



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 Report

January 5, 2006

Prepared by:

Inga Hsieh
University of North Texas
ikh0003@unt.edu

Kathleen Murray
University of North Texas
krmurray@unt.edu

With contributions from:

Cathy Hartman
Samantha Hastings
William Moen

Special thanks are extended to the project's curators
for their assistance in completing the survey.

Contents

1 Introduction 6

2 Methods 6

 2.1 Purpose 6

 2.2 Participants..... 7

 2.3 Survey Instrument Development..... 7

 2.4 Online Survey Development 7

 2.5 Institutional Review Board for the Protection of Human Subjects 8

 2.6 Data Collection 8

 2.7 Data Analysis 8

3 Key Results 9

 3.1 Respondents and Digital Collections 9

 3.2 Collection Policies 11

 3.3 Selection of Web Materials 15

 3.4 Curation of Web Collections..... 19

 3.5 Preservation of Web Collections 21

 3.6 Web Archiving Service Requirements..... 23

4 Discussion 29

 4.1 Building Web Collections 29

 4.2 Web Archiving Service Requirements..... 31

 4.3 Closing 32

Appendix A. Collection Development Framework for Web Archives..... 33

Appendix B. Survey Participants 34

Appendix C. Survey Instrument 36

Appendix D. Glossary 55

Appendix E. Letter of Consent..... 58

Appendix F. Survey Results 59

 Section A. About Your Collections 59

 Section B. Selection: Policy, Identification, & Acquisition..... 64

 Section C. Curation: Description, Organization, Presentation, Maintenance, & Deselection.... 80

Section D. Preservation 83

Section E. Curator User Interface 85

Appendix G. Respondents' Most Important Digital Collections 95

Table of Figures

Figure 1 - Organizational Support & Web Archive Creation (Q5 & Q11) 13

Figure 2 - Financial Challenges in Building Web Collections (Q27)..... 13

Figure 3 - Technical Challenges in Building Web Collections (Q28)..... 14

Figure 4 - User Acceptance of Privacy and Technical Issues (Q12 & Q13) 15

Figure 5 - Planned Level of Selection (Q14) 16

Figure 6 - Commercial and Foreign Material Sources (Q15 & Q17) 16

Figure 7 - Authenticity Concerns for Archived Materials (Q24 & Q25)..... 18

Figure 8 - Methods of Searching Web Archives (Q39 & Q40) 19

Figure 9 - End User Understanding of Deselection Criteria (Q35 & Q37-39) 21

Figure 10 - User Expectations of Preservation Practices (Q40 - 43) 22

Figure 11 - Level of Threat to Authenticity by Migration Type (Q44) 23

Table of Tables

Table 1 - Percentage of Web-published Materials in Digital Collections (Q4) 10

Table 2 - Digital Formats for Material Types Included in Policies (Q8) 11

Table 3 - Acceptable Digital Formats (Q9) 12

Table 4 - Intellectual Property Considerations (Q19) 17

Table 5 - Deselection Criteria and End User Understanding (Q33 & Q35-39) 21

Table 6 - Importance of Crawl Attributes in Selection Decisions (Q45) 24

Table 7 - Importance of Crawl Definition Parameters (Q47) 25

Table 8 - Importance of Realtime Data Reporting During Crawls (Q50) 25

Table 9 - Importance of Collection-Level Attributes for Collection Management (Q52) 26

Table 10 - Importance of Object-Level Attributes for Collection Management (Q54) 27

Table 11 - Desired Level of Descriptive Metadata (Q56) 27

Table 12 - Importance of Attributes as End User Access Points (Q57) 28

1 Introduction

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

The *Needs Assessment Toolkit*¹ created for the Web-at-Risk project describes the project's needs assessment activities and includes data collection tools, which are designed to identify the needs and requirements of curators, web-content producers, and end users with regard to the Web Archive Service. Additionally, information gathered by some of the data collection tools will help to identify curators' requirements for the web crawler and its crawl analyzer tool, which will be developed as part of the project.

Each of the assessment activities described in the *Needs Assessment Toolkit* was designed to follow a collection development framework for web archives. (See Appendix A.) This report contains a data analysis of the survey results. Results from focus group discussions and interviews with content providers and end users are presented in separate reports.

The remainder of this report includes:

- Methods – design, implementation and execution of the survey
- Results – description of significant results, including figures and tables
- Discussion – conclusions and questions from results
- Appendices – framework, participants, survey, glossary, detailed results

2 Methods

2.1 Purpose

The web-based needs assessment survey served two purposes: (a) to identify end user and curator needs that might impact collection development for web archives and (b) to identify functional requirements for the Web Archiving Service's crawler and associated tools in the areas 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

¹ Murray, K. R. (2005, May 31). *Needs Assessment Toolkit: Guidelines & Data Collection Tools*. Retrieved December 6, 2005, from the University of North Texas Web-at-Risk Project Web site: http://web2.unt.edu/webatrisk/na_toolkit/deliverable_na_toolkit_final_krm_31may2005.pdf

2.2 Participants

Survey respondents were the 22 curatorial partners involved in the Web-at-Risk project. All curators volunteered to participate and many are or will be involved in building web collections for the project.

In all, 16 surveys were submitted. Ten curators submitted individual surveys while 12 curators submitted a total of six surveys, each of which represented a joint effort between two curators. (See Appendix B.)

In answering survey questions, curators served a dual role, representing end user needs in addition to their own. Most curators collaborated with other professionals at their institutions or organizations to obtain the information necessary to complete the survey.

