legislative web sites that include a robots.txt file to disallow information capture. He suggested that it would be of great help if data from all states were resident on an archive.

Table 3. Importance of Web-Published Information Sources

Rank	Information Source
1	Journals & Periodicals
1	Databases
2	Government Records or Documents
3	Newspapers
4	Proceedings of Meetings & Symposia
5	Doctoral Dissertations & Master's Theses
5	Brochures
5	Technical & Research Reports
6	Unpublished Work & Publications of Limited Circulation
7	Videos
8	Audio files
9	Books

Retention of Archived Materials

> *"If they're being collected [in an archive], I can't imagine it would not be permanent!"*

Participants generally viewed archived materials as collections that were retained "forever". This opinion was particularly expressed in regard to four of the five information sources they rated as most important to their research and professional activities: journals and periodicals, government information, databases, and newspapers.

Table 4. Relative Retention Time for Archived Materials

Rank	Information Source
1	Journals & Periodicals
1	Government Records or Documents
2	Databases
3	Newspapers
4	Videos
5	Audio files
5	Books, Brochures
6	Proceedings of Meetings & Symposia
6	Doctoral Dissertations & Master's Theses
6	Unpublished Work & Publications of Limited Circulation
7	Technical & Research Reports

> *"The moment I say 10 years back [is long enough to retain journal articles] I'd need one from 12 years back. You never know when you're going to need one of those."*

Table 4 ranks information sources according to participants' opinions of each source's relative value over time, from 1 year to more than 10 years. While most participants thought each source provided value for 10 or more than 10 years after publication, a few selected fewer years for three of the sources. Specifically, some participants did not think that either proceedings of professional meetings or unpublished works needed to be retained beyond three years. One person thought technical and research reports only needed to be retained for five years. (It is worth noting that technical and research reports were not a common information source in most participants' disciplines.)

3.2 Authenticity of Archived Materials

<u>Major Concerns</u>

In terms of citing archived materials in their research and scholarship, two primary concerns emerged:

1. Trust in the authenticity of archived materials and

2. General acceptance among colleagues of such citations as 'authentic sources'.

Participants were well-aware that web-published materials can be altered. Several expressed concern regarding whether they could trust that the original source materials had not been altered or manipulated in some way. A few had no problem citing certain trusted web sources, although one described their trust as always involving "a leap of faith".

> *"One problem is that it's essentially a copy and also it's a copy that could be manipulated. So, I'd want some security provision or some way to know that nobody in the archive or hacking in from the outside had changed the content of the web sites."*

While one participant couldn't imagine any problem with citing a web archive as the source for the data used in their research, a few thought this was not yet generally accepted within their area of research or discipline. They were concerned that others might question their scholarship. When in doubt and for critical sources, some researchers cite print sources in favor of web sources.

<u>Creating Trust</u>

A common criterion for trusting the authenticity of materials in an archive is the reputation of the publisher, such as the New York Times, or the archive provider, such as the Library of Congress. In general, the attribution of "trusted source" was ascribed by participants to the federal government and to universities.

> *"I would want the archive from an institution that I have faith and confidence in: if it's done in the university [or] the federal government that would satisfy me."*
>
> *****
>
> *"I put a lot of trust in the track record of the institution. If it was a Library of Congress sponsored project or a University of California sponsored project, I'm sure I would have complete faith in it. But if it was Joe's History web site I would have a whole lot less faith."*
>
> *****
>
> *"Realistically, I would most likely trust the imprimatur of the library."*

Another common criterion for establishing trust in archives of web-published and digital materials was the fact that the original physical source material or artifact existed somewhere and could be viewed if desired. In this regard, it would helpful if web archives included references to the locations of original sources.

> *"I want to put the official bibliographic reference on all the documents on my site [and] refer to where you can find this document, in which volume. That's important. That refers to where I can find it in print."*

*****

> *"There has to be some way of having access to the original. I wouldn't be comfortable with anything else."*

While there appeared to be a common understanding that ultimately it was up to the individual researcher to evaluate the authenticity of their information sources, there was a strong desire expressed by some participants that some "authority" create a standard verification process that could be used to establish the authenticity of web archives. A 'verification stamp' from such an authority would ensure that archived materials were authentic, thus making citations of these materials more trustworthy.

