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

Content Identification, Selection, and Acquisition Path

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.

Web Archive Creation Process: Web Published Materials

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?