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

Needs Assessment Toolkit:

Guidelines & Data Collection Tools

May 31, 2005

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.

University of North Texas:	Inga Hsieh, Cathy Hartman, William Moen, Samantha Hastings, Arlene Weible, and Valerie Glenn
California Digital Library:	Patricia Cruse, Tracy Seneca, and Suzanne Samuel
New York University:	Leslie Myrick

Contents

1 Introduction

The Web-at-Risk project is one of eight digital preservation projects funded in 2004 by the Library of Congress. Each of the projects represents a collaborative effort to preserve for future generations born-digital or digitized cultural heritage materials and collections.

The Web-at-Risk 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 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 surrounding political movements. The project work will be conducted along four paths of overlapping activities.

Web-at-Risk Project: Work Paths	
1.	Content Identification, Selection, and Acquisition
2.	Content Harvest and Analysis
3.	Content Ingest, Retention, and Transfer
4.	Partnership Building

One focus of the Content Identification, Selection, and Acquisition (CISA) path is to produce tools and guidelines to assist curators and other information professionals in the development of web archives. This Needs Assessment Toolkit delivers a set of assessment tools along with guidelines for using them. The assessment tools are intended to identify the needs of both content producers and data users, that is, those who publish or supply the materials for web-based collections and those who will access and use those materials. Additionally, some of the assessment tools will identify curators' requirements for the web crawler and its crawl analyzer tool, which will be developed in the project's Content Harvest and Analysis path.

1.1 Web Archive Service

The Web-at-Risk project is concerned with building a Web Archiving Service that will enable curators to build, store, and manage collections of web-published materials in distributed repositories located at the three funded project partner sites. (See Figure 1.) Web-published materials include a wide range of material types from text documents to streaming video to interactive experiences. They are both dynamic and transient and are at risk of disappearing. The Web Archiving Service will provide tools and guidance for curators to create collection plans tailored to their unique environments and to select and manage the web-published materials comprising their collections.

As part of the Web Archiving Service, the Web-at-Risk project will build repositories for the long-term storage and maintenance of the libraries' collections of web-published materials. While some of the materials in the collections may exist in other forms, the goal of the Web Archiving Service is to preserve the web versions for posterity so that users may access the materials long after the web sites where they were originally published cease to exist.

At least seven collections of web-published materials are planned. Some will build on existing collections and some will create new collections of web-published materials. Within the web archives themselves, materials will mirror the originally published web sites. One distinction among the planned archives might include patron access to the web archive, that is, a designation of an archive as either light (open access) or dark (controlled access).

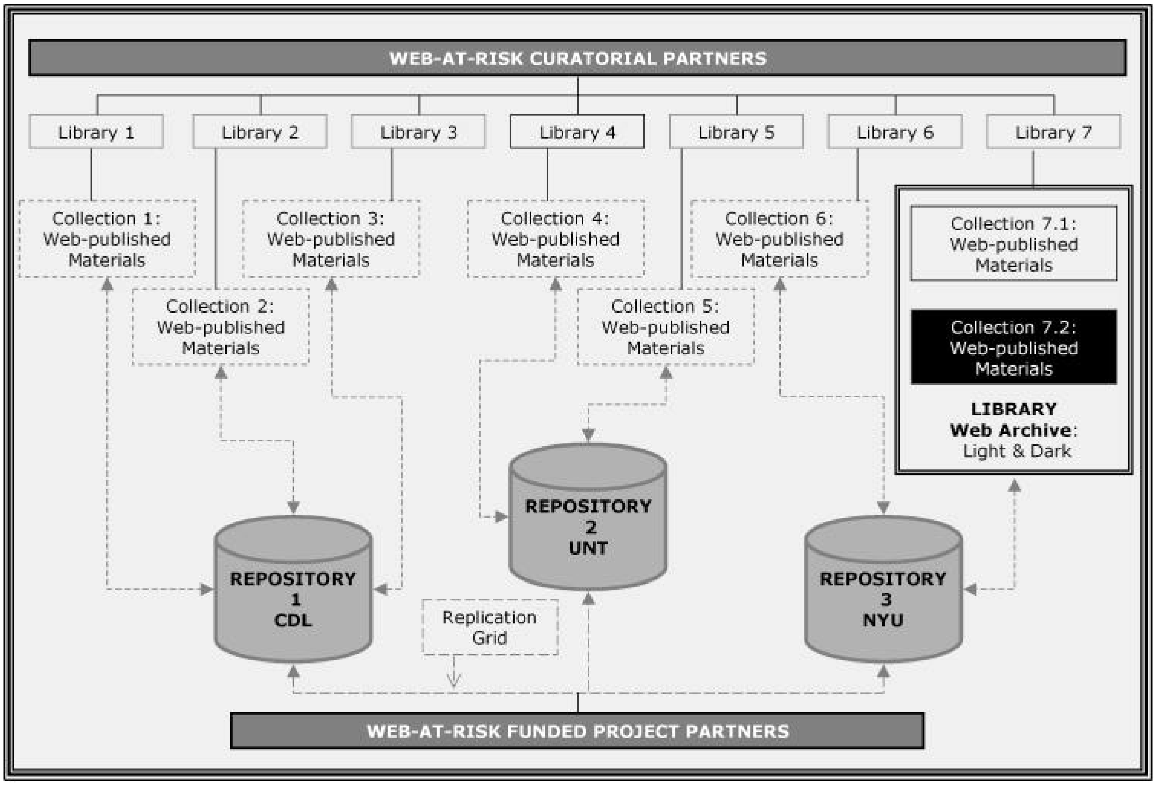
WEB-AT-RISK CURATORIAL PARTNERS
Library 1
Library 2
Library 3
Library 4
Library 5
Library 6
Library 7
Collection 1: Web-published Materials
Collection 3: Web-published Materials
Collection 4: Web-published Materials
Collection 6: Web-published Materials
Collection 7.1: Web-published Materials
Collection 2: Web-published Materials
Collection 5: Web-published Materials
Collection 7.2: Web-published Materials
LIBRARY Web Archive: Light & Dark
REPOSITORY 2 UNT
REPOSITORY 1 CDL
REPOSITORY 3 NYU
Replication Grid
WEB-AT-RISK FUNDED PROJECT PARTNERS

Figure 1. Web Archiving Service: Conceptual View

1.2 CISA Path Context: The Web Archive Development Process

Many concepts and terms related to web archives have different meanings in web archive applications and projects. (A glossary is included in Appendix 1.) It is helpful to establish the context and define the terms as they are used in the CISA path and in this toolkit. The overall context for this toolkit is the library, with a focus on large academic libraries. (See Figure 2.)

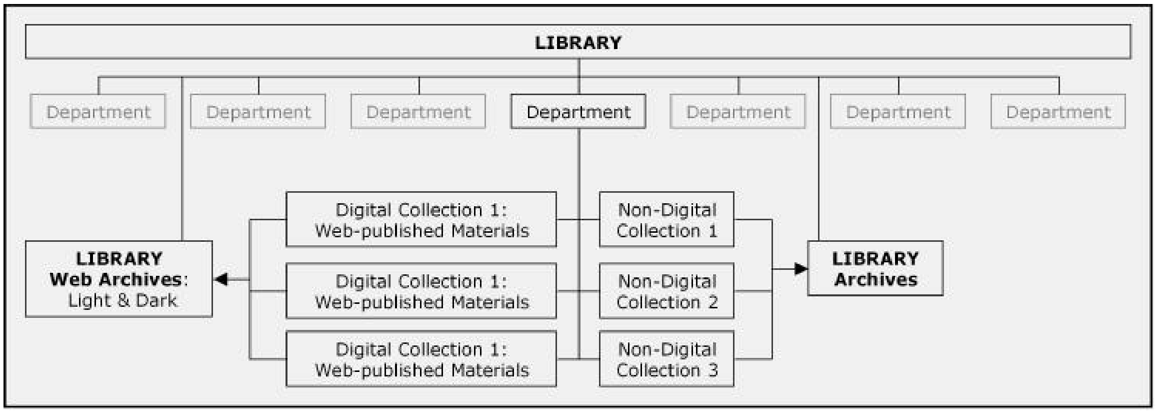
LIBRARY
Department
Department
Department
Department
Department
Department
Department
Digital Collection 1: Web-published Materials
Digital Collection 1: Web-published Materials
Digital Collection 1: Web-published Materials
Non-Digital Collection 1
Non-Digital Collection 2
Non-Digital Collection 3
LIBRARY Web Archives: Light & Dark
LIBRARY Archives

Figure 2. Library Context for Collections and Archives

Several departments exist within a library and each department has curators or other librarians who are charged with collection development responsibilities for traditional (non-digital) collections as well as for digital collections, which may consist of web-published materials (web sites) or other digitized materials such as letters or images. Traditional collection development activities encompass material selection, acquisition, management (including deselection and removal), and preservation (including the transfer of materials from circulating collections to library archives and remote storage facilities). Collection development for web-published materials includes similar functions. Patron access to a library's web archives can be designated as light or dark.

A web archive is a collection of web-published materials for which an institution has either accepted long-term preservation and access responsibility or made arrangements with a third-party to do so. Some of these materials may also exist in other forms but the web archive captures the web versions for posterity and provides access in keeping with the archive's user access policies.

Web archive development includes three major phases: selection, curation, and preservation. Table 1 briefly explains each of these phases. Establishing policies and practices to guide activities and decisions in each of the phases is an important ongoing activity. Although there is a general trajectory of tasks in the web archive development process, with selection occurring early on and preservation toward the end, tasks are not necessarily completed in a linear fashion. By breaking down curators' work into a series of steps, a functional view of web archiving emerges. This view enables the creation of tools targeted to meet curators' needs at various stages of the web archive development process. The assessment tools included in this toolkit have been developed within this context.

Table 1. Functional Areas of the Web Archive Development Process

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.

1.3 Overview of Assessment Activities

The toolkit consists of implementation guidelines and data collection tools for three types of assessment activities.

1. Survey
2. Focus Groups
3. Interviews

The purpose of the survey is to identify both end user and curator needs that might impact collection development activities. Curators completing survey questionnaires are the curatorial partners involved in the Web-at-Risk project. (Section 2.2 lists the curatorial partners.) Most of the curators are or will be involved in building web archives at their institutions. Curators will serve a dual role, representing end user needs in addition to their own.

The focus groups will elicit the needs and issues end-users and librarians have in relation to web archives. Librarians who participate will represent end user needs in addition to their own. It is expected that these needs and issues will translate into requirements that inform guidelines for the Web Archiving Service. For the two focus groups at a national level, participants will be volunteers solicited from among the membership of the Law and Political Science Section (LPSS) of the Association of College and Research Libraries (ACRL) attending the American Library Association (ALA) conference in Chicago in June 2005 or from members attending the Government Printing Office - Federal Depository Library Program (GPO - FDLP) Council Meeting in October 2005 in Washington, DC. For the three local focus groups, participants will be volunteers recruited by each of the three funded project partners at their respective institutions.

Individual interviews will be conducted with both end users and content producers. End user interviews will elicit the needs users have regarding access to the content of web archives. The participants are expected to be researchers, either faculty or graduate students, in departments within the subject domain of the Web-at-Risk project, for example, political science faculty or graduate students. Interviews with content producers are expected to identify the concerns and needs of publishers of web content in the subject domain of this project, for example, state government web site publishers.

1.4 Institutional Responsibilities

The following persons have responsibilities for coordinating the assessment activities at their respective institutions.

University of North Texas	Kathleen Murray
California Digital Library	Patricia Cruse
New York University	Leslie Myrick

Table 2 indicates responsibility among the three funded project partners (CDL, UNT, and NYU) for conducting the assessment activities and suggests the number of participants to be included in each activity. For example, the University of North Texas is responsible for gathering data via the survey from the project's curatorial partners. Likewise, it is the responsibility of each funded project partner to identify focus group participants and key informants for individual interviews. In all, user needs will be assessed from maximums of 21 curators via a survey, 30-50 librarians via focus groups, 9-15 individual interviews with end users, and 6-9 individual interviews with content producers. This document provides tools for conducting these activities.

Table 2. Assessment Activities, Responsibilities, and Timelines

	Survey (Jun - Jul 2005)		**Focus Groups** (Jun - Aug 2005)		**Interviews** (Aug - Sep 2005)	
Participants	Responsibility	#	Responsibility	#	Responsibility	#
Curators	UNT	21				
Librarians – National (ACRL & FDLP*)			UNT	2 Groups of 6 - 10		
Librarians - Local			1. UNT 2. CDL 3. NYU	1 Group of 6 – 10 per Partner		
End Users					1. UNT 2. CDL 3. NYU	3-5 per Partner
Content producers					1. UNT 2. CDL 3. NYU	2-3 per Partner

Notes. UNT - University of North Texas; CDL - California Digital Library; NYU - New York University; ACRL- American Library Association - Association of College and Research Libraries; FDLP – Government Printing Office - Federal Depository Library Program.
*FDLP focus group will be held in October.