> *"I've been thinking -- add a symbol or icon on docs that says 'We have double-checked this against the original source' -- like on Ebay -- 'ID Verified' -- 'Authenticity verified by a human being.'"*

<u>Dealing with Modifications</u>

> *"I can imagine these hypothetical [situations] where authenticity could be called into question, where an archivist is making editorial judgments about how the material is being used."*

Most participants thought removal of parts of a web page by an archive provider would compromise its authenticity. The major concern with selective removal of parts of web pages was for the potential loss of important contextual information. As one participant stated, such alterations could potentially "limit one's understanding of the subject". The information context of web pages was highly valued by many participants, some of whom likened it to the importance of context for analysis of print materials.

> *"I think placement is significant in the same way placement of an article in a newspaper is important to know when you're analyzing that article. The way in which a piece -- a document -- is situated on a web site is relevant."*

*****

> *" . . . the web site seems like a single 'document' unto itself. So taking parts of it out seems problematic. . . . I guess it seems more like blocking out certain parts of a letter someone writes. Obviously, there are large parts of it that are not that interesting to most people but they still reflect something about the document as a whole."*

A few participants identified circumstances in which selective removal of parts of web pages might be understandable: downsizing very large image formats, removal in accord with policies of a trusted provider (e.g., potentially offensive material), and copyright infringement. Consistency in editorial policy on the part of the archive provider generally appears to increase trust in the authenticity of the archive's web sites. One participant offered this solution for embedded video content that an archive might remove from a web page:

> *"You might not be able to preserve an embedded video but you probably could archive that video separately and have some indication that this was embedded."*

Echoing this suggestion, participants generally endorsed archive practices that would alert visitors to changes made to the original source materials. One participant noted this "needs to be done and would mitigate the problem with authenticity." Participants thought archives should tag web pages to indicate changes and should provide documentation explaining modifications. One participant wanted such documentation to include assessments of the impact of material format changes. Another participant wanted 'maps' of original web sites to be provided and thought this might provide sufficient context for analysis of sites from which some content was removed.

3.3 Interacting with Materials in an Archive

Mirroring Web Sites

Participants expressed a variety of expectations regarding replicating or mirroring web sites' interactivity within an archive. A few thought the archive should definitely mirror the interactivity of original web sites. A few thought this would be "convenient" but they neither expected this functionality nor always required it when interacting with archives. Others were more interested in specific information content (e.g., images or transcripts) from web sites and were less concerned with replication of the sites' original interactivity. One participant thought it was important for an archive to make it absolutely clear that visitors were not interacting with the original, live web sites. Another participant appreciated the difficulties of mirroring the functionality embedded in older web sites and summed up his expectations of a web archive this way:

> *"I would either extract just the docs or data you need and toss out the navigation structure or completely duplicate the web site the way it was. . . . including the database, the navigation . . ."*

Handling Active Elements & Dynamic Content

Email Links

A few thought the 'ephemeral' nature of email addresses made disabling them in an archive understandable and, in one person's opinion, desirable. At the other end of the spectrum, one person thought an archive should attempt to keep the links operational by validating email addresses and replacing invalid addresses when possible. Some participants thought it was of research interest to be able to identify the original email addresses (i.e., the 'mailto:' targets) in source web pages.

> *" . . . if making it [the archived web site] authentic for the experience,[a] link to email would come with [a] pop-up saying it no longer is active but this is what it looked like . . . to preserve the authenticity, the experience."*

Hyperlinks

Most thought hyperlinks should be preserved in an archive, although there was not general agreement regarding to what extent the content of linked materials should be included in an archive. A few thought as much as possible of the linked content should be preserved in an archive while most thought only critical content whose absence might misrepresent the meaning or value of the source web site should be included. Most agreed that even if hyperlinks were disabled or no longer valid, the targeted addresses of the original hyperlinks are of research interest and should be preserved and made accessible to visitors. A few participants thought visitors would be tolerant of non-functional links in an archive, or at least more tolerant than they would be in live web sites.