2.3 Survey Instrument Development

The survey instrument was created by project team members at the University of North Texas (UNT) and reviewed for clarity and comprehensiveness by project principals from UNT, as well as the California Digital Library (CDL) and New York University (NYU). The survey instrument was revised based on their feedback and subsequently implemented in a web-based format.

The survey consisted of 58 questions divided into five sections and addressing the following areas:

- Section A. Respondents' Background & Collections
- Section B. Selection Needs: Policy, Identification and Acquisition
- Section C. Curation Needs: Description, Organization, Presentation, Maintenance and Deselection
- Section D. Preservation Needs
- Section E. Curator User Interface Requirements

Curators either selected responses from a pre-defined list of possible answers or entered free-form textual answers. At the end of the survey, respondents were asked to provide any additional comments they would like. Appendix C contains the final survey instrument.

2.4 Online Survey Development

The web-based version of the survey was implemented using HTML, PHP, and MySQL. Participants used their standard web-browsers to complete the online survey. To enhance navigation of the online survey, the survey instrument's five sections were further sub-divided to create eight sections. Each sub-section was preceded by a brief introductory paragraph describing the context for the questions that followed.

Client-side JavaScript was used to provide an interactive glossary of terms and definitions. When respondents selected terms, definitions were displayed in a separate frame at the bottom of the browser window. Appendix D contains the glossary from the online survey.

Project curators at UNT tested the online survey instrument. These curators met with the survey designers and navigated the online survey instrument while commenting on question clarity, survey layout, and usability. Feedback from these tests was incorporated into the final version of the online survey, which was made available to participants on June 26, 2005 with a requested completion date of July 15, 2005.

2.5 Institutional Review Board for the Protection of Human Subjects

Although survey completion involved no risk to participants, approval was obtained from UNT's Institutional Review Board for the Protection of Human Subjects (IRB) in accordance with UNT policy prior to making the survey available. Participants were presented with a letter of consent before beginning the survey and were instructed to close their browser window if they did not want to participate. See Appendix E to view the consent letter.

2.6 Data Collection

Each curator was assigned a user name and password to access the online survey. Upon accessing the survey Web site, participants were advised to print a hardcopy of the survey instrument to review, as necessary, with their colleagues before completing the survey online.

Prior to logging in participants were presented the consent letter. If users agreed to the terms of the survey as described in the consent letter, they were presented with a login screen. After logging in to the survey, they were presented with a second opportunity to print a hardcopy of the survey instrument as well as the opportunity to print a hardcopy of the glossary of terms used in the survey. Proceeding from this screen took the participants to the first section of survey questions.

Upon submission of each section, responses were stored in a MySQL database. If a participant was forced to abandon the survey for technical or other reasons, he or she could reenter the survey at a later time and would be positioned at the beginning of the last unsubmitted section. Participants were not permitted to re-access any submitted survey section.

2.7 Data Analysis

Questions in each section of the survey were first analyzed individually. Where appropriate, response sets were removed prior to analysis. For the most part, descriptive statistics (e.g., numbers and percentages of responses) were used to analyze the data.

Due to the small number of respondents and the categorical nature of most of the data, statistical calculations were used infrequently. In a few cases, Spearman's Rho was calculated to evaluate the relationships between responses to two questions. A significance level of .05 was required in each case.

3 Key Results

This section reports the key results of the needs assessment survey. Detailed descriptive data for each question are included in Appendix F.² Survey results are presented in the following order, which essentially corresponds to the structure of the survey instrument.

- 3.1 Respondents' and Digital Collections
- 3.2 Collection Policies
- 3.3 Selection of Web Materials
- 3.4 Curation of Web Collections
- 3.5 Preservation of Web Collections
- 3.6 Crawler Interface Requirements

Symbols used throughout this report include:

- Q_n - Question number n of the survey
- N - Total number of responses
- n - Total number of responses when a subset of responses is examined
- M - Mean
- r_s - Spearman's Rho correlation coefficient
- p - Significance

3.1 Respondents and Digital Collections

Section A of the survey obtained background data regarding the respondents, their collections, and their experience with digital archives. This data provides a context for interpreting the survey results.

3.1.1 Characterization of Respondents

The survey respondents were the curatorial partners involved in the Web-at-Risk project. The majority work in academic libraries, while one curator works in a state library. Their collections concern a range of materials, including federal, state and local information, international materials, labor and policy information, and other resources for the social sciences (Q1). End users of the respondents' collections include community members, university students and faculty, government and non-government agencies, and lawyers and other professionals (Q2).

3.1.2 Existing Digital Collections

Slightly more than half (56%) of the respondents ($N=16$) indicated that they currently maintain digital collections. The digital collections respondents considered most important are listed in Appendix G (Q3).

In order to identify the types of web-published³ materials respondents were already collecting, they were asked to estimate the percentage of various material types in their most important digital collections that were web-published (i.e., in those collections identified in Q3).⁴ Nearly half (44%) of the respondents ($N=9$) indicated that more than 75% of the government documents in these collections were web-published. All respondents ($N=8$) indicated that less than 25% of the videos in these collections were web-published.

² Reported results reference their corresponding survey question numbers (e.g., Q3 or Q23). Refer to Appendix F for detailed descriptive data for any question.

³ Web-published materials are materials that are accessed and presented via the World Wide Web.