1.5 Institutional Review Board for the Protection of Human Subjects (IRB)

It is the responsibility of each institution engaged in assessment activities to gain the approval of their institutional review board for the specific assessment activities they undertake. While these assessment activities involve minimal risk to participants and may be exempt from full IRB review, those determinations must be made by each institution's IRB in accordance with their procedures.

It is important to submit requests for IRB review in advance of engaging in assessment activities. Additionally, each institution should ensure that IRB approval is gained in advance of the timeframes for assessment activities. The appendices include the assessment instruments and example consent forms for each assessment activity. These may be helpful in preparing IRB review applications.

1.6 Timeframe

Data gathering by project partners commences in June 2005 and is scheduled for completion in October 2005. (See Table 3.)

Table 3. Timeframe for Needs Assessment Activities

	June	July	August	September	October
Survey	X	X			
Focus Groups: National	X				X
Focus Groups: Local	X	X	X		
Interviews: End Users			X	X	
Interviews: Content producers			X	X	

1.7 Submitting Data and Reports

Submit Data & Reports via Email
to the Web-at-Risk Assessment Analyst

Kathleen Murray

krmurray@unt.edu

Data gathered from all activities using the tools provided in this toolkit should be submitted to the Assessment Analyst. It is highly desirable that data be submitted as soon as possible after an activity is completed but no later than the completion dates listed in Table 4.

Table 4. Completion Dates for Needs Assessment Activities

Activity	2005 Completion Dates
Surveys	July 15
Focus Groups: National	October 31
Focus Groups: Local	August 31
Interviews: End Users	September 30
Interviews: Content producers	September 30

1.8 Getting Help

If you need further information or have any questions, please contact the following persons at the University of North Texas.

Kathleen R. Murray
Assessment Analyst, CISA Path
Web-at-Risk Project
Postdoctoral Research Associate
University of North Texas

krmurray@unt.edu

Cathy N. Hartman
Project Manager, CISA Path
Web-at-Risk Project
Head, Digital Projects Department
University of North Texas

chartman@library.unt.edu

2 Survey Questionnaire

2.1 Purpose

The purpose of surveying the curators is twofold.

2.1.1 The first purpose is to identify end user and curator needs that specifically impact collection development for web archives in the functional areas listed below. These needs will inform a Common Collection Planning Framework, which is a future deliverable of this project and which is intended to be a set of guidelines to assist curators and others in the development of web archives.

1. Policy Setting
2. Selection
    a. Selection of Materials
    b. Acquisition of Materials
3. Curation
    a. Description of Materials
    b. Organization of Materials
    c. Presentation of Materials
    d. Maintenance of Materials
    e. Deselection of Materials
4. Preservation
    a. Preservation of Materials

2.1.2 The second purpose is to identify functional requirements for the Web Archiving Service crawler and crawler analyzer tools. These tools will enable curators and others to conduct activities vital to the creation, maintenance, and preservation of web archives. The core activities are listed below.

1. Content crawling
2. Crawl progress monitoring
3. Crawl quality assessment
4. Management and description of crawled content
5. Searching and browsing of crawled content
6. Preservation of crawled content

2.2 Participants

Participants include the 21 curatorial partners involved in the Web-at-Risk project. All participants volunteered to participate in the project and many are or will be involved in building web archives at their institutions.

Public Policy and Political Movements	
Gabriella Gray	Curator Online Campaign Literature Archive Young Research Library UCLA

Public Policy and Political Movements	
Ronald J. Heckart	Director Institute of Governmental Studies Library Institute of Governmental Studies UC Berkeley
Terence K. Huwe	Director Library and Information Resources Institute of Industrial Relations UC Berkeley
Peter Filardo	Tamiment Archivist Tamiment Library New York University
Michael Nash	Head Tamiment Library & Robert F. Wagner Labor Archives New York University
Nick Robinson	Librarian Institute of Governmental Studies Library Institute of Governmental Studies UC Berkeley

Local, State, Federal, and International Government Information	
Sherry DeDecker	Head Government Information Center Davidson Library UC Santa Barbara
Charles Eckman	Head Social Sciences Resource Center Green Library Stanford University
Valerie Glenn	Electronic Resources Coordinator Government Documents Department University of North Texas Libraries
James R. Jacobs	Local, State, and International Government Information Librarian Social Sciences and Humanities Library UC San Diego
Kris Kasianovitz	Reference and Instruction Local and State Government Information Librarian Young Research Library UCLA
Amy Kautzman	Head, Research Reference and Collections Doe/Moffitt Libraries UC Berkeley
Linda Kennedy	Head Government Information and Maps Department Shields Library UC Davis

Local, State, Federal, and International Government Information	
Ann Latta	State and Local Documents Bibliographer Social Sciences Resource Center Green Library Stanford University
Janet Martorana	Local & California Documents / Environmental Sciences Librarian Davidson Library UC Santa Barbara
Lucia Orlando	Government Information Librarian University Library UC Santa Cruz
Richard Pearce-Moses	Director Digital Government Information Archives and Public Records Arizona State Library
Lynne Reasoner	Government Publications Librarian UCR Libraries UC Riverside
Juri Stratford	Government Information Librarian Shields Library UC Davis
Yvonne Wilson	California and Orange County Government Information Librarian Langson Library UC Irvine
Arlene Weible	Head of the Government Documents Department University of North Texas Libraries

2.3 Methods

The web-based survey will be developed and tested by project team members at the University of North Texas. The survey will be online and available for completion by curatorial partners in June and July 2005. Approval of the IRB at UNT for administering the survey will be obtained in advance of this date. The assessment analyst at UNT will analyze survey data.

Analysis will consist of both qualitative and quantitative methods to identify curator and end user needs as well as curators' requirements for the Web Archiving Service tools. These needs will inform creation of (a) a Common Collection Planning Framework and (b) the tools for the Web Archiving Service, which are being created under the auspices of this project to assist curators in the development of web archives.

2.4 Tools

The research consent form for the web-based survey will be a letter accessed by participants prior to completing the survey. (See Appendix 2.) The survey questionnaire consists of 59 questions organized in five sections. (See Appendix 3.) The first section asks respondents to identify their current collections. The middle three sections pose questions relating to the major functional areas of web archive development: selection, curation, and preservation. The last section asks questions regarding the Web Archiving Service tools.

2.5 Timeframe

The online survey will be made available in mid-June 2005. Surveys should be submitted by July 15, 2005. Curators will be notified when the survey is available via the NDPPCURATORS-L mailing list (Web-at-Risk Curators List).

3 Focus Group Interviews

3.1 Purpose

The focus groups will identify needs and issues users and librarians have in relation to web archives. Librarians who participate will serve a dual role, representing end user needs in addition to their own.

Needs and issues identified during the five focus group sessions will be provided to the Web-at-Risk project team for further analysis. It is expected that these needs and issues will translate into requirements that inform guidelines for a web archiving service, which is a goal of the project. It is also quite likely that librarians participating in the focus group will identify needs, issues, requirements, or activities that might inform local plans or strategies for developing web archives.

3.2 Participants

The University of North Texas will conduct two focus groups with librarians from national organizations:

- American Library Association: Association of College and Research Libraries – Law and Political Science Section
(ALA: ACRL - LPSS)

- Government Printing Office: Federal Depository Library Program
(GPO: FDLP)

In both cases, participants will be volunteers solicited from among the membership attending the ALA conference in Chicago in June or the FDLP Council Meeting in October in Washington, DC. In order to widen participation beyond academic libraries, efforts will be made to recruit focus group participants from public libraries at the FDLP meeting.

Three additional focus groups with librarians will be conducted by each of the three funded project partners at their respective institutions: University of California, University of North Texas, and New York University. It is the responsibility of the funded project partners to identify focus group participants.

Participants are anticipated to be librarians representing their own needs as well as the needs of the end users they serve. Each group will include 6-10 participants who should have some commonalities, for example, institution, experience, or job focus.

3.3 Methods

Focus group planning should begin immediately since some lag time will exist between inviting participants to the meeting and confirming their attendance. Additionally, approval of the Institutional Review Boards at UNT, the CDL, and NYU for conducting the focus groups should be obtained prior to the focus group sessions. (Appendix 4 is an example participant research consent form.)

As much as possible efforts should be made to follow this methodology for each focus group session. Consistency in methodology will promote confidence in results. The focus group process begins with reviewing the tools in Appendices 4 through 9. This review will provide facilitators with general focus group guidelines as well as an overview of the topics to be addressed in the session. This background may help identify potential participants.

Once potential participants are identified, they need to be invited to participate and their agreement to participate should be confirmed. Logistics for the meeting are important. Attention to logistical details can translate into a comfortable and enjoyable experience for participants.

The agenda for the 2 – 3 hour focus group session opens with introductions, proceeds through topical discussions outlined in the facilitator's guide, and concludes with participants completing a brief questionnaire and being offered a gift to thank them for their time and for sharing their ideas and experiences. Depending on the length of the session, it may desirable to offer a break at the mid-point. It may not be possible to address all of the topics in the discussion guide in the session. Care should be taken by facilitators to discuss topics in the order provided. By doing this, it is likely that topics related to the selection and curation processes, which are a major focus of the CISA project path, will be addressed.

If the session is not recorded and transcribed, it is helpful to have two individuals take notes during the session. The questionnaire used by the facilitator to guide the discussion can serve as a general outline for note takers, who ideally will record their notes on a computer. Attention should be paid to recording examples, key quotes, observations or impressions of group dynamics, and summary conclusions or ideas that emerge.

Immediately following the session, note takers should augment their notes for clarity. In consultation with the note takers, the facilitator should create a summary report that identifies key needs and issues for each topic discussed by the group. Upon completion and no later than August 31, 2005, all notes and the facilitator's summary report should be sent to the Assessment Analyst.

Methodological details and key points are listed below. The appendices provide additional details and tools for conducting focus group sessions.

3.3.1 Letter of Invitation

- Include the purpose of the focus group within the context of the Web-at-Risk project and the National Preservation Program
- Explain
  - Voluntary participation
  - Confidentiality of input
  - Logistics: who, what, where, when
- See Appendix 5 for a sample letter

3.3.2 Letter of Confirmation

- Identify logistics: who, what, where, when
- Include stimulus questions
- Provide contact information
- See Appendix 6 for a sample letter

3.3.3 Logistics

- Recruit
  - Facilitator
  - Note taker(s)

- Reserve
  - A quiet, comfortable room
  - Recording equipment (optional)

- Procure
  - Light refreshments
  - Table tents (if tables will be used) or name tags to help personalize and facilitate the discussion
  - White board or flip chart
  - Black markers
  - Thank you gifts

3.3.4 Meeting Agenda

- Opening by facilitator
  - Introduce self
  - Thank participants
  - Identify the facilities and exits
  - State the purpose
  - Describe the agenda & timeframe
  - Obtain signatures on & collect participant research consent forms (Appendix 4)

- Introductions by participants

- Topical discussions
  - Follow Focus Group Discussion Guide (Appendix 8)
  - Topics are arranged in order of importance
  - Time constraints may not allow discussion of all topics
  - Allow time at end for closing activities

- Closing by facilitator
  - Handout & request completion of participant questionnaire
  - Thank participants
  - Offer thank you gifts as participants turn in their completed participant questionnaires

3.3.5 Recording the Discussion

- Consider audio taping the group meeting and having a verbatim transcript made

- Have 2 note-takers in the meeting

- Instructions for note-takers
  - Record notes on a computer using the Focus Group Discussion Guide as an outline
  - Identify and add any issues or discussion areas not addressed in the focus group questionnaire
  - Take detailed notes of the conversation, including:
    - Quotes that illustrate points
    - Agreements among participants
  - Record and identify any subjective observations or impressions of participants

- Instructions for facilitator
    - Take hand-written notes if possible using the Focus Group Discussion Guide as an outline
    - Transcribe and augment any notes into a computer immediately after the meeting
    - Collect and review notes from note-takers
    - Write a brief summary characterizing the group and the session
    - Write a 2-3 page summary (as appropriate) of each major discussion topic, including themes, key points, agreements, and issues
    - Write a 2 page summary of the group discussion

3.3.6 Submitting Documentation, Notes, & Summaries

- Facilitator or coordinator at the institution hosting the focus group should submit the following to the Assessment Analyst as soon as possible after the focus group session
    - Signed Research Consent Forms
    - Letter of Invitation
    - Letter of Confirmation
    - Participant list
    - Notes from note-takers and facilitator
    - Facilitator's
        - Summary characterization of the group
        - Summaries for each discussion topic
        - Group summary
    - Completed Participant Questionnaires

3.4 Tools

Appendices 4 – 9 contain six tools for coordinators and facilitators. The first tool is the consent form. The second and third tools are letters of invitation and confirmation. The fourth provides general guidelines for facilitating the discussion. The fifth tool is the discussion guide. It includes the nine discussion topics and should be used by the facilitator to guide the discussion and by the note-takers to record their notes and observations. The sixth tool is a participant questionnaire to be completed by participants at the close of the meeting. The purpose of the participant questionnaire is to capture information to characterize the participants and to augment or confirm key points about topics of interest. Each of the tools is contained in an appendix.