> *"The AFL web site has a huge system of links, grouped by different unions, different labor organizations, and different activist organizations -- political organizations. I think that knowing who the AFL-CIO selected to link to is important even if you don't have the ability to go to [those] web site[s]."*

*****

> *"It might be enough to see that the link was there. You don't actually need to keep the link active."*

*****

> *"I think it would be interesting to preserve those links. . . . I think that the web sites are interesting documents from the standpoint of how the union is conceiving of itself and trying to communicate to its members through this new medium."*

Forms & Programmatically-Generated Web Pages

In order to offer the most "faithful representation of web pages" whose forms were no longer operational in an archive, a few participants thought it would be good practice to provide screen shots of the original forms.

> *"Maybe take a screen shot of the form so that when people get to the form, they can see what it looked like -- can't be filled in but this is what it looked like."*

*****

> *"In the case where [the archive] could not mirror the original site exactly, maybe you have just a static sort of screen shot of what the original search page looked like and explain here's why it doesn't look like this any more and doesn't have the same functionality."*

A few participants thought that any code or script-based functionality that could be replicated in an archive should be replicated. Only two participants had experience with the complexity of replicating programmatically-generated web pages in an archive and both expressed frustration when dealing with archives that failed in their attempts to do so. One participant indicated that in the absence of the original server-side code and scripts (e.g., Perl and PHP), it is not possible to replicate the functionality in an archive.

Customized Web Pages & Privacy

A few participants thought retaining the functionality to generate customized web pages would be desirable if possible. One person was concerned about privacy and confidentiality and thought this functionality should be disabled. One participant suggested that in the absence of access to both the cookies stored on visitors' personal computers and the back-end databases from the content provider, as well as the server-side code and scripts required to generate the customized web pages, it would be unlikely that this functionality would be replicated in an archive.

Some privacy concerns were expressed in regard to archives storing personal information that visitors might have previously submitted. A few participants suggested web archives adopt existing archival guidelines regarding the suppression of personally identifiable information for some period of time. One historian saw value in obtaining and retaining information about visitors to original web sites when the sites were archived:

> *"I think the hardest thing to deal with as an historian with any publication is knowing about the readers, knowing how widely this [publication] was disseminated, who accessed it. So I think anything that you can -- any information that you can archive about people who access these web sites is valuable. . . . There's clearly an ethical issue -- clearly a legal issue -- which is not my expertise."*

3.4 Searching an Archive

Search Criteria

Participants agreed that 'topic or subject' and 'full-text using any keyword' are the two most important search criteria they would use to locate materials in an archive. Table 5 lists the order of importance for other search criteria. All of the criteria ranked as 1, 2, or 3 were rated high in importance by most of the participants. A few participants rated 'organizational name' and 'material format' as high in importance. 'Project name' was rated low in importance by most participants. Alternatively, a few participants added browsing the archive via a subject directory structure as a desirable interface for locating materials.

Table 5. Relative Importance of Search Criteria

Rank	Search Criteria
1	Topic or subject
1	Full-text using any keyword
2	Author
3	Title
3	Original URL
3	Publication date
4	Organizational name
4	Format
5	Project name

Search Results

In terms of search results showing all of the versions (or captures) of a web site in an archive, one person thought that could be "overwhelming" while another person thought it was an "extremely important" feature. Effectively analyzing search results is related to how the results are structured and to what information they include. Some participants suggested models that were successful for them, for example, the search results from Google or the Wayback Machine. Others named specific attributes of materials they might use to evaluate results. (See Table 6.)

Of the attributes or evaluation criteria suggested by participants, neither Google nor the Wayback Machine includes creation date, organization, or author in their search results. Google does include additional hyperlinks to cached web pages and a follow-on search capability.

Table 6. Evaluation Criteria

Attribute	**Google**	**Wayback Machine**
	Result from Simple Keyword Search	*Result from URL Search*
Title of the web page	Page Title (e.g., from HTML <title> tag) - hyperlinked to the web site	
Original URL	URL	
Description	Brief text extracted from web page	
Date created		
Date archived		Dates for each capture of a URL - hyperlinked to the web site
Organization		
Author		
	Non-Descriptive Information	*Presentation of Results*
	Hyperlink to cached web page	Table with columns for each year: 1996 - 2006
	Hyperlink to conduct a follow-on search and retrieve similar pages	Number of Pages - i.e., Total number of captures (or instances) in the archive for a URL by year
		Date for each page - Month, day, year of each capture

Multiple Versions

Most participants envisioned one summary for each web site resulting from a search. Multiple instances or versions of a web site would be included in the summary result. One participant described this as a “serialized” description that would include one description for each web site with each version or instance listed along with its capture date and the ‘trigger’ or rationale that prompted each capture. Another participant would “eventually like to have separate records [for each capture] but as a front page to them one record would be fine.” The result of a URL search of the Wayback Machine does identify each instance of a capture; however there is currently no descriptive information provided for the web site.