⁴ Based on the responses to Q4 (e.g. the high percentage who selected Journals & Periodicals), respondents may not have limited their responses to their list of 'most important digital collections' from Q3.

When the responses to question four were consolidated into three categories,⁵ the material types most frequently collected from web sources for inclusion in digital collections were: 'Journals & Periodicals,' 'Government Records,' 'Technical & Research Reports,' and 'Proceedings of Meetings & Symposia'. (See Table 1.)

Material Type	0%	1-50%	51-100%
Government Records	11.1	33.3	55.6
Journals & Periodicals	33.3	11.1	55.6
Technical & Research Reports	25.0	25.0	50.0
Proceedings of Meetings & Symposia	37.5	12.5	50.0

Table 1 - Percentage of Web-published Materials in Digital Collections (Q4)

Given that survey respondents work primarily in government documents positions, it is not surprising that the collections identified by the respondents as their most important digital collections (Q3) frequently contain government and research materials, supporting the high-frequency of web-published 'Government Records' and 'Technical & Research Reports' reported in Table 1. The percentage of web-published 'Journals & Periodicals' in respondents' digital collections may be explained by the inclusion of certain government agencies' publications in their digital collections, such as the Texas Register, which is a weekly publication from the Office of the Texas Secretary of State. Lastly, 'Proceedings of Meetings & Symposia' are generally web-published and, because they are likely to be both of value to researchers and at risk of being removed from organizational websites over time, it is not surprising that curators would include these in local digital collections.

3.1.3 Current Digital Archiving Activity

Slightly more than one-third (36%) of the respondents ($N=14$) indicated that they were actively maintaining a digital archive of one or more of their unlicensed digital collections (Q5). The underlying software or management tools used for these archives are (Q6):

- eScholarship Repository⁶
- CONTENTdm⁷
- LOCKSS⁸
- ISL/UIUC SafetyNet Software
- Proprietary systems
- Ad-hoc systems and methods

Those who reported maintaining digital archives were asked to identify the two greatest hurdles they encountered in creating their archives (Q7). Responses fell into these five broad categories:

1. Difficulties getting allocation for staff and finding staff with the appropriate skill set
2. Attracting and sustaining interest in digital archiving projects
3. Technical limitations
4. Metadata
5. Costs

⁵ Old category(-ies) = New category: '0%' = '0%'; '<25%' & '26-50%' = '1-50%'; '51-75%' & '>75%' = '51-100%'

⁶ <http://www.cdlib.org/programs/escholarship.html>

⁷ <http://contentdm.com/> or <http://www.oclc.org/contentdm/default.htm>

⁸ <http://lockss.stanford.edu/>

3.2 Collection Policies

3.2.1 Policies and Practices for Existing Digital Collections

Several questions addressed how current collection policies and practices are shaping the content of existing digital collections. Respondents were asked about the types of materials⁹ specifically included in or excluded from their collections by their institution's collection policies or practices (Q8). The material types specifically included in collection policies or practices roughly fell into the three groups listed in Table 2.

Group	N	Material Types	%
1	15	<ul style="list-style-type: none"> • Journals & Periodicals • Books & Brochures • Government Records • Technical & Research Reports 	60-67%
2	15	<ul style="list-style-type: none"> • Databases • Newspapers • Image Files 	47-53%
3	14-15	<ul style="list-style-type: none"> • Proceedings of Meetings & Symposia • Videos • Unpublished Works & Publications of Limited Circulation • Audio Files • Doctoral Dissertations & Master's Theses 	29-40%

Table 2 - Digital Formats for Material Types Included in Policies (Q8)

Where respondent's policies and practices did not specifically include digital formats¹⁰ of a given material type, respondents generally indicated that inclusion or exclusion of digital formats of that type was not specified by their existing policies and practices. For example, Table 2 shows that approximately two-thirds of the respondents' institutional collection policies and practices specifically included digital formats of 'Government Records'. Although not explicitly stated in Table 2 the reader can assume that most or all of the remaining one-third of the respondents indicated that inclusion or exclusion of digital formats of 'Government Records' was not specified by their existing policies and practices.

Also notable is that slightly more than one-quarter (27%) of the respondents (N=15) indicated that their institution's collection policies and practices do not specify inclusion nor exclusion for digital formats of any of the material types indicated. Conversely, one respondent indicated that their institution's collection policies and practices specifically include digital formats for all indicated material types.

Question nine examined the acceptability of specific digital material formats in respondents' digital collection policies or practices. Table 3 lists the digital formats most often accepted. For each format listed, over half of the respondents indicated the format was acceptable with no limitations.

⁹ 'Material type' refers to the form or genre of the content of a digital object (e.g. journal, image, video, dissertation, etc.).

¹⁰ 'Digital format' refers to the way the contents are encoded for use by a computer and is frequently designated by the extension of a file (e.g. .doc for a Microsoft Word Document).

In general, policies and practices do not exclude specific digital formats. The digital format most often excluded from policies and practices was MacWrite (mw). One third (33%) of the respondents' policies and practices exclude this format.

Digital Format	%	N
Adobe Portable Document Format (pdf)	80	15
Rich Text Format (rtf)	73	15
Images (jpeg, jpg, gif, png, tif)	73	15
Text (ans, txt)	73	15
Web Pages (htm, html, asp, jsp, php)	73	15
Microsoft Excel (xls)	67	15
Microsoft Word (doc)	67	15
Audio (mp3, wav, midi, ra)	64	14
Video (mpeg, ra, mov, rm)	53	15

Table 3 - Acceptable Digital Formats (Q9)

3.2.2 Challenges Presented by Web Collections

Most (81%) respondents (N=16) reported they had at least some support from their organization for creating a web archive. However, three respondents (19%) reported having very little support or no support from their organizations (Q11).