1. Appendix 4. Research Consent Form: Focus Groups & End User Interviews
2. Appendix 5. Letter of Invitation to Focus Group Participants
3. Appendix 6. Letter of Confirmation to Focus Group Participants
4. Appendix 7. Focus Group Facilitator Guide
5. Appendix 8. Focus Group Discussion Guide
6. Appendix 9. Focus Group Participant Questionnaire

3.5 Timeframe

The focus groups arranged by the three funded project partners should be held in the June – July timeframe. Documentation, notes, and summaries from the meetings should be provided to the Assessment Analyst no later than August 31, 2005.

4 End User Interviews

4.1 Purpose

End user interviews are an opportunity to gather in-depth information from patrons and end users in a one-on-one setting. End user interviews will elicit the needs users have regarding access to the content of web archives. These needs will inform the guidelines for a web archiving service.

General goals of the interview include identifying:

- The scope, type, and format of web-published resources that are important to the research needs of your end users
- Issues surrounding selection and acquisition of these materials for a web archive
- End users' requirements for access to the contents of web archives

4.2 Participants

It is expected that each of the three funded project partners (CDL, NYU, and UNT) will identify three to five patrons or end users at their respective institutions for one-on-one interviews. The participants are expected to be researchers, either faculty or graduate students, in departments within the subject domain of the Web-at-Risk project, for example, political science faculty or graduate students. Selection should be based in part on the participant's capacity as a key informant, that is, a person possessing knowledge in research using web-published materials as well as good communication skills.

4.3 Methods

Approval of the Institutional Review Boards at UNT, the CDL, and NYU for conducting the interviews should be obtained in advance. Prior to beginning each interview, participants should sign a research consent form. (See Appendix 4.)

Once possible participants are identified, they should be invited either by telephone or mail. The method of contact may vary according to the interviewer's relationship with the person. A sample letter of invitation is included in Appendix 10. After the participant (interviewee) has agreed to the interview, a confirmation letter should be sent. (See Appendix 11.)

Interviewer tips are provided in Appendix 12 and an end-user interview questionnaire is provided in Appendix 13. The questionnaire includes participant background questions as well as questions in five topic areas. The questionnaire serves as a general guide to anticipated topics of interest. It is likely that the interview will elicit information that is not addressed in the questionnaire. The role of the interviewer is to listen to the individual being interviewed, to pursue topics that emerge, and to record what is said.

4.3.1 <u>Letter of Invitation</u>

- Include the purpose of the interview within the context of the Web-at-Risk project and the National Preservation Program

- Explain
  - Voluntary participation
  - Confidentiality of input
  - Time involved
  - Interviewer contact information

- See Appendix 10 for a sample letter

4.3.2 <u>Letter of Confirmation</u>

- Identify logistics: who, what, where, when
- Include stimulus questions
- Provide contact information
- See Appendix 11 for a sample letter

4.3.3 <u>Recording the Interview</u>

- Consider audio taping the interview and having a verbatim transcript made
- Have a note-taker in the interview
- Instructions for note-takers
  - Create notes on a computer using the End User Interview Questionnaire as an outline
  - Identify and add any issues or discussion areas not addressed in the questionnaire
  - Take detailed notes of the conversation, including any quotes that particularly illustrate points
  - Record subjective observations or impressions of participants and identify them as such
- Instructions for interviewers
  - Obtain participant's signed consent
  - Take hand-written notes if possible using the End User Interview Questionnaire as an outline
  - Transcribe and augment any notes into a computer immediately after the meeting
  - Collect and review notes from the note-taker
  - Write a 2-3 page summary (as appropriate) of each interview
    - Describe the interviewee's background from questionnaire
    - Identify key points, needs, requirements, and issues
    - Focus on web archives and access to the archives
  - Write a 2 page summary of all interviews conducted at your institution
    - Identify and describe common needs or issues
    - Identify and describe common access needs and requirements

4.3.4 <u>Submitting Notes & Summaries</u>

- Either the interviewer or coordinator at each institution should submit the following to the Assessment Analyst as soon as possible after the end user interviews are conducted
  - Signed consent forms
  - Notes from note-takers and interviewer
  - Summaries of each interview
  - Combined summary of all interviews

4.4 Tools

- Appendix 4. Research Consent Form: Focus Groups & End User Interviews
- Appendix 10. Letter of Invitation to End Users

- Appendix 11. Letter of Confirmation to End Users
- Appendix 12. Interviewer Tips
- Appendix 13. End User Interview Questionnaire

4.5 Timeframe

The interviews at the three funded project partners' institutions should be conducted in the August – September timeframe. Notes and summaries from the meetings should be provided to the Assessment Analyst no later than September 30, 2005.

5 Content Producer Interviews

5.1 Purpose

Unless web-published materials to be included in a web archive are free of intellectual property considerations, intellectual property agreements will be required between the archiving institution and the information provider or content producer(s). As part of the web archive creation process, it will be necessary for curators or other information professionals to identify the issues, needs, and requirements of content producers and web publishers.

As part of the Web-at-Risk project, guidelines for the creation of web archives will be created and an understanding of content producers' needs will provide valuable input to that effort. It is also possible that interviews with specific content producers may have immediate application for an institution as they develop plans and strategies for the creation of specific web archives.

5.2 Participants

It is expected that each of the three funded project partner institutions (CDL, NYU, and UNT) will identify two to three content producers for one-on-one interviews. The participants are expected to represent the interests and needs of web publishers in the domain of this project, for example, federal or state government entities such as the National Library of Medicine or the Government Printing Office.

Selection should be based in part on a participant's capacity as a key informant, that is, a person knowledgeable of the publishing and ownership issues involved in building web archives. Their perspective should offer insight into content producers' needs relative to web archives of web-published materials.

5.3 Methods

Approval of the Institutional Review Boards at UNT, the CDL, and NYU for conducting the interviews should be obtained in advance. It is expected that interviews will be conducted by telephone. Prior to beginning each interview, interviewers should obtain verbal consent from participants using the script in Appendix 14.

To avoid possible duplication of content producers targeted for interviews, the three funded project partner institutions will each contribute the names of content producers to a common list. The CISA Assessment Analyst will compile the list and ensure that duplicates are removed from the list by common agreement among the partners using the NDIIPPWEB-L mailing list (NDIIPP Web Archiving Project List). Example letters inviting content producers to participate in the study and confirming their participation are in Appendices 15 and 16 respectively.

The interview will be semi-structured in nature. This means that each interview will utilize the questionnaire provided in Appendix 17, however, the questionnaire serves as a general guide to anticipated topics of interest. Adhering to the order and wording of questions is not expected. The role of the interviewer is to listen to the individual being interviewed, to pursue topics that emerge, and to record what is said.

5.3.1 Recording the Discussion

- Consider audio taping the interview and having a verbatim transcript made. Since the interview may take place over the telephone, special equipment might be required. Additionally, the person being interviewed must be informed that the interview is being recorded and must consent to the recording.

- Include a note-taker in the interview if feasible.

- Instructions for note-takers
  - Create notes on a computer using the Content Producer Interview Questionnaire as an outline
  - Identify and add any issues or discussion areas not addressed in the questionnaire
  - Take detailed notes of the conversation, including any quotes that particularly illustrate points
  - Write subjective observations or impressions of participants and identify them as such

- Instructions for the interviewer
  - Take hand-written notes if possible using the Content Producer Interview Questionnaire as an outline
  - Transcribe and augment any notes into a computer file immediately after the meeting
  - Collect and review notes from the note-taker
  - Write a 2-3 page summary (as appropriate) of each interview
    - Briefly describe the interviewee: name, position, organization, etc.
    - Identify key points, needs, requirements, and issues
    - Focus on intellectual property and preservation issues
  - Write a 2 page summary of all interviews conducted at your institution
    - Identify and describe common needs or issues

5.3.2 Submitting Notes & Summaries

- Either the interviewer or coordinator at each institution should submit the following to the Assessment Analyst as soon as possible after the interviews are conducted
  - Notes from note-taker and interviewer
  - Summaries of each interview
  - Combined summary of all interviews

5.4 Tools

Appendix 14.	Research Consent Script for Telephone Interviews
Appendix 15.	Letter of Invitation to Content Producers
Appendix 16.	Letter of Confirmation to Content Producers
Appendix 17.	Content Producer Interview Questionnaire

5.5 Timeframe

The interviews conducted by the three funded project partners’ institutions should be completed in the August – September timeframe. Notes and summaries from the interviews should be provided to the Assessment Analyst no later than September 30, 2005.

Appendix 1. Glossary

Acquisition	For digital materials, see Capture
Authenticity	The genuineness of a digital object. Verification of authenticity requires ascertaining that the object is what it claims to be or is what the metadata associated with the object asserts it to be. Authenticity of a digital object is determined in several ways including checksums, provenance, and digital signatures.
Automated Capture Tool	See Crawler
Baseline Metadata	Baseline metadata is machine-generated and captured by a crawler at the time of data capture.
Born-digital	Created originally in digital format (i.e., a machine-readable format). Examples include scientific databases, sensory data, digital photographs, and digital audio and video recordings. A born-digital resource may or may not have a counterpart analog format but, if it does, the digital version existed prior to the counterpart.
Capture	The process of copying digital information from the web to a repository for collection or archive purposes.
Collection	A group of resources related by common ownership or a common theme or subject matter. A web collection consists of one or more crawls that capture a group of related websites (e.g., candidate websites for state election campaigns). Collections are owned and/or maintained by an organization or institution.
Crawl	The content associated with a web capture operation that is conducted by a crawler.
Crawler	Software that explores the web and collects data about its contents. A crawler can also be configured to capture web-based resources. It starts a capture process from a seed list of entry point URLs (EPUs).
Curation Process	Collection development for web-published materials includes the selection, curation, and preservation processes. In this context, the curation process involves description, organization, presentation, maintenance, and deselection of the materials in the collection.
Dark Archive	A digital archive to which no end user access is permitted.
Dark Web	See Deep Web
Deep Web	Resources available via the World Wide Web that are invisible to or inaccessible by crawlers. These resources may be invisible or inaccessible to crawlers because they (a) are contained in a database or other data store, (b) require information collected from the end-user before they are created, or (c) are password protected.
Digital Archive	A digital collection for which an institution has agreed to accept long-term responsibility for preserving the resources in the collection and for providing continual access to those resources in keeping with an archive's user access policies.
Digital Collection	A collection consisting entirely of born-digital or digitized materials.

Digital object	Also called a digital information object. Digital objects can be interactive works (e.g., video games), sensory presentations (e.g., music or audio), documents, and data. Two types of digital objects included in digital archives are: surrogates of information objects in various original formats, (e.g., print books or audio tapes) and born-digital objects.
Dynamic Web Page	A web page created automatically by software at the web server. The page may be (a) personalized for the user based on identification via login or based on cookies stored on the user's computer, (b) tailored to fulfill a specific request made by the user, or (c) code-generated (e.g., using php, jsp, asp, or xml). Information used for personalization or tailoring of pages may be retrieved in real-time from a database or other data store.
Emulation	A method by which newer software interacts with older resources and displays the result using the same commands and formatting that the software that created the resource used. Emulation provides a means of allowing a digital resource to be preserved without altering its binary format.
Enriched Metadata	Enriched metadata is generally specific to an organization and contains a mixture of baseline metadata and human-generated metadata added subsequent to data capture.
Entry Point URL	A URL appearing in a seed list as one of the starting addresses a web crawler uses to capture content. Also called a targeted URL.
External Link	A hyperlink which takes the user to a new website. For a web archive, an external link is one that takes the user out of the archived collection.
Fixity	The extent to which an archived object remains unchanged over time regardless of access and movement due to copying. One common fixity mechanism used to establish and protect the integrity of a digital object (or data) is the result of a cyclical redundancy check (CRC). Redundancy checks are sometimes referred to as checksums.
Harvest	See Capture
Invisible Web	See Deep Web
Light Archive	A digital archive accessible to end-users.
Migration	A method of preserving digital materials and access to those materials by copying or reformatting the materials while preserving their intellectual content.
Persistent Name	A unique name assigned to a web-based resource that will remain unchanged regardless of movement of the resource from one location to another or changes to the resource's URL. Persistent names are resolved by a third party that maintains a map of the persistent name to the current URL of the resource.
Repository	The physical storage location and medium for one or more digital archives. A repository may contain an active copy of an archive (i.e. one that is accessed by end users) or a mirror copy of an archive for disaster recovery.