Multiple Material Types

For materials captured in multiple types (e.g., text, audio, and video), most participants wanted a single summary search result that described the available types. It was important to some participants that they be able to select which type they accessed. One participant suggested visitors be linked to text by default but be presented with icons for easy access to other material types. This would address one participant’s concern that he “wouldn’t accidentally try to open a 99 MB file.”

One participant suggested that if specific data contained in a captured web site is extracted and also retained in an archive, then there should be a separate summary (or descriptive record) for the data. However, he cautioned that it was “too hard to combine both of these types in one archive” or if they were both included in an archive “there should be clear lines demarcating” each type (i.e., the web site as one material type and the extracted data as another type).

3.5 Preservation of Archived Materials

Retaining Multiple Material Types

Even when considering cost factors, such as the hardware and software required for presentation and storage of materials in an archive, all of the participants thought that it was either ‘very important’ or ‘extremely important’ to retain multiple types of original materials (e.g., a video file, an audio file, and a text file of the same content). Even if the quality of the material was poor, most felt it should be retained because “whatever you keep is better than nothing”.

The significance and implications of not retaining multiple types of an item in an archive, or an item contained in a web site, appears to be specific to a researcher’s area of study. For most researchers, the implications are more substantial depending on what material type is retained. In particular if only text is retained, participants thought information of interest to different researchers would be lost. For example, linguists researching variations in speech and psychologists studying non-verbal behavior would be thwarted if only transcripts were available.

> *“I like to have my own personal archives of the speeches of the president. When he walks off the podium, shakes hand, there’s something to that. There’s data there. Whose hand does he shake? Do they stand? How long do they clap? Is there a kiss on the cheek, a whispering of something to whom? That’s what you lose when you don’t keep the original format.”*

Another concern is that information can be “lost” or altered information can be introduced when the source material is recreated as a different material type, for example, when a transcript of an audio file is created. This concern was echoed by a participant regarding copying source data; retaining source data is not critical as long as no data is lost in the copy process.

> *“There’s always the danger that you lose information of some sort. As an oral historian that’s something I think about. A lot of oral historians should be ultimately concerned with the transcript but the transcript is ‘garbage in garbage out’. So if you mess up the first version of a transcript no matter how many times you improve it -- if you’re misquoting the original -- you’re just repeating the same mistake.”*

While recognizing the value and subtleties in different material types, for example text, video, and audio, one participant thought text materials should be kept if a retention decision among the various material types had to be made. Another participant stated his preference for databases to be retained as flat files.

Factors to Consider in Removing Archived Materials

Frequency of Access

This factor has two facets: low usage and high usage. While most participants recognized that low usage might well influence decisions about which materials to remove from an archive, they did not think low usage should be the sole factor. As one person stated, “I don’t know if use should be the decisive factor because that changes over time.” Another person expressed a similar view this way:

> *“In the perfect world you’d say: ‘While even though no one has tried to look through this recording in the last 20 years, it doesn’t mean that 20 years from now someone [won’t] absolutely have to hear it.”*

*****

> *“Some of the best historical work is based on unusual discoveries of sources that people don’t usually use.”*

One participant recognized that high usage could also be a factor in managing an archive, in particular the bandwidth required for multiple transmissions of large files. To address this issue, he suggested that large image files might be converted to smaller files -- as long as their authenticity is not compromised.

Multiple Formats

One person suggested that in some cases different formats of the same material (e.g., a video formatted for viewing with different plug-ins) was for the convenience of users and was not important for either archival or research purposes. He suggested the decision to retain multiple formats in an archive should rest with the archive provider. If forced to remove materials, another participant thought it might be good practice to retain the original format and the most recent format.

Multiple Versions

In general, participants thought that the existence of multiple versions of a web site was not in and of itself a factor to consider when making deselection decisions. Some participants thought the merit of multiple versions could be gauged by the importance of their contribution to the body of knowledge about a particular subject.