Although not statistically significant ($r_s = .401, p = .16$), the data does suggest a possible relationship between the level of support for web archive creation and maintenance within an organization and the respondents' current archiving efforts the (see Figure 1). Not surprisingly, those organizations with at least some support for creating archives are more likely to engage in this activity than are those with little or no support (Q5 and Q11).

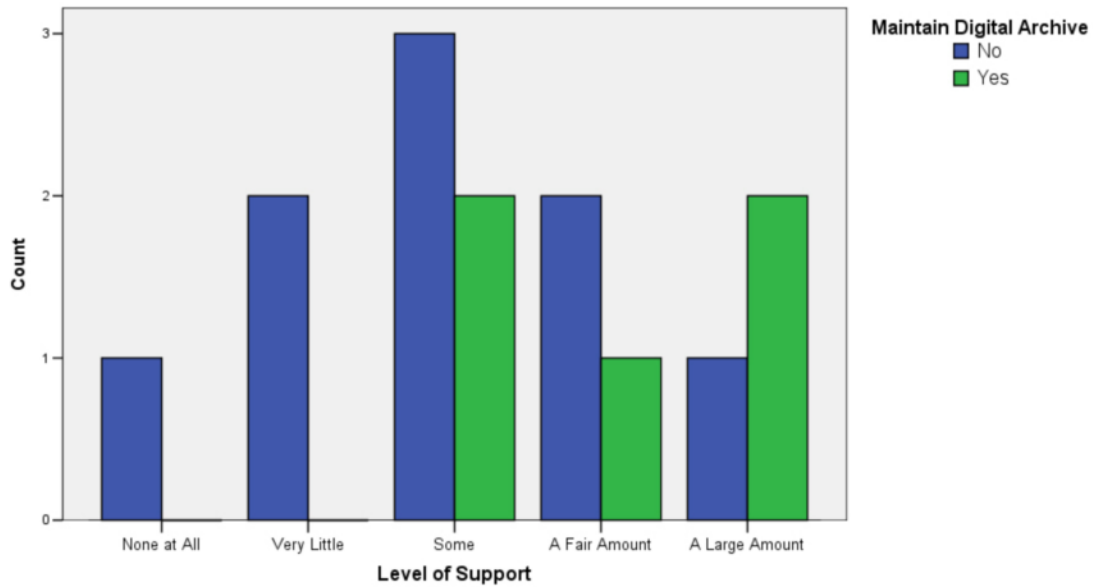


Figure 1 - Organizational Support & Web Archive Creation (Q5 & Q11)

Figure 2 shows the respondents' estimates of the magnitude of the financial challenges they will face when they create or add to their collections for the Web-at-Risk project. The horizontal line indicates a count value of eight, which is one half of the 16 respondents. Over half of the respondents thought the following three areas would be either very or extremely challenging financially: cataloging (75%; N=16), preservation (65%; N=16), and IT support (60%; N=15).

Fifty percent of respondents thought staff training (N=16) would be either very or extremely challenging from a financial perspective and indicated that the initial investment in hardware and software (N=16) to implement their collection would be somewhat challenging. Additionally, over half of the respondents identified needs assessment (67%; N=15) and network access (56%; N=15) as areas that they anticipate posing little financial challenge (Q27).

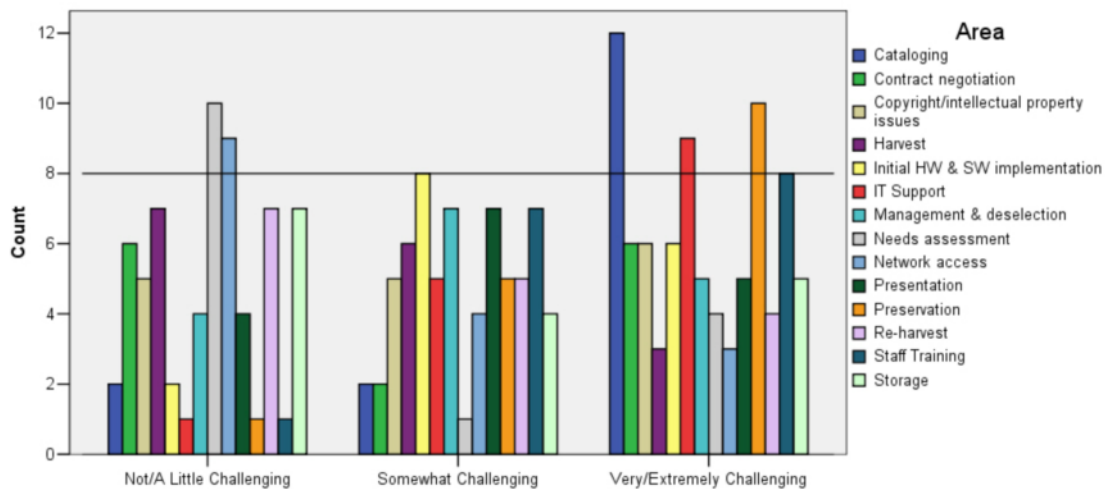


Figure 2 - Financial Challenges in Building Web Collections (Q27)

Similarly, Figure 3 shows the respondents' estimates of the magnitude of the technical challenges they will face when they create or add to their collections for the Web-at-Risk project. The horizontal line indicates a response value of eight, representing one half of the 16 respondents.