Seed List	One or more entry point URLs from which a web crawler begins capturing web resources. Curators, or others responsible for building collections of web-based resources, specify seed lists for specific crawls.
Spider	See Crawler
Targeted URL	See Entry Point URL
Visibility	The extent of end user access allowed to a digital archive.
Web Archive	A collection of web-published materials that an institution has either made arrangements for or has accepted long-term responsibility for preservation and access in keeping with an archive's user access policies. Some of these materials may also exist in other forms but the web archive captures the web versions for posterity.
Web Archive Service	Enables curators to build collections of web-published materials that are stored in either local and/or remote repositories. The service includes a set of tools for selection, curation, and preservation of the archives. It also includes repositories for storage, preservation services (e.g., replication, emulation, and persistent naming), and administrative services (e.g., templates for collection strategies, content provider agreements, and repository provider agreements.)
Web-published materials	Web-published materials are accessed and presented via the World Wide Web. The materials span the cultural heritage spectrum and include a range of material types from text documents to streaming video to interactive experiences. Web-published materials are both dynamic and transient. They are at risk of disappearing. Web archives preserve web-published materials.

Appendix 2. Web-Based Survey Research Consent Letter

Title of Study: Web at Risk: A Distributed Approach to Preserving our Nation's Political Cultural Heritage - Content Identification, Selection, and Acquisition (CISA) Path

Dear Survey Respondent:

The Web-at-Risk project is one of eight digital preservation projects funded in 2004 by the Library of Congress. The Web-at-Risk 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 collections of web-published materials. As you may be aware, the content will be collected largely from US federal and state government agencies, but will also include political policy documents, campaign literature, and information surrounding political movements.

The Content Identification, Selection, and Acquisition (CISA) path of the project will produce tools and guidelines to assist curators and other information professionals in the development of web archives. We need your input to identify (a) your needs and concerns and the needs of your end users and (b) the functional requirements for the web crawler and associated tools being developed as part of this project.

It is expected that the needs and issues identified as a result of this survey will inform guidelines for a web archiving service. Implementation of these guidelines by curators will help ensure that the collections built as a part of this project address curator and end user needs. It is also quite likely that curators completing the survey will identify needs, issues, requirements, or activities that might inform their local plans or strategies for developing web archives.

Survey data will be accessible only to project researchers and analysts. While lists of participants may be published to acknowledge individual contributions to the project or for documentation of the breadth of contributions to the research, no public or published analysis or reports will identify individual respondents in such a way that responses can be attributed to them.

This research study has been reviewed and approved by the UNT Institutional Review Board (IRB). The UNT IRB can be contacted at (940) 565-3940 or sbourns@unt.edu with any questions regarding the rights of research subjects.

Your participation in this study is completely voluntary. If you have any questions about this study, please contact Kathleen R. Murray, Ph.D., CISA Path Assessment Analyst, by sending email to: krmurray@unt.edu.

Thank you very much for your help with this study.

Kathleen R. Murray, Ph.D.
Assessment Analyst, Web-at-Risk Project
Postdoctoral Research Associate
University of North Texas

Appendix 3. Needs Assessment Survey

[Questionnaire begins on next page.]

DigitalPRESERVATION
NATIONAL DIGITAL INFORMATION INFRASTRUCTURE AND PRESERVATION PROGRAM

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

Content Identification, Selection, and Acquisition Path

Needs Assessment Survey

Purpose: The purpose of this assessment is twofold:

1. To identify curator and end-user needs that impact the collection development process for web archives

2. To identify the requirements for the Curator User Interface (CUI) to the web crawler and associated tools in the following functional areas:

   a. Content crawling
   b. Crawl progress monitoring
   c. Crawl quality assessment
   d. Management and description of crawled content
   e. Searching and browsing of crawled content
   f. Preservation of crawled content

Directions: The survey will be completed online. Curators participating in the study may find it helpful to review the text version of the survey prior to completing the online version.

Help: A table outlining the functional areas of the web archive development process can be found at the end of the survey. Please note that as curators in the Web-at-Risk project you are not responsible for all of these functional areas (e.g., maintenance activities). A Glossary of terms used in the survey will be available online. (See also Appendix 1 of this Toolkit.)

Please feel free to contact Kathleen Murray, Assessment Analyst for the Web-at-Risk project, with any questions you may have.

NDIIPP Information: The National Digital Information Infrastructure and Preservation Program (NDIIPP) at the Library of Congress is a program initiated and funded by the US Congress in 2000. In 2004 the program provided funding to eight collaborative projects to carry out the goal of establishing a national network of partners committed to the digital preservation of cultural heritage materials. More information is available at: http://www.digitalpreservation.gov/

Web-at-Risk
Project Information: The Web-at-Risk 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 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 surrounding political movements.

Section A. About Your Collections

To help us understand your needs better, please describe the collections that either you manage directly or your staff manages.

1. What is the overall focus of your collections, including both digital and print materials?

2. Who are the end users of your collections?

3. Please list and briefly describe four of your most important digital collections.

1. Name	Location or URL
Brief Description	
2. Name	Location or URL
Brief Description	
3. Name	Location or URL
Brief Description	
4. Name	Location or URL
Brief Description	

4. For each material type, estimate the percentage of items in your most important digital collections that are web-published.

		0%	<25%	25 – 50%	51 – 75%	>75%
a.	Journals & Periodicals					
b.	Books & Brochures					
c.	Databases					
d.	Newspapers					
e.	Videos					
f.	Audio files					
g.	Image files					
h.	Technical & Research Reports					
i.	Proceedings of Meetings & Symposia					
j.	Doctoral Dissertations & Master's Theses					
k.	Government Records					
l.	Unpublished Works & Publications of Limited Circulation					
m.	Other: ______________________________					
n.	Other: ______________________________					
o.	Other: ______________________________					

5. If any of your unlicensed digital collections contain web-published materials, do you currently maintain a digital archive for the long-term preservation of these collections? (Select one.)

a. _____ Yes
b. _____ No (Skip questions 6 & 7. Go to the next page.)

6. What best describes the underlying software or management tools your archive(s) uses? (Select all that apply.)

a. _____ Web / HTML interface to mirrored websites

b. _____ Content Management System (CMS)

→ Please specify: ______________________

c. _____ Institutional Repository Software (e.g., DSpace, Eprints, Fedora)

→ Please specify: ______________________

d. _____ Other

→ Please specify: ______________________

7. Please describe the two greatest hurdles you encountered in creating your archive(s).

1. ____

____

2. ____

____

Section B. Selection: Policy, Identification, & Acquisition

Answers to the following questions will help determine the impact of user needs on collection policies and practices.

8. Indicate if your collection policies or practices specifically include or exclude support of digital formats for the following material types.

Material Types		Include (√)	Exclude (√)		Not Specified (√)
a.	Journals & Periodicals				
b.	Books & Brochures				
c.	Databases				
d.	Newspapers				
e.	Videos				
f.	Audio files				
g.	Image files				
h.	Technical & Research Reports				
i.	Proceedings of Meetings & Symposia				
j.	Doctoral Dissertations & Master's Theses				
k.	Government Records or Documents				
l.	Unpublished Work & Publications of Limited Circulation				
m.	Other: ________________________________				
n.	Other: ________________________________				
o.	Other: ________________________________				

Additional Comments:

____

____

____

9. Indicate the acceptability of each of the following digital formats in your digital collection policies or practices. (Examples of limits: Only certain types of audio formats are acceptable or only video files under a specified size are acceptable.)

Digital Format		Acceptable (√)	Acceptable within Limits (√)	Not Acceptable (√)		Not Applicable (√)
a.	Adobe Portable Document Format (pdf)					
b.	Adobe PostScript (ps)					
c.	Lotus 1-2-3 (wk1, wk2, wk3, wk4, wk5, wki, wks, wku)					
d.	Lotus WordPro (lwp)					
e.	MacWrite (mw)					
f.	Microsoft Excel (xls)					
g.	Microsoft PowerPoint (ppt)					
h.	Microsoft Word (doc)					
i.	Microsoft Works (wks, wps, wdb)					

Digital Format		Acceptable (√)	Acceptable within Limits (√)	Not Acceptable (√)		Not Applicable (√)
j.	Microsoft Write (wri)					
k.	Rich Text Format (rtf)					
l.	Shockwave Flash (swf)					
m.	Audio (mp3, wav, midi, ra)					
n.	Images (jpeg, jpg, gif, png, tif)					
o.	Text (ans, txt)					
p.	Video (mpeg, ra, mov, rm)					
q.	Web Pages (htm, html, asp, jsp, php)					
r.	Supporting Code (css, js)					
s.	Other: ______________________________					
t.	Other: ______________________________					

10. Do contractual, depository, or other arrangements or responsibilities affect the types or formats of materials in your digital collections? (Select one.)

a. ______ Yes
b. ______ No

11. Indicate the level of support in your organization for creating a web archive.

None at All	Very Little	Some	A Fair Amount	A Large Amount
1	2	3	4	5

12. Indicate the level of acceptance your end users would have if web-published materials were not archived due to privacy concerns. For example, a management decision could be made not to archive personal testimony records from public hearings if release forms were not obtained from the individuals testifying.

Not Accepting	A Little Accepting	Somewhat Accepting	Very Accepting	Extremely Accepting		Don't Know
1	2	3	4	5		X

13. Indicate the level of acceptance your end users would have if web-published materials were not archived due to technical roadblocks, such as dynamic web pages or password-protected materials.

Not Accepting	A Little Accepting	Somewhat Accepting	Very Accepting	Extremely Accepting		Don't Know
1	2	3	4	5		X

For the following questions, think about a collection of web-published materials you are planning to create or add to as a part of the Web-at-Risk project. If you have not identified specific source materials, consider materials of interest to the primary end users of your collection and the web-based sources your end users accept as credible and authoritative.

14. At what level will you primarily select source materials for your planned web archive? (Select one.)

a. ______ Object level (Example: images or movies)
b. ______ Web page level (Example: .html, .xml, etc.)
c. ______ Logical document level (Example: article spanning multiple .html files)
d. ______ Website level (Example: all content within a URL)
e. ______ Organizational level (Example: websites within an agency's top-level URL)

15. Are you definitely planning to collect materials from any commercial sources, for example, news sites?

a. ______ Yes
b. ______ No

If yes, please describe the commercial information source(s) and list their respective URLs, if known.

Source Description	Source URL

16. Briefly describe any circumstances in which you might collect commercial source materials?

____

____

____

____

17. Are you planning to collect materials from sources outside the United States?

a. _____ Yes →
b. ______ No

If yes, please describe the information source(s), indicate if the content is commercial or not, and list respective source URLs, if known.

Source Description	Commercial Content	Source URL
	Y N	
	Y N	
	Y N	
	Y N	
	Y N	

18. What other web-based information sources and publishers you are considering for possible inclusion in your collection? Example: Web sites of Chambers of Commerce in Texas, which are published by local city governments.

____

____

____

____

19. Describe the major intellectual property considerations you anticipate for access, use, and reproduction of the source materials in your planned collection.

____

____

____

____

20. Considering the source materials for your planned collection, estimate how often they change or are updated.

Not at All	A Little	Somewhat	Quite Often	At Least Daily		Don't Know
1	2	3	4	5		X

21. After the initial acquisition of web-published materials for your collection, do you plan to re-acquire the materials at certain intervals? (Select one.)

a. _____Yes
b. _____ No

If yes, at what interval do you plan to re-acquire the materials?

____

22. Web pages often contain links to other web sites that are outside of the publishing control of the web site owner. Is the content from the first level of external links important to include in your collection? (Select one.)

a. _____ Yes
b. _____ No

23. Over time, it is likely that some external links in the web archive will no longer be operational (i.e., no longer lead to their originally intended destinations). How would you ideally like an archive to deal with these broken links? (Select one.)

a. _____ Allow selection and let browser provide standard messages for broken links
b. _____ Allow selection but provide custom messages for broken links
c. _____ Deny selection but leave text with no notification of broken links
d. _____ Deny selection but leave text with notification of broken links
e. _____ Other

If other, please explain.

____

____

24. Would it concern you if an archived web page were altered to include additional metadata? (Select one.)

a. _____ Yes
b. _____ No
c. _____ Don't Know

25. Which of the following might endanger the authenticity of materials in a web archive? (Select all that apply.)

a. _____ Multiple versions captured at different points in time
b. _____ Addition of enhanced metadata to captured materials
c. _____ Multiple formats of the same object (e.g., .txt and .pdf)

26. For your planned collection, who will have final responsibility for ensuring the authenticity of web-published materials? (Select one.)

a. ______ Content provider
b. ______ Web archive creator or curator
c. ______ End users
d. ______ Other

If other, please explain.