> *"You're archiving the web pages of the [Dean] campaign -- an absolutely critical part of the campaign -- you'll want to have a daily shot of the dot com or whatever. But if you're talking about just (multiple versions) for the Organization of American Historians web site -- probably doesn't change that often. Even if it does [where's the significance in] what those differences might be? In the Dean campaign that [web site] was such a vital part of the campaign."*

A few respondents addressed this factor from the perspective of capturing web sites: How often should a web site be captured? What events should trigger a capture? For example, quarterly captures of a union web site might generally suffice for an archival record. However, the record would be incomplete without capturing web sites during certain events, such as a strike, contract negotiation, or internal election.

Length of Time in the Archive

While one participant suggested that removal of materials based on their time in the archive might be related to what the materials were, most agreed this should not be a factor for removing materials from an archive. In fact, 'change over time' is precisely the concern of many scholarly researchers and removing materials based on this factor might be a disservice to research in general.

Other Factors

> *"It seems to me an archivist has to consider other things, which is to anticipate the future historical value, to use their own subjective sense of what's interesting and that is purely subjective."*

One participant suggested the following factors should influence decisions to retain or remove archived materials:

- Something that seems extraordinary (e.g., an unusual event)
- An unusual kind of record (e.g., an expensive autobiography or a rare diary)
- The source (e.g., a person of importance at one time)
- Something likely to generate interest (e.g., a great unpublished collection of cartoons)

4 Discussion

Web Content v. Web Sites

Many of the materials these scholars use in support of their research, teaching, and other professional activities are from reputable web-published collections: newspaper articles, historical photographs, national labor statistics, and national census data. Large organizations, such as agencies of the federal government like the Department of Labor, the Library of Congress, and the National Archives and Records Administration, newspapers such as the New York Times and the Chicago Tribune, and universities such as the University of Virginia, as well as university collaborations such as the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan, publish and provide access to these important collections. In interacting with this class of web site, scholars are dealing largely with the content of web sites and not with the web site as an object of study unto itself.

This is largely true of scholars who utilize content from web sites of smaller organizations and state government agencies, such as political scientists concerned with opinion poll research results or committee memberships of state legislatures. These researchers capture specific content from web sites for their scholarly research interests and are not studying web sites per se. They appear to be more experienced in dealing with the frustrations of web-published information for which the publishing organization has not taken preservation responsibility, for example the susceptibility of their source web sites to disappearance.

For researchers using web-published content in their scholarly research, long-term preservation of the content is of great importance. The overall context of the web site containing the content is of lesser importance. Their needs can be met by archiving the content without compromise to its integrity and with a capability for researchers to locate materials of interest within the archive.

Some researchers investigate the content of web sites, including analysis of the overall context in which content is placed on a web site and on a web page. For these researchers, as well as those concerned with web sites as the objects under investigation, long-term preservation of the web site as a whole is of great importance.

Transitional Times: The Web Site as 'Document'

Citations to web sources in scholarly research are not always respected and some researchers have yet to use web-published resources in their research publications. However, while most of the scholars interviewed in this project are not currently studying web sites as source materials unto themselves, in the tradition of correspondence or newspapers as source materials, they are quite sensitive to the growing importance of web sites as cultural artifacts that will inform future research. They recognize the central importance of web sites in elections, as exemplified by the Howard Dean campaign, and in organizations, such as international labor unions and their local affiliates. And they are asking what it means from a scholarly research perspective to treat web sites as original source 'documents'. It is from this perspective that concerns arise regarding the authenticity of web page content and preservation of web page 'context'.

Authenticity

The Web raises concerns about the authenticity of source materials for many researchers, although most use web-published materials from federal government sources without question. A common criterion for trusting the authenticity of materials in an archive is the reputation of the archive provider. In general, the attribution of "trusted source" was ascribed by participants to the federal government and to universities.

Beyond this, archives can establish and promote trust in their materials by providing bibliographic references to the locations of original source materials and by documenting the provenance of born-digital materials, including format migrations over time. Further, it would be helpful to establish an independent practice standard for web archives and a certification authority. Such an independent authority could certify a web archive's conformance to standard practices thereby providing researchers with some assurance of the authenticity of archived materials.