At least half of the respondents felt that the following areas would present substantial technical challenges: metadata creation (81%; $N=16$), the dynamic nature of web materials (75%; $N=16$), password-protected source materials (73%; $N=15$), and encrypted source material (69%; $N=16$) (Q28).

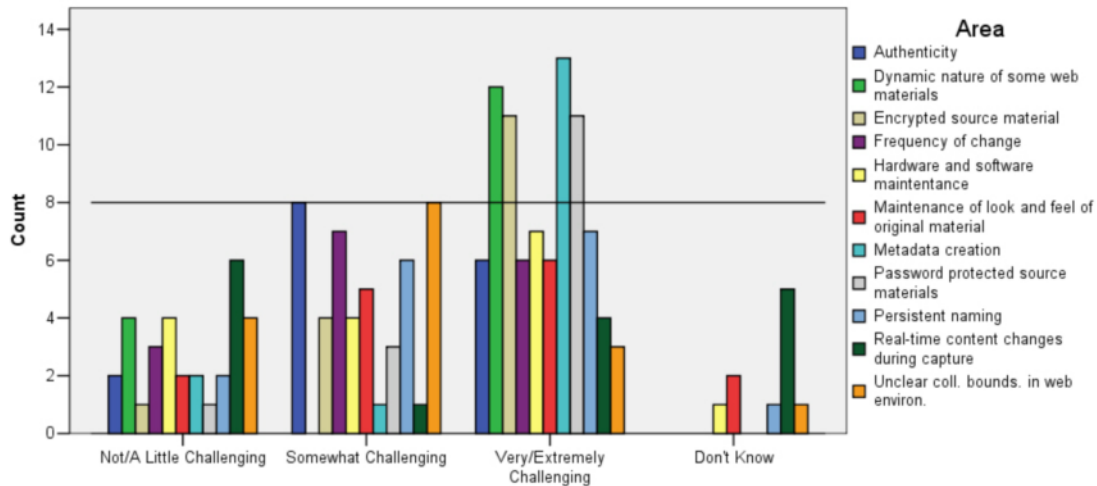


Figure 3 - Technical Challenges in Building Web Collections (Q28)

Recognizing that there would be situations in which materials could not be archived due to privacy issues or technical reasons, we asked respondents about their end users' acceptance of such practices. As Figure 4 illustrates, respondents do not expect their end users to be very accepting of an organization's failure to archive materials for privacy reasons (Q12) or technical roadblocks (Q13).

While 25% of the respondents ($N=16$) did not know how accepting their end users would be of an organizational practice to not archive web sites due to privacy concerns, none of the respondents thought their end users would find privacy issues an extremely acceptable reason for not archiving web sites. Conversely, 56% of respondents expected their end users would be either somewhat, a little, or not at all accepting of this practice. One respondent elaborated on their response to this question via email:

In our opinion, the public will expect to find the information they are looking for without concern for privacy issues (although they might take exception if it is their privacy being violated). ... The public will expect us to find a way to capture the materials. They won't accept excuses of, 'It was password protected.' They'd expect us to find a way around the technical barriers. At the same time, the public often doesn't know what's not been captured, so the issue may never arise.

With regard to technical roadblocks preventing the archiving of web-published materials, two of the 16 respondents were unable to estimate their end users' level of acceptance. The remaining respondents felt that their end users would be somewhat, a little, or not at all accepting of this

practice. None of the respondents thought their end users would be either very accepting or extremely accepting of technical challenges preventing the archival of web sites.

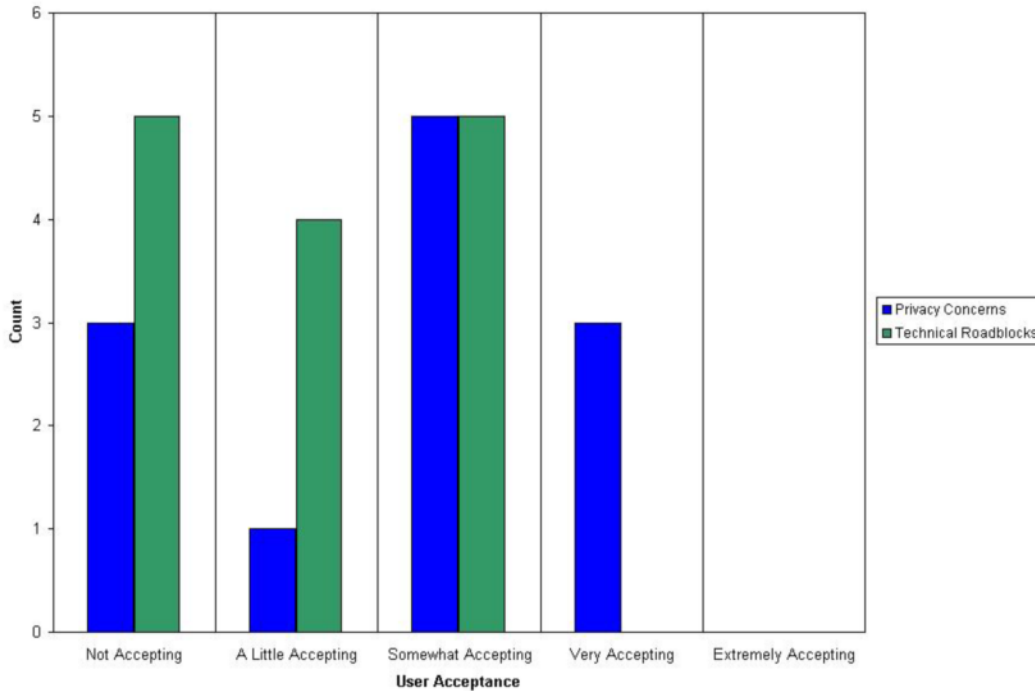


Figure 4 - User Acceptance of Privacy and Technical Issues (Q12 & Q13)