____

____

27. As you consider creating your collection, estimate the magnitude of the financial challenge facing your organization in each of the following areas.

	Not Challenging	A Little Challenging	Somewhat Challenging	Very Challenging	Extremely Challenging
Needs assessment	1	2	3	4	5
Contract negotiation	1	2	3	4	5
Copyright/intellectual property issues	1	2	3	4	5
Initial hardware & software implementation	1	2	3	4	5
Harvest	1	2	3	4	5
Network access	1	2	3	4	5
Storage	1	2	3	4	5
Cataloging	1	2	3	4	5
Presentation	1	2	3	4	5
Re-harvest	1	2	3	4	5
Management & deselection	1	2	3	4	5
Preservation	1	2	3	4	5
IT Support	1	2	3	4	5
Staff Training	1	2	3	4	5

28. As you consider creating your collection, estimate the magnitude of the technical challenge facing your organization in each of the following areas.

	Not Challenging	A Little Challenging	Somewhat Challenging	Very Challenging	Extremely Challenging		Don't Know
Hardware and software maintenance	1	2	3	4	5		X
Unclear collection boundaries in the web environment	1	2	3	4	5		X
Maintenance of look and feel of original material	1	2	3	4	5		X
Metadata creation	1	2	3	4	5		X
Password protected source material	1	2	3	4	5		X
Encrypted source material	1	2	3	4	5		X
Authenticity	1	2	3	4	5		X
Persistent naming	1	2	3	4	5		X
Dynamic nature of some web materials	1	2	3	4	5		X
Frequency of change	1	2	3	4	5		X
Real-time content changes during capture	1	2	3	4	5		X

Section C. Curation: Description, Organization, Presentation, Maintenance, & Deselection

Answers to the following questions will help identify both the metadata requirements for the organization and presentation of archival materials and the impact of user needs on ongoing archival maintenance activities.

29. Our end users will want to use any word(s) to search the full-text of the web archive.

Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
1	2	3	4	5

30. Our end users will want to search or browse web archive materials by subject categories or topics.

Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
1	2	3	4	5

31. It is important for our end users to interact with archived materials in a fashion that mirrors the website(s) at the time of capture.

Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
1	2	3	4	5

32. Our end users will require access to the materials in our web archives into the foreseeable future.

Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
1	2	3	4	5

Answers to the following questions will help identify the impact of end user needs on material deselection activities.

33. Which of the following criteria for deselection of materials from your web archive will you use? (Select all that apply.)

a. _____ Usage data thresholds
b. _____ Sensitive or offensive material
c. _____ Copyright violation
d. _____ Fraud
e. _____ Storage costs

34. What additional deselection criteria will you use?

____

____

____

35. In general, end users understand if materials are removed from public access or web archives based on how frequently the materials are used.

Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
1	2	3	4	5

36. End users generally understand how copyright protection applies to web-published materials.

Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
1	2	3	4	5

37. In general, end users understand the removal of materials from public access or web archives based on published or known policy guidelines pertaining to potentially sensitive or offensive materials.

Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
1	2	3	4	5

38. In general, end users understand if materials are removed from public access or web archives for legal reasons such as fraud.

Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
1	2	3	4	5

39. In general, end users understand the removal of materials from public access or web archives for financial reasons such as storage costs.

Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
1	2	3	4	5

Section D. Preservation

Answers to the following questions will help identify user expectations that impact web archive preservation activities.

40. End users accept updated versions of web materials supplanting previous versions.

Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
1	2	3	4	5

41. End users expect unique persistent names to identify each version, type, and format of materials in web archives.

Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
1	2	3	4	5

42. It is generally acceptable to end users that retention of multiple versions of web-published materials is dictated by the degree of change from version to version.

Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
1	2	3	4	5

43. It is important to end users that web archive content is replicated in another geographic location.

Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
1	2	3	4	5

44. To ensure access, archived materials may be migrated to new software versions and different formats, platforms, or operating system environments. For each of the following migration events, estimate the threat to the authenticity of archived materials.

	No Threat	Small Threat	Moderate Threat	Significant Threat	Extreme Threat		Don't Know
Migration to new version of same software (e.g., from version 2 to 6 of Microsoft Word)	1	2	3	4	5		X
Migration to different format (e.g., text to pdf)	1	2	3	4	5		X
Migration to different hardware platforms	1	2	3	4	5		X
Migration to different operating system environments	1	2	3	4	5		X
Migration to different file system within an operating system environment	1	2	3	4	5		X

Section E. Curator User Interface

In the Web-at-Risk project, a web archive contains the results of web crawls. Curators initiate crawls by identifying entry-point URLs and other crawl parameters, Curators also build collections by specifying which crawls from the archive to include in collections. Crawls are associated both with the curator who originated them and the collections that contain them. It is possible that some crawls will be included in more than one curator's collection.

The project is creating tools and services to assist curators in their activities at three points in the collection development, or curation, process:

1. After materials are identified for inclusion but prior to final selection
2. When specifying parameters for a crawl
3. During a crawl

Answers to the following questions will help identify functional requirements for a curator's interface to the web archive services and crawler tools being created as part of the Web-at-Risk project.

45. Imagine you have identified potential web-published source materials for your collection as well as targeted URLs (or entry-point URLs) for a crawler to begin the capture process. How important is it for you to evaluate each of the following attributes of the crawl prior to finalizing your selection decisions?

Total crawl size				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Content object types (image, audio, video, etc.)				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Content object formats (html, jpeg, gif, pdf, etc.)				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Total file size by type				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
# Links to external URLs				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Content URLs within the targeted or entry-point URLs				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
# Broken Links within targeted or entry-point URLs				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Failures by # and Type (timeouts, server errors, unsupported schemes such as 'mailto')				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5

46. List any additional attributes you think are important to your material evaluation and selection process.

____

____

____

47. When you define a crawl or capture process, how important is it for you to specify each of the following parameters?

Frequency of the crawl (daily, weekly, monthly, etc.)				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Time period over which to repeat crawl (1 month or 6 months at specified frequency)				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
# Levels within targeted or entry-point URLs to capture				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Depth of links to external URLs to capture				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Compliance with robot exclusions (obey or ignore)				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Content object types to capture (image, audio, video, etc.)				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Content object formats to capture (html, jpeg, gif, pdf, etc.)				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5

48. List any additional parameters you think are important to specify for a crawl.

____

____

____

49. When you configure the crawler at the start of a capture process, how important will it be to exclude web materials based on specific parameters, for example, to exclude materials based on a certain file type?

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

50. As the crawler is capturing materials in accord with the parameters you specified, how important is it that someone monitoring the capture process receives real-time data about each of the following parameters of the materials being captured?

Crawl completion status by targeted or entry-point URL						
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important		Don't Know
1	2	3	4	5		X
Total size captured						
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important		Don't Know
1	2	3	4	5		X
Content object types captured (image, audio, video, etc.)						
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important		Don't Know
1	2	3	4	5		X
Content object formats captured (html, jpeg, gif, pdf, etc.)						
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important		Don't Know
1	2	3	4	5		X
Total file size by object type & format						
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important		Don't Know
1	2	3	4	5		X
Errors encountered by error code (200, 300, 400, 404, 500, etc.)						
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important		Don't Know
1	2	3	4	5		X

51. List any other parameters you think are important for the crawler to report during your material capture process.

____

____

____

Information and data about crawls and the objects they captured can be used to:

- Assist curators as they select crawls from the archive to include in their collections
- Create metadata records
- Establish baseline fixity or data authenticity at the bit level for on-going maintenance
- Analyze the dynamic nature of the archive’s materials

52. Indicate the importance of each of the following collection-level attributes to the overall collection development process, including crawl selection and ongoing collection management activities.

Curator for each crawl in the collection				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Crawl completion date(s)				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Targeted or entry-point URLs for each crawl				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Content URLs within targeted or entry-point URLs for each crawl				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Parameters of each crawl				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Total size of each crawl				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Total collection size by type & format				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
# Errors encountered for each crawl by error code (200, 300, 400, 404, 500, etc.)				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Measurement of content change over time				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5

53. List any other collection-level attributes you think are important for the overall selection and management of a collection in a web archive.

54. Content objects within a collection can be interactive works (e.g., video games), sensory presentations (e.g., music or audio recordings), documents, or data sets. Indicate the importance of each of the following attributes of archived content objects to the overall collection management process.

URL				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Size				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Type (image, audio, video, etc.)				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Format (html, jpeg, gif, pdf, etc.)				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Title				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Author				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Subject				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Description				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Creation date				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Object name (e.g., filename)				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Language				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Archived date				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Measurement of change over time				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5

55. List any other object-level attributes you think are important for the overall management of a collection in a web archive.

____

____

____

56. What level(s) of descriptive metadata is critical for the source materials in your planned collection? (Select all that apply.)

a. ______ Object level (Example: images or movies)
b. ______ Web page level (Example: .html or .xml files)
c. ______ Logical document level (Example: article spanning multiple .html files)
d. ______ Website level (Example: all content within a targeted or entry-point URL)
e. ______ Other: ____

57. The web crawler may capture the following attributes of web-published materials during harvesting. Indicate the importance of each attribute as an end user access point or search criteria for the web archive.

	Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
URL	1	2	3	4	5
Date/Time of Capture	1	2	3	4	5
Object Format/Type	1	2	3	4	5
Language	1	2	3	4	5
File Size	1	2	3	4	5
File Name	1	2	3	4	5
Author	1	2	3	4	5
Title	1	2	3	4	5

58. What additional search criteria will be important to your end users as they interact with your collection?

____

____

____

59. We welcome any additional comments you may have.

____

____

____

Web Archive Development Process: Functional Areas

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 4. Research Consent Form: Focus Groups & End User Interviews

Subject Name ________________________________ **Date** _______________

Title of Study: Web-at-Risk: A Distributed Approach to Preserving our Nation's Political Cultural Heritage - Content Identification, Selection, and Acquisition (CISA) Path

Researcher: [*Insert Researcher's Name*]
[*Insert Researcher's Title*]
[*Insert Institution's Name*]

Before agreeing to participate in this research study, it is important that you read and understand the following explanation of the purpose and benefits of the study and how it will be conducted.

Purpose of the Study

There are two assessment goals in the Content Identification, Selection, and Acquisition (CISA) path. The first is to identify the needs of both web content producers and the users of web archives, that is, those who publish or supply the materials for web-based collections and those who will access and use those materials. The second is to identify the needs of curators and others who will use the web crawler and its analyzer tools, which are being developed as part of this project.

Description of the Study

The Web-at-Risk project is one of eight digital preservation projects funded in 2004 by the Library of Congress. The Web-at-Risk 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 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 surrounding political movements. The project work will be conducted along four paths of overlapping activities.

1. Content Identification, Selection, and Acquisition
2. Content Harvest and Analysis
3. Content Ingest, Retention, and Transfer
4. Partnership Building

One focus of the Content Identification, Selection, and Acquisition (CISA) path is to produce tools and guidelines to assist curators and other information professionals in the development of web archives.

Procedures to be Used

The assessment activity this form applies to is:

1. _____ Focus group with librarians
2. _____ Face-to-face interview with end users

The focus group session will last approximately two hours. Interviews are expected to take one hour.

Description of the Foreseeable Risks

There are no foreseeable risks in these activities.

Benefits to the subjects or others

It is expected that the needs and issues identified in this assessment activity will translate into requirements that inform guidelines and tools for a web archiving service. These tools and guidelines will help ensure that the collections built as a part of this project address the needs of key stakeholders:

institutions, curators, librarians, researchers, end users, and content providers. It is also quite likely that librarians and end users participating in this research study will identify needs, issues, requirements, or activities that might inform local plans or strategies for developing web archives.

Procedures for Maintaining Confidentiality of Research Records
Notes or transcripts will be accessible only to researchers and analysts. While lists of participants may be published to acknowledge individual contributions to the project or for documentation of the breadth of contributions to the research, no public or published analysis or reports of assessment results will identify individuals in such a way that responses, contributions, or analytic results can be attributed to them.

Review for the Protection of Participants
This research study has been reviewed and approved by the Institutional Review Board (IRB) at [*institution's name*]. The IRB can be contacted at [*telephone number*] or [*email address*] with any questions regarding the rights of research subjects.

Research Subject's Rights
I have read or have had read to me all of the above.

[*Interviewer*] has explained the study to me and answered all of my questions. I have been told the risks and/or discomforts as well as the possible benefits of the study.

I understand that I do not have to take part in this study and my refusal to participate or my decision to withdraw will involve no penalty or loss of rights or benefits. The study personnel may choose to stop my participation at any time.

In case I have any questions about the study, I have been told I can contact [*interviewer or researcher*], [*title*], [*institution*], at [*telephone*] or [*email*].

