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 (page 20). 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.)

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, w ki, 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 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.