3.3 Selection of Web Materials

To provide a frame of reference for the questions regarding the selection of web materials (Q14-26), respondents were given the following directive: “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.” This forms the context for the analysis of responses in this section of the report.

3.3.1 Level of Selection

For web-published materials, the unit of selection is not obvious. It can vary widely from a single digital object to an entire website or group of sites owned by a single organization. Therefore, respondents were asked about the primary level at which they plan to select source materials for their web collection (Q14). Almost half (44%) of the respondents ($N=16$) indicated they plan to select source materials at the website level. However, 50% of the respondents plan to collect materials at a more granular level, specifically, the logical document level (19%), the web page level (19%), or the object level (13%). (See Figure 5.)

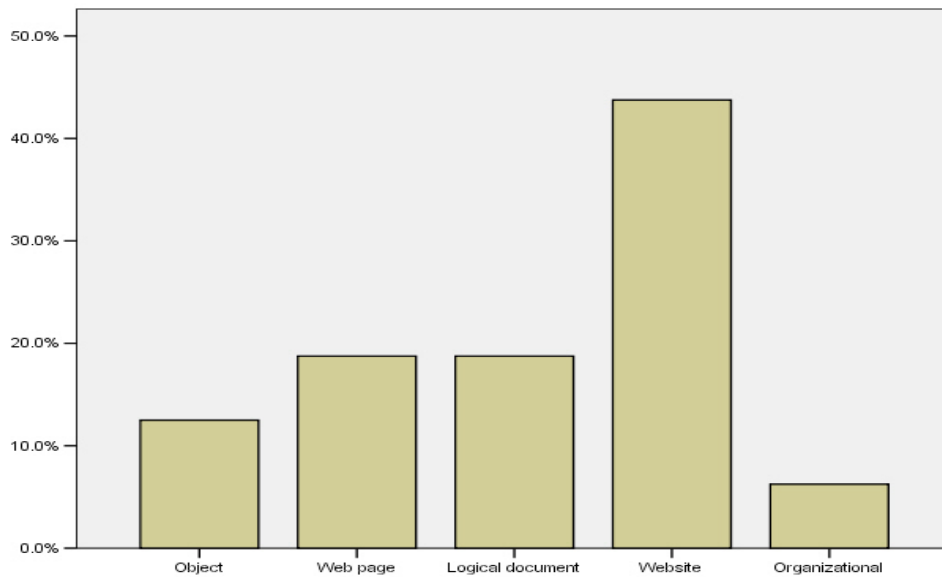


Figure 5 - Planned Level of Selection (Q14)

3.3.2 Material Sources

When asked about the source of the materials they planned to collect, two of the respondents ($N=16$) said that they definitely planned to collect from commercial sources and an additional nine of the respondents (56%) indicated they might collect commercial source materials (Q15). The major reasons for collecting from commercial sources include:

- Materials from media sources which are relevant to the collection
- Agency materials published or co-published by commercial entities
- Reports from think tanks or non-profit organizations

One-quarter of the respondents ($N=16$) planned to collect from sources outside the United States (Q17). (See Figure 6.)

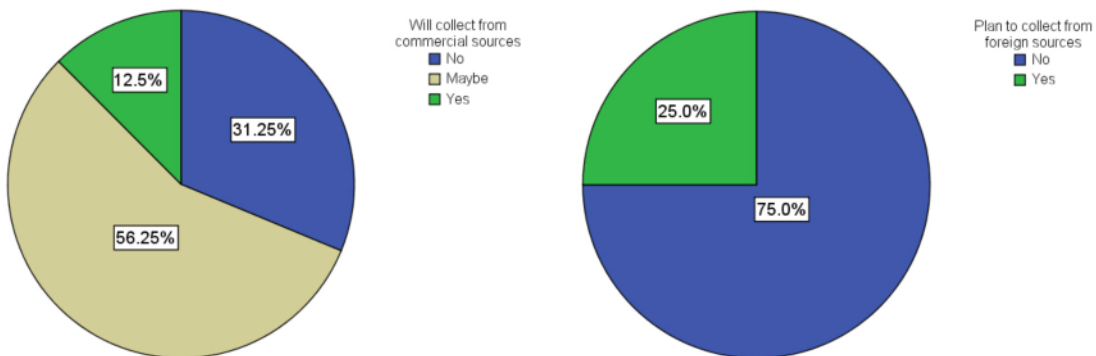


Figure 6 - Commercial and Foreign Material Sources (Q15 & Q17)

Respondents are also considering these web-based information sources for their collections (Q18).