I understand my rights as a research subject and I voluntarily consent to participate in this study. I understand what the study is about, how the study is conducted, and why it is being performed. I have been told I will receive a signed copy of this consent form.

Subject Signature ____________________________________ **Date** _______________

For the Investigator: I certify that I have reviewed the contents of this form with the subject signing above. I have explained the known benefits and risks of the research. It is my opinion that the subject understood the explanation.

Investigator Signature _________________________________ **Date** _______________

Appendix 5. Letter of Invitation to Focus Group Participants

Dear [*Potential Participant*]:

You are invited to participate in a focus group discussion about archiving web-published materials. This activity is being conducted by [*institutional or departmental name*] under the auspices of the Web-at-Risk project, which is one of eight digital preservation projects funded in 2004 by the Library of Congress. The Web-at-Risk 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 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 surrounding political movements. The project work will be conducted along four paths of overlapping activities.

The project is also producing tools and guidelines to assist curators, librarians, and other information professionals in the development of web archives. The focus group will elicit information regarding the needs and issues users and librarians have in relation to web archives. Librarians who participate will serve a dual role, representing end user needs in addition to their own.

Needs and issues regarding web archive development that are identified during the focus group discussion will inform archive development guidelines, which are another planned outcome of the Web-at-Risk project. It is also likely that librarians participating in the focus group will identify needs, issues, requirements, or activities that might inform local plans or strategies for developing web archives.

We hope that you are interested in taking part in this interesting and important activity. Identifying the needs of information professionals and end users will help ensure the development of useful and successful tools and guidelines for web archive development.

Your participation is completely voluntary and, while a list of participants may be published, comments will not be attributed to individuals in a manner that they can be recognized. The focus group meeting is scheduled for, [*date & time*], in [*building & room*]. Refreshments will be served and you will receive a gift of thanks for your participation.

If you are able to attend, please contact [*person*] at [*email & telephone number*] by [*date*]. [*Person*] will contact you to confirm your attendance and provide further details.

Thanks very much for your consideration!

[*Name of Facilitator or Contact Person*]

Appendix 6. Letter of Confirmation to Focus Group Participants

Dear [*Participant*]:

Thank you very much for agreeing to participate in the upcoming focus group discussion about web archiving. I think you will find it stimulating and your contributions will be greatly appreciated.

Meeting Details

Date [*Insert*]
Time [*Insert*]
Place [*Insert*]

The discussion will focus on the activities involved in the creation of collections of web-published materials, or web archive creation, from policy and material selection to maintenance and preservation. Web-published materials span the cultural heritage spectrum and include a range of material types from text documents to streaming video to interactive experiences. These materials are accessed and presented via the World Wide Web and because they are both dynamic and transient, they are at risk of disappearing.

A web archive preserves collections of materials published for the web. Some of these materials may also exist in other forms but the web archive captures the web versions for posterity and may provide users access to the materials long after the web sites themselves are no longer in existence. Two examples of web archives are:

CyberCemetery	PANDORA: Australia’s Web Archive
http://govinfo.library.unt.edu/	http://pandora.nla.gov.au/index.html

Listed below are some questions to stimulate your thinking prior to attending the focus group.

What web-published resources are your patrons or end users currently using for research (e.g., key sites, domains, databases)?

What types of materials (images, web pages, databases, blogs) would you include in a web archive for your patrons?

Who are the owners or publishers of the content you might include in a web archive?

What relationships exist between the library and the owners that might affect web archiving?

What current collection policies apply to web archives?

What administrative guidelines, technical capabilities, or resource constraints might impact the creation of a web archive in your library?

If you are unable to attend the meeting, please let me know as soon as possible [*email, telephone number*].

I look forward to meeting you!

[*Name of Facilitator or Contact Person*]

Appendix 7. Focus Group Facilitator Guide

- Project Background
  - The Web-at-Risk project is a 3-year collaborative effort of the California Digital Library, the University of North Texas, and New York University. The project is one of eight digital preservation projects funded in 2004 by the Library of Congress.
  - The Web-at-Risk project will build archives comprised of web-published materials produced by US federal and state government agencies, as well as political policy documents, campaign literature, and information surrounding political movements.
  - The project is also producing tools and guidelines to assist curators, librarians, and other information professionals in the development of collections of web-published materials.
- Objectives of Focus Group
  - The focus group will elicit information regarding the needs and issues users and librarians have in relation to web archives. Librarians who participate will serve a dual role, representing end user needs in addition to their own.
  - Needs and issues identified during the focus group discussion will inform guidelines for web archive development, which are another planned outcome of the Web-at-Risk project.
  - It is also likely that librarians participating in the focus group will identify needs, issues, requirements, or activities that might inform local plans or strategies for developing web archives.
- Atmosphere
  - Assure confidentiality: A list of participants will be published but comments will not be attributed to individuals in a manner that they can be recognized.
  - Explain the structure for the group: topics to be addressed and timeframes.
  - Foster a sense of ease and rapport among the participants. Allow individuals to introduce themselves, for example by providing 3 types of information: name, work involvement, and experience with web archiving. This helps to provide structure in the beginning.
  - Use table tents or nametags to personalize the discussion and increase the level of ease and familiarity among the participants.
- Interviewer Conduct
  - Contribute as little as necessary to the discussion.
  - Use and allow participants to use available white boards or flip charts to illustrate ideas
  - Use simple probes for more information: for example, “Tell me more about that.”
  - Avoid contradictions.
  - Avoid communicating judgments either verbally or non-verbally.

- Tune into and use non-verbal cues and information.
- Invite less verbal individuals to add contributions.
- Summarize consensual views, needs, opinions, problems, issues, etc. at close of each topic and end of meeting.
- Gain concurrence to summary.

- Discussion Guide
  - Be familiar with the discussion guide prior to the session.
  - Follow the topics as presented in the discussion guide.
  - Cover all topics if time allows.
  - Introduce each topic in an organized and logical fashion.
  - Abandon topics of inquiry that:
    - Appear irrelevant to the participants
    - Encounter a dead end in the discussion
    - Consume an inappropriate amount of time
  - Add significant areas of inquiry that emerge.
  - Avoid closed-ended questions, for example questions that elicit ‘yes’ or ‘no’ answers, when seeking information.
  - Use closed-ended questions when gaining concurrence to summations, for example a restatement of a major issue or need that emerges in the discussion.
  - Adapt to problems with language or terminology.

Appendix 8. Focus Group Discussion Guide

[Guide begins on next page.]

Focus Group Discussion Guide

OPENING

Purpose

Notes to Facilitator

The purpose of these questions is (a) to create a comfortable atmosphere in which people feel valued for their participation, (b) to establish the context for the discussion, and (c) to provide the facilitator with information about the group. This information will help the facilitator modify questions and guide the discussion in directions relevant to the participants and the topics.

Be prepared to offer definitions of key concepts and examples as outlined in the boxes below.

Begin the session by introducing yourself and having the participants introduce either themselves or one another using the four points listed below. Follow the introductions with the recommended dialogue for initiating the focus group session.

<u>Web Archive Creation Process: Web Published Materials</u>

Selection

1. Policy Creation or Modification
2. Identification
3. Acquisition or Harvesting

Curation

4. Description
5. Organization
6. Presentation
7. Maintenance
8. Deselection

Preservation

9. Persistent Naming
10. Format Migration
11. Content Replication
12. Authentication

Key Concepts

Web-Published Materials

Web-published materials are accessed and presented via the World Wide Web. The materials span the cultural heritage spectrum and include a range of material types from text documents to streaming video to interactive experiences. Web-published materials are both dynamic and transient. They are at risk of disappearing. Web archives preserve web-published materials.

Web Archive

A web archive is a collection of web-published materials published. Some of these materials may also exist in other forms but the web archive captures the web versions for posterity and may provide users access to the materials long after the web sites themselves are no longer in existence.

Web Archive Examples

CyberCemetery
http://govinfo.library.unt.edu/

“The University of North Texas Libraries and the U.S. Government Printing Office, as part of the Federal Depository Library Program, created a partnership to provide permanent public access to the Web sites and publications of defunct U.S. government agencies and commissions. This collection was named the "CyberCemetery" by early users of the site.”

PANDORA: Australia’s Web Archive
http://pandora.nla.gov.au/index.html

“PANDORA, Australia's Web Archive, is a growing collection of Australian online publications, established initially by the National Library of Australia in 1996, and now built in collaboration with nine other Australian libraries and cultural collecting organisations.”

Participant Introductions

A. Name, organization, position
B. Collection responsibilities
C. Patrons or users served
D. Experience with web archives

Discussion Context

When a library creates a web archive it is important to ensure user or patron needs inform their plans. The purpose of this group discussion is to elicit your needs and thoughts regarding web archives. As librarians you are in a good position to represent the needs of your patrons.

Our discussion will address needs and issues in the major phases of web archive creation with a particular emphasis on the selection of web-published materials.

For discussion purposes, imagine you are chair of a committee charged with creating a web archive in your library. Your job is to uncover your patrons' needs for a web archive and to identify how those needs impact your institution's existing collection development policies and functional activities. Additionally, you are responsible for identifying issues of any sort (e.g., technical, legal, resource, or administration) related to creating and maintaining the archive.

TOPIC 1: COLLECTION POLICY FOR A DIGITAL ARCHIVE OF WEB-PUBLISHED MATERIALS

Notes to Facilitator

Policy factors influencing the creation of a web archive include political mandates, organizational mission, financial parameters, and technical capabilities.

The purpose of these questions is to identify areas that need to be addressed in a policy for a web archive.

1. Identifying the materials as well as the users and owners of the materials for your proposed web archive are obvious first steps in creating the archive. Unique policy issues and organizational guidelines can then be addressed in this specific context. Briefly describe the targeted users and proposed materials, including the material owners, for the web archive you plan to create.

Notes to Facilitator

This is the opening discussion to set a context for discussions that follow. It allows users to describe the context of their proposed archive. If necessary, stimulate discussion using this outline of ideas.

Users
- Who are your current end users and potential future end users?
- What are their information needs?

Owners
- Who are the information producers or publishers?
- What are the relationships between the library and the content producers (e.g., depository relationships)?

Materials
- What is the subject focus for archive?
- What are the sources of materials: web pages, databases, blogs, etc.?
- What types of materials: text, images, movies, etc. are in the source materials?

2. What financial constraints impact the development of your web archive? How might these be addressed in your web archive policy?

3. What technical constraints impact the development of your web archive? How might these be addressed in your web archive policy?

4. Who within the organization has responsibility for the following items relative to web archive? How might these responsibilities be addressed in your web archive policy?

   - Selection of materials
   - Acquisition of materials
   - Technical support
   - Patron & end user support
   - Cataloging or metadata creation
   - Preservation of materials

- Contracts with content producers
- Copyright permissions
- Privacy issues
- Presentation of materials

5. What details and issues need to be addressed in contractual agreements with the content producers?

> Notes to Facilitator
>
> Suggest the following areas if necessary:
>
> - Specifications of the data (materials) to be archived
> - Minimum metadata to be provided with the data (e.g., description of structure and meaning of data sets)
> - Data delivery specifications: protocols, verification procedures, and frequency
> - Support and maintenance
> - Intellectual property and copyright

6. How is copyright permission for web-published materials addressed in your organization?

> Notes to Facilitator
>
> If necessary, suggest the following ways that copyrighted materials might be handled:
>
> - Archives would be considered a ‘fair use’ application
> - Ignored if materials are not explicitly copyrighted
> - Copyright clearance always requested
> - Required if materials are copyrighted

7. How do your existing copyright policies apply to:

   a. Embedded content such as images or audio files in a web page
   b. Reformatted materials or materials migrated (i.e. copied from one hardware platform to another) to work with newer software and hardware technologies

8. What types of information in your web archive might elicit privacy concerns? How will privacy issues be addressed in your web archive policy?

> Notes to Facilitator
>
> Prompt if necessary with “What about audio files containing personal reflections or data used by information publishers to personalize an individual’s web experience?”

TOPIC 2: IDENTIFICATION OF WEB-PUBLISHED MATERIALS

Notes to Facilitator

Identification of web-published materials for a web archive is impacted by the focus of the archive, the unit of material selection, delineation of web boundaries, copyright obligations, and authenticity of materials.

The purpose of these questions is to articulate issues and needs regarding the identification and characterization of the web-published materials targeted for inclusion in a web archive.

1. How will you evaluate materials for appropriate content? For example:

    a. Consistency with subject matter of interest
    b. Conformance with content policies of the organization

2. Is any content from targeted sources excluded based on policy decisions, for example, dynamic content or streaming video? If yes, what type of content?

3. Discuss the unit of selection for the materials in the archive. What unit is necessary to meet end users' requirements?

    a. Specific objects within web page
    b. Specific web pages within site
    c. Logical document level (multi-page article)
    d. Agency or project web site (e.g., http://www.nih.gov)
    e. Domain (e.g., .mil or .gov)
    f. Other ________

4. Web pages often contain links to other web sites, which are outside of the publishing control of the web site owner. Is the content from external links important to include in your archive? If yes, explain why it is important.