Content v. Context

Participants discussed replicating the experience or 'look-and-feel' of source web sites, including active web page elements (email links, hyperlinks, and forms) and methods for programmatically generating web pages (individual customization and database generation). They offered a range of opinions and suggestions regarding how this functionality should be handled in an archive. In general, participants seemed to grasp that an archive might not be able to functionally replicate the interactivity of source web sites, although this insight came as a surprise to some participants.

The extent to which original web site interactivity, or the 'ghost' of it (e.g., non-functional email links and hyperlinks with identifiable addresses), was retained in an archive related to the purpose for which scholars' use archived materials. For some, specific content is more important than its context and placement on a web site. For others, context and placement are critical; a web site is viewed as a single entity in the same way a printed document is a single entity and content placement has relevance in the same manner that content placement in a printed newspaper does. Researchers studying organizations or movements or elections as reflected in their web sites find importance in the web site in its entirety and require the most faithful replication of web sites in archives as well as documentation of all modifications.

Curation and the Archive Provider

The participants in this set of interviews articulated a wish list of 'requirements' for presentation of the materials in a web archive and for searching an archive. Satisfying these requirements will enable scholars to cite archives as source materials in their research and will promote trust in archived materials as authentic sources both for scholarly research, teaching, and other professional work. To meet these requirements, it appears that curation of materials in a web archive will require a set of tools for modifying, tagging, and describing the materials.

Presentation

- Reference the locations of original source materials
- Make the addresses of no longer functioning mail links and hyperlinks visible/accessible
- Provide site maps for web sites
- Provide screen shots of non-operational forms with explanations of why the forms do not work

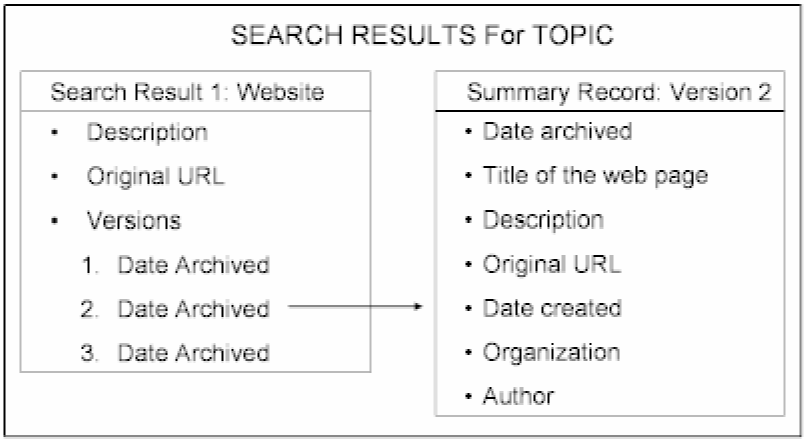
Searching

- Topic or subject
- Full-text

Search Results

Researchers would prefer a single summary record identifying all versions of a particular item as well as a summary record for each version that would include: (See Figure 1.)

- Title of the web page
- Original URL
- Description
- Date created
- Date archived
- Organization
- Author



SEARCH RESULTS For TOPIC

Search Result 1: Website

- Description
- Original URL
- Versions
  1. Date Archived
  2. Date Archived
  3. Date Archived

Summary Record: Version 2

- Date archived
- Title of the web page
- Description
- Original URL
- Date created
- Organization
- Author

Figure 1. Illustration of the Relationship between Search Result Summary Records

Web archive providers understand the distinction between capturing web sites and replicating the sites' original interactivity. Tools and technologies for the former are operational today while the latter poses many challenges. Content providers must determine what materials on their web sites, including server-side code and back-end databases, they want to entrust to an archive. Copyright permissions, access limitations, and maintenance requirements are among the details that must be negotiated between content providers and archive providers.

Material Deselection

This group of participants, reflecting the wider scholarly research community, expressed a good deal of caution regarding removing materials from a web archive. It is never clear which materials will be of research interest and value in the future. Participants generally viewed archived materials as collections that were retained "forever".

*****

"I say what makes a good archivist, like a historian, is to have a sense of something being important and special even if it's not immediately obvious."