- Local governments
- Inter-governmental organizations
- Non-government or quasi-governmental organizations that do advocacy work
- Non-government or quasi-governmental organizations that do policy work
- Academic institutions

3.3.3 Intellectual Property Considerations

Respondents were asked to describe the major intellectual property considerations they anticipated in regard to their planned web collections (Q19). All of the respondents (N=16) mentioned copyright or ownership as a major consideration. Table 4 lists the three categories of copyright considerations respondents anticipate. Half of the respondents are not certain if they will need copyright permission for the materials they plan to collect. In addition to copyright considerations, privacy, commercial reuse, and cultural sensitivity considerations were mentioned by a few respondents.

Intellectual Property Considerations
Copyright (N=16)
Definite need to gain permission to collect (n=4) <ul style="list-style-type: none"> • For government agency/office or IGO website • For publications from non-profit sources • For publications from news sources • For publications from educational sources
No need to gain permission to collect (n=4) <ul style="list-style-type: none"> • Archive will operate under fair use provisions • Permission granted in existing agreements
Questionable need to gain permission to collect (n=8) <ul style="list-style-type: none"> • Publications by private consulting firms commissioned by a government agency • Sites that repackage government material • Publications by international government organizations • Publications by state and local government
Privacy (n=2)
Commercial reuse (n=1)
Cultural sensitivity (n=1)

Table 4 - Intellectual Property Considerations (Q19)

3.3.4 Frequency of Change in Source Materials

Most (81%) of the respondents (N=16) estimated that their intended source materials would change either somewhat often, quite often, or at least daily (Q20). Likewise, the majority (81%) of respondents (N=16) plan to reacquire the materials at certain intervals (Q21). There is a somewhat significant relationship ($r_s = .552, p = .03$) between these two factors. Specifically, it is likely that if planned source materials regularly change, respondents will plan to reacquire the materials at certain intervals.

3.3.5 External Links

Web-published materials generally contain links to other web-published materials within a site or external to a site. A slight majority (56%) of the respondents ($N=16$) indicated it was important for content from the first level of external links¹¹ to be included in their collections (Q22).

A large majority (88%) of respondents ($N=16$) thought that users should be allowed to select broken links, that is, links that point to locations outside the archive but no longer work. Half of the respondents thought that a browser should provide a standard message for broken links and the other half thought that a custom message should be provided (Q23).

3.3.6 Authenticity

The ease with which digital materials, including web-published materials, can be copied and reformatted raises serious concerns about the authenticity of archived materials. We posed some questions to survey respondents in order to ascertain their concerns and thoughts about authenticity in a web archive environment.

Respondents generally were not concerned about altering web pages to add metadata (Q24). As illustrated in Figure 7, this finding was supported by the results of a question regarding archival practices that might endanger the authenticity of materials (Q25). Only three (20%) respondents ($N=15$) indicated that the addition of enhanced metadata to captured materials might endanger the authenticity of those materials.

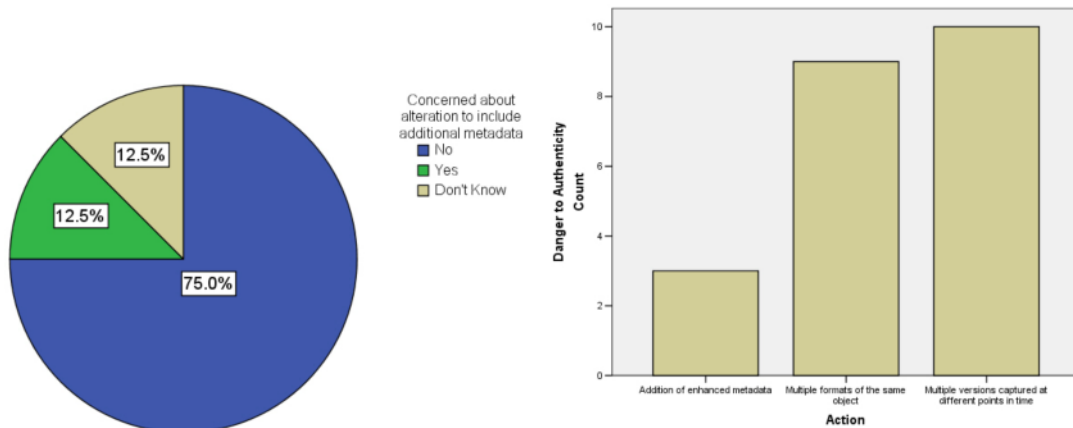


Figure 7 - Authenticity Concerns for Archived Materials (Q24 & Q25)

Respondents were more concerned about the threat to authenticity caused by multiple versions of materials captured at different points in time (67%, $N=15$) and the capture of multiple formats of the same object (60%, $N=15$). One respondent submitted this comment via email:

If a website has both a txt and pdf version of the same file, they may both be authentic. The presense [sic] of two versions does not raise risks of authenticity so much as reliability; if the content differs, someone might argue (in litigation) that they relied on one, not the other.

¹¹External links refer to links to other web sites that are outside of the publishing control of the web site owner.

If you're talking about keeping two versions of a document, with one being created by the archival repository, the situation may be slightly different. Authenticity may be a problem unless you can distinguish which was taken from the website and which was created by the repository. If you can demonstrate the process by which the copy was made and when the transformation was done, I don't think there would be much, if any, threat to authenticity.

When asked who is ultimately responsible for ensuring the authenticity of web-published materials in an archive, half of the respondents ($N=16$) believed that the content provider is responsible. However, a number of respondents (31%) felt that this responsibility lies with the curator or creator of the archive. None of the respondents felt that the end user was wholly responsible for the authenticity of web-published materials in an archive (Q26).