5. How is the content from externally linked web sites evaluated for inclusion in this web archiving effort? What boundaries need to be established?

6. Discuss your concerns (or your patrons' concerns) with the authenticity of the materials targeted for the archive. If authenticity is an issue, how will it be evaluated? Who is responsible for evaluating or certifying the authenticity of the materials?

7. The tools used to analyze web-published materials targeted for a web archive can gather data about various characteristics of the targeted materials: for example, name, size, format, levels within a web site, number and targets of external links.

    Discuss how you might use this data as you evaluate targeted materials in the following areas.

    a. Consistency of materials with the archive's subject area
    b. Conformance to policies regarding data type and format
    c. Storage requirements
    d. Presentation requirements
    e. Human resource requirements
    f. Hardware & software requirements

TOPIC 3: ACQUISITION OF WEB-PUBLISHED MATERIALS

> Notes to Facilitator
>
> Agreements and licensing arrangements should be in place prior to materials being acquired. Web-published materials are acquired or ‘harvested’ using crawling tools, which either globally capture or selectively capture materials.
>
> The purpose of these questions is to identify agreements that need to be in place prior to material acquisition and to identify the features of a web crawler that will result in the meaningful capture of web-published materials.

1. What agreements, contracts, subscriptions, copyright clearances, or licensing arrangements are required prior to the acquisition of your web-published materials?

2. During and after acquisition of materials, a web crawler can provide information that might be useful in deciding whether to archive the materials acquired. Imagine that you are monitoring the acquisition of materials, either as they are harvesting or subsequent to the harvest. Discuss how each of the following characteristics of web materials will help you determine whether or not they should be archived?

   a. File Names
   b. Material Types (image, audio, video, etc.)
   c. Material Formats (html, jpeg, gif, pdf, etc.)
   d. File Size by Type
   e. Others

3. Many web-published resources change frequently. Your archive will likely involve updating the materials? What will trigger the re-harvesting of materials for your planned archive?

> Notes to Facilitator
>
> Suggest the following discussion areas if necessary:
>
> - Frequency or degree of change in source materials
> - Elapse of a pre-defined time period
> - Request or notification from data supplier or web publisher
> - Data based on some metric derived from crawler data

TOPIC 4: DESCRIPTION OF WEB-PUBLISHED MATERIALS

Notes to Facilitator

Baseline metadata is acquired at the time of data capture or harvesting. Enriched metadata is generally specific to an organization and contains a mixture of machine-generated and human-generated data added subsequent to capture.

The purpose of these questions is to identify required metadata categories and to address issues relative to (a) multiple versions of single items harvested over time and (b) multiple formats of a single unit.

1. You've already discussed the unit of selection for your planned archive. What is the unit of description for the web archive, for example, the domain, web site, web page, logical document (multiple web pages), or object? How might the unit of description differ from the unit of selection?

2. There are several types of metadata specified for archived materials: descriptive, technical administrative, structural, and preservation.

   a. Have you considered or identified a standard metadata scheme for your planned web archive? What are the pros and cons of a standard scheme for your archive?

Notes to Facilitator

Most repositories for web archives will use one of the following metadata schemes:

- MARC
- Dublin Core
- METS
- OAIS

   b. Considering the unit of description, how might crawler-captured data be used as metadata?

Notes to Facilitator

If necessary suggest the following data elements possibly resulting from a harvest:

- URL
- Date/Time of Capture
- Object Format/Type
- Language
- File Size
- File Name
- Author
- Title

   c. What additional metadata elements might facilitate curation of digital materials?

3. Will it be necessary for metadata elements to be added through post-harvest analysis of the archived materials? Consider and discuss:

    a. Subject identification via automated full-text analysis of web pages

    b. Intellectual or human analysis of the archived materials

4. What are the trade-offs between enhanced metadata and the effort required to produce it?

    a. Is the effort worth it?

    b. Does your organization have the resources for the effort?

5. It is likely that multiple versions of archived materials will be harvested over time. Should multiple versions be treated as separate items (i.e., have separate metadata records) or should they be described within a single metadata record? Discuss the reasons for your preference.

6. It is also likely that multiple formats of materials might be harvested. Should multiple formats of a given item be treated as separate items (i.e., have separate metadata records) or should they be described within a single metadata record? Discuss the reasons for your preference.

TOPIC 5: ORGANIZATION OF WEB-PUBLISHED MATERIALS

Notes to Facilitator

Web archives are typically organized to mirror the web at the time of capture. Post-capture, they may be further organized, for example, by subject. An example of this type of organization is the BUBL Information Service, http://bubl.ac.uk

The purpose of these questions is to determine the requirements for the organization of the web archive based on users' needs. How the archive is organized will relate to how end users can access or search the archives' contents. The following are likely search criteria:

- URL
- Date/Time of Capture
- Object Format/Type
- Language
- File Size
- File Name
- Author or Publisher
- Title
- Original Publication Date
- Subject
- Full Text

1. How will your patrons or end users expect to interact with the web?

2. What search criteria should a keyword search cover?

3. What would be the minimum acceptable search criteria for advanced searches?

4. How important is it to classify materials in a web archive based on some classification system? How feasible is this?

TOPIC 6: PRESENTATION OF WEB-PUBLISHED MATERIALS

Notes to Facilitator

Presentation of web archive materials is related to their capture and organization. In general, archives of web-published materials either (a) mirror the web experience of the materials at the time of capture or (b) are re-presented in accord with indexing functionality, which is typically by subject or topic, by author, or by title.

The purpose of these questions is to identify the requirements users have for locating, retrieving, viewing, and interacting with materials in the web archive.

1. Consider information publisher (content producer) constraints that might impact the presentation of materials in the web archive?

   a. What access restrictions exist?

   b. How are these restrictions enforced and monitored? What authentication mechanisms are needed?

   c. How might access requirements differ amongst varying materials in the archive?

2. How will the results of archive searches be displayed?

Notes to Facilitator

If necessary suggest the following display options for the results of a search or query of the archive:

- Search result list with hyperlinks
- Brief metadata record
- Expanded metadata record

3. Some active links within web pages will no longer be active in a web archive. How might you implement custom 404 error messages to deal with the following?

   a. Hyperlinks to non-archived materials within the site

   b. Hyperlinks to non-archived materials external to the site

Notes to Facilitator

404 errors occur whenever a user requests a nonexistent or non-archived web page. The default error message from the web server is generally something like: "Not found: The requested URL was not found on this server."

These messages might be customized by the web archive to provide end users with more useful information regarding the linked URL, for example, user-friendly guidance on how to locate the resource external to the archive or possibly an explanation of how come the URL content does not exist within the web archive.

4. Various active elements within web pages will no longer be active in a web archive. Consider how your presentation of materials will deal with the following?

   a. Non-functional ‘mailto’ hyperlinks

   b. Non-functional interactive forms

5. Discuss the importance of presenting end users with multiple formats of the same object (e.g., an image in bmp, tif, or jpeg format)? How important in their selection process is this?

6. How will users assess the authenticity of materials retrieved from the archive?

7. How will software and hardware requirements for viewing or interacting with the materials in the archive be presented and made accessible?

TOPIC 7: MAINTENANCE OF WEB-PUBLISHED MATERIALS

Notes to Facilitator

Several maintenance functions are critical to ensuring the successful use of materials in web archives: software and hardware training for archive support staff and end users; hardware and software maintenance, performance optimization, backups, and upgrades; and duplicate detection.

The purpose of these questions is to identify maintenance issues and the information needed for librarians and curators to successfully carry out web archive maintenance activities.

1. As web archives age and hardware and software platforms change, it may be necessary to maintain expertise on older platforms in order to train your end users on the older software and hardware platforms. How feasible is this for your organization?

2. All web archives require storage as well as the hardware and software to maintain the archive. Describe the maintenance challenges to your organization in the following areas:

   a. System security
   b. Backups
   c. Software upgrades
   d. Hardware upgrades
   e. Performance optimization

3. It is anticipated that the process of harvesting materials to build a web archive will result in duplication of some materials. Likewise, subsequent harvests will likely yield duplicates of existing materials. Discuss the trade-offs of duplicate detection and duplicate storage.

4. Maintenance of a web archive occurs during each harvest as the crawler reports data about the web sites it is harvesting. Identification of the data elements a curator needs to review for an in-progress crawl prior to admitting materials into the web archive is important. Additionally, data available subsequent to a crawl may assist librarians and curators in their maintenance activities.

   Discuss and come to a consensus regarding the importance of each of the following attributes for post-harvest maintenance activities of completed crawls.

URLs				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Number of Files				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Total Capture Size				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Average File Size by Type				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5

Type (image, audio, video, etc.)				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Format (html, jpeg, gif, pdf, etc.)				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Errors Encountered (200, 300, 400, 404, 500, etc.)				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5
Rate of Change by Selected Criteria				
Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
1	2	3	4	5

5. What additional attributes would help with ongoing web archive maintenance activities?

Other ________________________________

Other ________________________________

Other: ________________________________

Other: ________________________________

TOPIC 8: DESELECTION OF WEB-PUBLISHED MATERIALS

> Notes to Facilitator
>
> Removal of web archive materials can be for several reasons: duplication, errors, legality, or nature of the materials (e.g., offensive materials). Risks of removal and retention need to be weighed against storage costs.
>
> The purpose of these questions is to identify the issues involved in (a) deselection of materials from a web archive, (b) the mechanisms for deselection, (c) the methods for deselection, and (d) users' criteria for deselection.

1. How reasonable is it to consider the unit of selection to be the unit of deselection? For example, if the unit of selection and description within an archive is a web page, what issues emerge when a decision is made to deselect a particular image from a page?

> Notes to Facilitator
>
> If necessary, prompt with the following questions.
>
> - Is the authenticity of the source material compromised?
> - Should agreements with information publishers or data providers address deselection?
> - Would the page be altered or tagged in some way to alert users to the modification?
> - How might your end users react?

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

   a. Is it critical to your end users to retain each format type?

   b. How do collection policies need to address the concept of multiple formats?

3. If an archive is created from source materials that are frequently modified and the archive intends to capture versions of the source materials over time, then periodic harvests of web materials will be conducted. Considering your end users and their needs, how accessible do multiple versions of materials in the web archive need to be? What are the implications of meeting your users' needs for deselection and storage?

> Notes to Facilitator
>
> If necessary, prompt with the following possible answers:
>
> - My users need access all harvested updates
> - My users need only the most current update
> - My users need access to various versions based their unique criteria

4. When a decision is made to remove objects or materials or entire web sites from a web archive, how should the deselected objects or materials be handled? Is it feasible to store them remotely? What are the issues that emerge and their implications for your organization?

TOPIC 9: PRESERVATION OF WEB-PUBLISHED MATERIALS

> Notes to Facilitator
>
> Preservation challenges are numerous. They include persistent naming, format migration and/or emulation, inventory management, volatility, and storage.
>
> The purpose of these questions is to identify the needs of your end users regarding their expectations for interacting with source materials in the archive and to identify the implications for preservation activities.

1. Web archives will likely consist of materials harvested from particular web locations at different points in time. How should materials be named within an archive? What naming issues emerge?

> Notes to Facilitator
>
> If necessary, prompt with the following questions.
>
> - How critical are unique object identifiers or persistent names? For example, should a web page or image harvested on different occasions have a unique name associated with it for each harvest?
> - What if there was no change in the object from time to time?
> - Will users expect to identify all versions of an object in the archive from a single name search?

2. Your end users may expect their interactions with the web archive to mirror their interactions with the source web sites, that is, they may expect the archive to present the same 'look and feel' and behave in the identical fashion. Obsolescence of both hardware and software platforms presents challenges to satisfying this end user expectation.

   Currently, there are two primary ways in which this "mirroring" is being achieved: emulation and migration. Discuss each one and identify any known or potential problems they present for your end users. What are the implications for preservation of the web archive?

> Notes to Facilitator
>
> Definitions if needed.
>
> - Emulation: Original end user experience is preserved on new platforms that provide access to original materials by emulating older platforms
>
> - Migration: Materials are recreated as necessary for presentation on new hardware and software platforms

3. How will hardware and software platforms be preserved to enable presentation of materials over time? What issues or needs emerge if this direction is chosen?

4. Inventory management is an important preservation function. What is the scope and definition of inventory management for a web archive?

5. Usage reports can assist in identifying demand for materials in a collection. In a web archive, what would be expected in a usage report? How do you think this information would be helpful in conducting other web archive functions, such as format migration and/or emulation, presentation, maintenance, deselection, and storage decisions?

6. Replication of archives in different geographic locations is proposed as one method to ensure an archive remains available, stable, and trustworthy. What do you consider the benefits or potential downsides of this method?