Appendix A. Collection Development for Web Archives

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

Appendix B. Participants

Jim Battista, Ph.D.
Assistant Professor - Political Science
University of North Texas

Joan M. Clay, Ph.D.
Professor - Hospitality Management (Business & Law)
University of North Texas

William Jones, Ph.D.
Associate Professor - History (African American & Labor History)
University of Wisconsin - Madison
Scholar in Residence - Schomburg Center for Research in Black Culture, New York Public Library

Tony Michels, Ph.D.
Associate Professor - History (History of the Jewish American left)
University of Wisconsin
Scholar in Residence - Tamiment Library & the Goldstein-Goren Center for American Jewish History at NYU (2005-2006)

Todd Moye, Ph.D.
Assistant Professor - History (20th century American History)
University of North Texas

Gerhard Peters, Doctoral Candidate - ABD
Graduate Student - Political Science
UC Santa Barbara

Kimberly Philips-Fein, Ph.D.
Assistant Professor - History (History of American Business & Labor)
NYU Gallatin School for Individualized Study

Appendix C. End User Interview Questionnaire

Background Information

- What is the name of your department?
- What is your current position, academic status, or title?
- How many years have you been in this position?
- Do you use web-published materials in your:

  a. Research activities? ______ Yes ______ No
  b. Professional activities? ______ Yes ______ No

- Have you ever tried to retrieve a critical document or a file from the Web that was no longer there? How often has this happened?
- Think about one of those incidents and describe the circumstances? How severe was the loss?

Topic 1. Selection of Materials for a Digital Archive

1. Which of the following information sources in your discipline are accessible on the web? How important are these web-published information sources in your discipline, for either research or professional information?

		Importance		
	Web Accessible?	High	Medium	Low
Journals & Periodicals				
Books, Brochures				
Databases				
Newspapers				
Videos				
Audio files				
Technical & Research Reports				
Proceedings of Meetings & Symposia				
Doctoral Dissertations & Master’s Theses				
Government Records or Documents				
Unpublished Work & Publications of Limited Circulation				

2. Are there other web-accessible information sources that are important in your discipline?

3. For each type of web-accessible information in your discipline, how long does it provide significant value after publication?

	# Years				
	< 1	1-3	5	10	>10
Journals & Periodicals					
Books, Brochures					
Databases					
Newspapers					
Videos					
Audio files					
Technical & Research Reports					
Proceedings of Meetings & Symposia					
Doctoral Dissertations & Master’s Theses					
Government Records or Documents					
Unpublished Work & Publications of Limited Circulation					
Other:					

Topic 2. Authenticity of Materials in the Archive

4. Suppose that source materials originally published on the web are no longer available except as ‘digital copies’ in an archive. What issues arise if you cite an archive as the material source?

5. How will you judge the authenticity of materials retrieved from a web archive? For example: “Is this item ‘really’ a transcript of the 1986 US House hearing on gun control?”

6. What is the impact to you of the removal of some parts of a web page from the source material before it is archived, for example, a particular image from a page?

    a. Is the authenticity of the source material compromised?

    b. What if it was removed in accord with university or organizational policy?

7. What do you think about an archive altering or tagging web pages in some way to alert archive users to a modification of the original page?

8. What can the archive do in a web page or a file to build your confidence and trust in the authenticity and credibility of materials?

9. Changes to materials can be expected due to copyright requirements or software migration. Discuss how keeping records of changes and the reason for the changes might help build trust and confidence in archived materials.

Topic 3. Interacting with Materials in the Archive

10. How do you expect to interact with the web archive? Is it important that the archive interaction mirror your experience of the original 'live' website?

11. Some web pages include active elements such as hyperlinks and interactive forms that are no longer active in archived materials. How should the archive handle each of the following previously active elements? How should they be presented to users?

    a. Email Links
    b. Links to Materials Accessible in or from the Archive
    c. Data Collection Forms
    d. Can you think of others?

12. Some web sites store personal information about their visitors in order to customize pages for the user when they visit the site. How do you expect customized web content to be handled in a web archive?

13. Some web pages are generated upon request either programmatically or using information retrieved from a database. How do you expect this dynamic web content to be handled in a web archive?

Topic 4. Searching the Archive

14. What information do you expect to find included in a summary of results from a search of a web archive? For example, what are the minimum attributes you would expect? What additional attributes would be nice to have?

15. How important are each of the following to you in locating and selecting web materials using a search engine?

	Importance		
	High	Medium	Low
Topic or subject			
Title			
Author			
Original URL			
Publication date			
Organizational name			
Project name			
Format			
Full-text using any keyword			
Other: ___________________			
Other: ___________________			

16. Many web-published materials are frequently modified and a web archive may capture different versions of the same source materials over time. Considering your needs, how important would it be for you to locate different instances of the same item harvested at different points in time?

17. Should there be a separate summary for each version of an item in the archive or should multiple versions be listed in a single summary?

When a web archive is created, it is predictable that some objects will be archived in multiple formats or media types. For example, archives of testimony before a commission might include a video file, an audio file, and a text file.

18. How about each format of an item? Should there be a separate record for each format?

Topic 5. Preservation of Archived Materials

19. Considering cost factors such as the hardware and software required for presentation and storage of materials, how important is it to you that a web archive retains multiple formats of original materials, for example, a video file, an audio file, and a text file with the same content.

Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5

What implications do you see if the original formats of an item are not saved? For example, if a video recording of testimony is created and then transcribed to a .pdf file, what are the implications if only the .pdf version is saved? Does it make a difference if the original video recording is of poor quality?

20. What implications can you anticipate for current or future researchers if only certain file types are retained, for example the text but not the audio or video files?

In question 22, material format is different from material type. Examples of material types are documents, images, audio files, or video files. A certain type of material could be formatted using one of many encoding methods. For example a document might be encoded using html or xml standards and an image might be encoded using jpeg or gif standards.

Additionally, multiple versions of a single material type, formatted using the same encoding standard, may exist within an archive. For example, multiple copies of a document encoded in PDF format could be captured over time.

21. Retaining or removing materials from an archive will involve trade-offs related to costs. In making these types of decisions, how much consideration should be given to:

a. Frequency of access in the archive
b. Material format
c. Multiple formats
d. Multiple versions
e. Length of time in the archive
f. Other factor

Appendix D. Responses to Questions 1, 3, and 15

1. Which of the following information sources in your discipline are accessible on the web? How important are these web-published information sources in your discipline, for either research or professional information?

Information Source	# Responses	Importance			Weighted Sum
		High	Medium	Low	
Journals & Periodicals	7	7	0	0	21
Databases	7	7	0	0	21
Government Records or Documents	7	6	0	1	19
Newspapers	6	6	0	0	18
Proceedings of Meetings & Symposia	7	4	1	2	16
Doctoral Dissertations & Master's Theses	5	3	1	1	12
Brochures	4	4	0	0	12
Technical & Research Reports	6	2	2	2	12
Unpublished Work & Publications of Limited Circulation	5	2	1	2	10
Videos	5	1.5	1.5	2	9.5
Audio files	4	1	2	1	8
Books	4	0	2	2	6

Notes:
a. Not all participants responded to each source.
b. Weights: High = 3; Medium = 2; Low = 1

3. For each type of web-accessible information in your discipline, how long does it provide significant value after publication?

Information Source	# Responses	# Years					Weighted Sum
		< 1	1-3	5	10	> 10	
Journals & Periodicals	6	0	0	0	0	6	30
Government Records or Documents	6	0	0	0	0	6	30
Databases	6	0	0	0	1	5	29
Newspapers	5	0	0	0	0	5	25
Videos	5	0	0	0	1	4	24
Audio files	4	0	0	0	0	4	20
Books, Brochures	4	0	0	0	0	4	20
Proceedings of Meetings & Symposia	5	0	2	0	0	3	19
Doctoral Dissertations & Master's Theses	4	0	0	0	1	3	19
Unpublished Work & Publications of Limited Circulation	5	0	2	0	0	3	19
Technical & Research Reports	3	0	0	1	0	2	13

Notes:
a. Not all participants responded to each source.
b. Weights: <1 = 1; 1-3 = 2; 5 = 3; 10 = 4; >10 = 5

15. How important are each of the following to you in locating and selecting web materials using a search engine?

Search Criteria	# Responses	Importance			Weighted Sum
		High	Medium	Low	
Topic or subject	6	6	-	-	18
Full-text using any keyword	6	6	-	-	18
Author	6	5	-	1	16
Title	6	4	1	1	15
Original URL	6	4	1	1	15
Publication date	6	4	1	1	15
Organizational name	6	3	2	1	14
Format	6	2	4	-	14
Project name	6	-	2	4	8

Notes:
a. One participant did not respond to this question.
b. Weights: High = 3; Medium = 2; Low = 1