Appendix 9. Focus Group 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!

Appendix 10. Letter of Invitation to End Users

Dear [*Potential Participant*]:

[*Library name or institution name*] is planning to create an archive of web sites of importance to our faculty and students. I would like to meet with you for an hour or two to discuss your ideas about archiving web-published materials for future use by researchers like you.

Web-published materials include a range of material types from text documents to streaming video to interactive experiences. These materials are accessed and presented via the World Wide Web and because they are both dynamic and transient, they are at risk of disappearing. Web archives preserve web-published materials. Some of these materials may also exist in other forms but the web archive captures the web versions for posterity and may provide users access to the materials long after the web sites themselves are no longer in existence.

We are targeting researchers in the areas of political science and history [*insert appropriate department or school within the institution*] who might use web sites of US federal and state government agencies in their research and professional activities. Our archive might also include web sites of political policy documents, campaign literature, or information surrounding political movements.

We are creating our web archive under the auspices of the Web-at-Risk project, which is one of eight digital preservation projects funded in 2004 by the Library of Congress. Each of the eight projects represents a collaborative effort to preserve born-digital or digitized cultural heritage materials and collections for future generations. We are working with a great team of experts from the California Digital Library, the University of North Texas, and New York University.

The needs and issues you articulate during our discussion will help us create guidelines and make choices for our web archive. I hope you are interested in taking part in this interesting and important activity. Your input will help ensure the creation of useful tools and guidelines for our web archive.

To get an idea of what web archives are like, take a look at these two examples:

CyberCemetery http://govinfo.library.unt.edu/	UCLA Online Campaign Literature Archive http://digital.library.ucla.edu/campaign/

Be assured your participation is completely voluntary and, while a list of participants may be published, comments will not be attributed to individuals in a manner that they can be recognized.

If you would like to contribute your ideas and experiences, please contact me at [*email, telephone number*] by [*date*]. I will get back with you to set up a convenient meeting time.

Thanks very much for your consideration!

[*Name of Interviewer*]

Appendix 11. Letter of Confirmation to End Users

Dear [*Participant*]:

Thank you very much for agreeing to meet with me. I think you will find it stimulating and your contributions will be greatly appreciated. Here are the details for our meeting.

Date	[*Insert*]
Time	[*Insert*]
Place	[*Insert*]

We'll talk about what web materials you and others in your discipline currently use. Our discussion will focus on how you would like to search and interact with a collection of web-published materials, or a web archive. We'll also touch on authenticity and preservation issues that directly affect the usefulness of archived materials.

Here are a few questions to stimulate your thinking prior to our meeting.

> What web-published resources do you currently use for research and professional interests (e.g., key sites, domains, databases)?
>
> What experiences have you had with Web resources disappearing? What were the consequences in your work?
>
> If an archive of web sites were available for your research, how would you ideally locate materials in the archive?

If you are unable to attend the meeting, please let me know as soon as possible [*email, telephone number*]. Feel free to suggest a more convenient time to meet.

I look forward to meeting with you!

[*Name of Interviewer*]

Appendix 12. Interviewer Tips

Introduction

Demeanor

One key to conducting a successful interview is neutrality. Your presence as an interviewer, even over the telephone, should not affect the participant's perception of questions or influence their responses. Be pleasant and appear curious and genuinely interested in what the participant has to say.

Questionnaire

Be familiar with the interview questionnaire and be prepared to adapt questions in a particular interview. Likewise be prepared to pursue the new directions that a participant identifies. Record responses as stated by the participant. Do not attempt to interpret responses or draw conclusions during the interview. Remember that each interview will contribute responses that, when compiled and analyzed, will yield a picture of participants' needs regarding web archives.

If you do not clearly understand what the participant is trying to communicate, ask them to restate their point or to give you an example that illustrates their point. If you remain somewhat unclear, try paraphrasing what you heard and ask the participant if that was their message or point.

Probing

If you judge that a response to an open-ended question is too minimal or general to be helpful, it is a good idea to attempt to elicit a more elaborate response. The first option in doing so is to remain silent. Often the interviewee will voluntary fill in the silence with an enhanced response. Minimal probes are also a good way to elicit more in-depth responses. Try the following probes:

- "Anything else?"
- "Tell me more."
- "How is that?"
- "In what ways?"

Appendix 13. End User Interview Questionnaire

Notes to Interviewer

Start the interview by (a) creating a comfortable atmosphere in which people feel valued for their participation and (b) establishing the context for the discussion. Be prepared to offer definitions of key concepts and examples as outlined in the boxes below.

<u>Opening</u>

"When a library creates a web archive it is important to ensure that user or patron needs inform their plans. The purpose of this discussion is to elicit your needs and thoughts regarding web archives. Our discussion will address needs and issues in the major phases of web archive creation with a particular emphasis on the selection of web-published materials and on users' needs for access to these materials."

<u>Key Concepts</u>

Web-Published Materials

Web-published materials are accessed and presented via the World Wide Web. The materials span the cultural heritage spectrum and include a range of material types from text documents to streaming video to interactive experiences. Web-published materials are both dynamic and transient. They are at risk of disappearing. Web archives preserve web-published materials.

Web Archive

A web archive is a collection of digitized or born-digital materials, many of which are initially made available on the web. Some of these materials may also exist in other forms but the web archive captures the digital versions for posterity and may provide users access to the materials long after the web sites themselves are no longer in existence.

<u>Digital Archive Examples</u>

CyberCemetery
http://govinfo.library.unt.edu/

"The University of North Texas Libraries and the U.S. Government Printing Office, as part of the Federal Depository Library Program, created a partnership to provide permanent public access to the Web sites and publications of defunct U.S. government agencies and commissions. This collection was named the "CyberCemetery" by early users of the site."

UCLA Online Campaign Literature Archive
http://digital.library.ucla.edu/campaign/

“Every American election produces thousands of campaign flyers, pamphlets, posters, and bumper stickers, generally called "campaign literature." These documents provide an important record of the campaign, its participants, issues, and tactics. The UCLA Online Campaign Literature Archive presents a subset of the materials in the complete Campaign Literature Collection. It contains copies of all archived websites plus scanned images of selected print materials. (Currently, this includes all elections from 1908 to 1939, and some materials from 1940 and 2000.) These are made available on the web for the use of scholars and students across the world.”

Opening. Participant Background & Need for Digital Archives

Facilitator:

Indicate which category best describes where the participant works/studies.

a. _______ K-12 School
b. _______ College or University
c. _______ Federally Funded Institution
d. _______ State Government Institution
e. _______ Local Government Institution
f. _______ Non-Profit Organization
g. _______ Corporate Institution
h. _______ Specify Other: ____________________________

1. What is the name of your department?

____

2. What is your current position, academic status, or title?

____

3. How many years have you been in this position? ______________

4. Do you use web-published materials in your:

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

5. 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?

____

____

____

____

6. 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					

	# Years				
	< 1	1-3	5	10	>10
Unpublished Work & Publications of Limited Circulation					
Other:					
Other:					
Other:					
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

20. 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?

____

____

____

21. 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.

22. 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 14. Research Consent Script for Telephone Interviews

Instruction to Researcher: At the beginning of telephone interviews, it is important that the subject is read the following explanation of the study and their verbal consent to participate is obtained and documented on this form by the interviewer.

I am [*interviewer*], [*interviewer's title*] at [*institution's name*].

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 collections of web-published materials. One important assessment goal in the project's Content Identification, Selection, and Acquisition (CISA) path is to identify the needs and concerns of web content producers, that is, organizations like yours who publish or supply the materials for web-based collections. It is expected that the needs and concerns identified will translate into requirements for the Web Archiving Service.

The interview is expected to take approximately one hour. Interview notes or transcripts will be accessible only to researchers and analysts. While a list of participants may be published, comments will not be attributed to individuals in a manner that they can be recognized.

This research study has been reviewed and approved by the Institutional Review Board at [*institution's name*]. The IRB can be contacted at [*telephone number*] or [*email*] with any questions regarding the rights of research subjects.

Your participation in this interview is voluntary and you can refuse to answer any questions or terminate the interview at any time without penalty. In case you have any questions about the study, please contact [*interviewer or researcher*], [*title or position*] at [*institution's name*], at [*email address*] or [*telephone number*].

Do you agree to participate in this interview? _____ Yes _____ No

For the Investigator or Designee: I certify that I have reviewed the contents of this form with the subject being interviewed. I have explained the known benefits and risks of the research. It is my opinion that the subject understood the explanation.

Investigator Signature _________________________________ Date _______________

Appendix 15. Letter of Invitation to Content Producers

Dear [*Potential Participant*]:

I would like to invite you to participate in a web-archive research study. Specifically, I would like to schedule a 1-hour telephone interview with you to discuss the needs and concerns web content producers have regarding web archives.

The study is being conducted under the auspices of the Web-at-Risk project, which is one of eight digital preservation projects funded in 2004 by the Library of Congress. The Web-at-Risk 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 library curators to build 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 surrounding political movements. Two examples of such web archives are:

CyberCemetery
http://govinfo.library.unt.edu/

UCLA Online Campaign Literature Archive
http://digital.library.ucla.edu/campaign/

One important goal of the project is to identify the needs and concerns of web content producers, that is, organizations like yours who publish or supply the materials for web-based collections. The needs and issues you articulate during our discussion will help us create guidelines and tools for the Web Archive Service.

I hope you are interested in taking part in this interesting and important study. Your input will help ensure the creation of useful tools and guidelines for our web archive service.

Be assured your participation is completely voluntary and, while a list of participants may be published, comments will not be attributed to individuals in a manner that they can be recognized.

If you would like to contribute your ideas and experiences, please contact me at [*email & telephone number*] by [*date*]. I will get back with you to set up a convenient meeting time.

Thanks very much for your consideration!

[*Name of Interviewer*]

Appendix 16. Letter of Confirmation to Content Producers

Dear [*Participant*]:

Thank you very much for agreeing to talk with me about the needs and concerns web content producers have regarding web archives. Your contributions will be greatly appreciated. Please reserve [*hour time frame*] for our discussion on [*day, date*]. I will call you at [*telephone number*].

The purpose of this interview is to explore the issues information publishers or content producers have regarding web archives. Our discussion will focus on the types of materials you publish and your ideas about archiving those materials in a third-party repository. We'll also touch on access, authentication, authenticity, and intellectual property issues that directly affect the creation of web archives.

Here are a few questions to stimulate your thinking prior to our discussion.

> What types of web materials do you currently produce? How frequently do these materials change?
>
> What type of authentication do you require for access to your materials?
>
> How should the following issues be addressed in agreements between content producers (information producers or publishers) and web archive providers?
>
> - Support and maintenance
> - Copyright and intellectual property

If you need to reschedule the meeting, please let me know as soon as possible [*email, telephone number*]. Feel free to suggest a more convenient time.

I look forward to talking with you!

[*Name of Interviewer*]

Appendix 17. Content Producer Interview Questionnaire

Notes to Interviewer

The purpose of this interview is to explore the issues information publishers or content producers have regarding web archives. Start the interview by (a) creating a comfortable atmosphere in which the person feels valued for their participation and (b) establishing the context for the discussion.

Opening

“When a library creates a web archive it is important to ensure that agreements are in place with web publishers or content providers. The purpose of this discussion is to elicit your needs and thoughts regarding web archives of your materials created by a third party, such as a university library.”

Topic 1. Materials

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

____

____

____

2. What types of materials are included?

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

____

____

____

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

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

Topic 2. Digital Archives

4. Describe your experience with digital archive providers.

____

____

____

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

____

____

____

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

____

____

____

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

____

____

____

Topic 3. Access to Materials

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

a. Publicly available

____

____

b. Access restrictions exist

____

____

c. Fair use guidelines applicable

____

____

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

____

____

____

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

a. To harvest your materials?

____

____

b. For end user access to your materials?

____

____

Topic 4. Authenticity of Archived Materials

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

____

____

____

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

____

____

____

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

____

____

____

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

a. Removal of an object that is linked into a web page

____

____

b. Deactivation of an email or hyperlink because it links outside of the archive

____

____

c. Are there other situations you can think of where modification of the source materials and alerting the end users to that modification is acceptable?

____

____

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

____

____

____

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

____

____

____

Topic 5. Intellectual Property of Archived Materials

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

____

____

____

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

____

____

____

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

____

____

____

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

____

____

____

Topic 6. Agreements with Archive Providers

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

____

____

____

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

____

____

____

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

a. Specifications of the data (materials) to be archived

____

____

b. Minimum metadata to be provided with the data (e.g., description of structure and meaning of data sets)

____

____

c. Data delivery specifications: protocols, verification procedures

____

____

d. Support and maintenance

____

____

e. Copyright and intellectual property

____

____

Closing

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

____

____

____