3.4 Curation of Web Collections

Section C of the survey was concerned with activities in the curation phase of web collection development. The questions sought to identify needs for organization, presentation, and ongoing maintenance of archived materials.

3.4.1 Searching Archived Materials

Figure 8 illustrates that respondents anticipate their end users will want the option of searching web archives using both full-text and subject categories. All respondents ($N=16$) agreed or strongly agreed with the statement "Our end users will want to use any word(s) to search the full-text of the web archive" (Q29). Respondents ($N=16$) showed only slightly less conviction about the statement "Our end users will want to search or browse web archive materials by subject categories or topics," with the majority indicating they either agreed (56%) or strongly agreed (38%)(Q30).

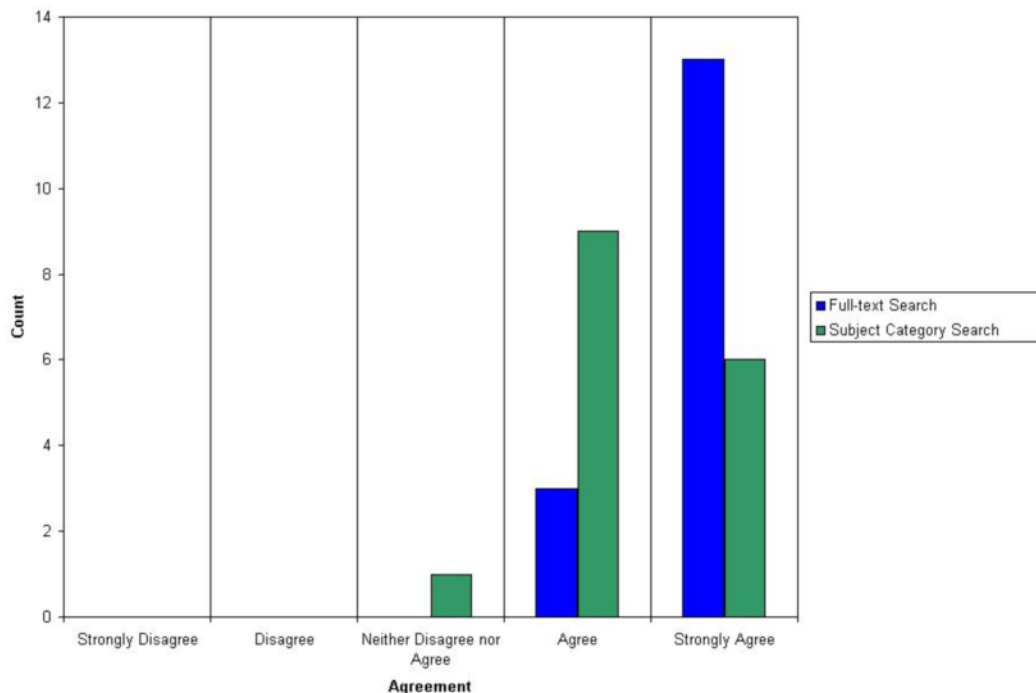


Figure 8 - Methods of Searching Web Archives (Q39 & Q40)

3.4.2 Presentation of Archived Materials

About half of the respondents ($N=16$) felt that it is important for their end users to interact with archived materials in a fashion that mirrors the source materials at the time of capture. Almost all of the remaining respondents (44%) neither agreed nor disagreed with this expectation (Q31).

To measure the relationship between curators' planned level of material selection (Q14) and end users' needs to interact with mirrored web sites in an archived collection (Q31), the responses for curators' planned level of selection were combined into two groups (i.e., 'website level' and 'organizational level' into one group and the more granular levels of selection into a second group) and the responses from question 31 were combined into three groups¹². A significant correlation was observed ($r_s = .517, p = .04$) between curators who thought their end users required a mirrored archive and those who were planning to collect at the website or organizational level. All of the respondents ($N=16$) agreed that their end users will require access to the materials in the web archives into the foreseeable future (Q32).

3.4.3 Deselection of Archived Materials

In order to determine what criteria curators might use for deselection of archived materials, respondents were instructed to select applicable deselection criteria from a predefined list (Q33). Most (93%) of the respondents ($N=15$) identified copyright violation as a criterion for deselection. Over half (73%) selected legal reasons, such as fraud. Nearly half of the respondents selected each of the remaining criteria for deselection: usage data thresholds (47%), storage costs (47%) and sensitive or offensive material (40%).

Respondents submitted the following additional deselection criteria (Q34):

- Value of material in relation to all available material
- Takedown requests from owners
- Data corruption
- Relevance to collection goals
- Availability elsewhere
- Duplication

One respondent pointed out that the criteria might be used in combination: "I would use a combination of the above criteria - if storage cost becomes too high, then I'd look at low use materials to determine deselection."

Figure 9 illustrates curators' expectations of end user acceptance of four deselection criteria: frequency of use (Q35), sensitive or offensive nature of materials (Q37), legal reasons (Q38), and financial reasons (Q39). All respondents ($N=16$) indicated their end users would accept removal of materials from an archive because of legal reasons such as fraud (Q38). In contrast, only one respondent ($N=16$) indicated their end users would accept removal of materials based on frequency of material use (Q35).

¹² Data were grouped as follows (new group = old group(s)): 'Disagree' = 'Strongly Disagree' & 'Disagree'; 'Neither Agree nor Disagree' = 'Neither Agree nor Disagree'; 'Agree' = 'Agree' & 'Strongly Agree'

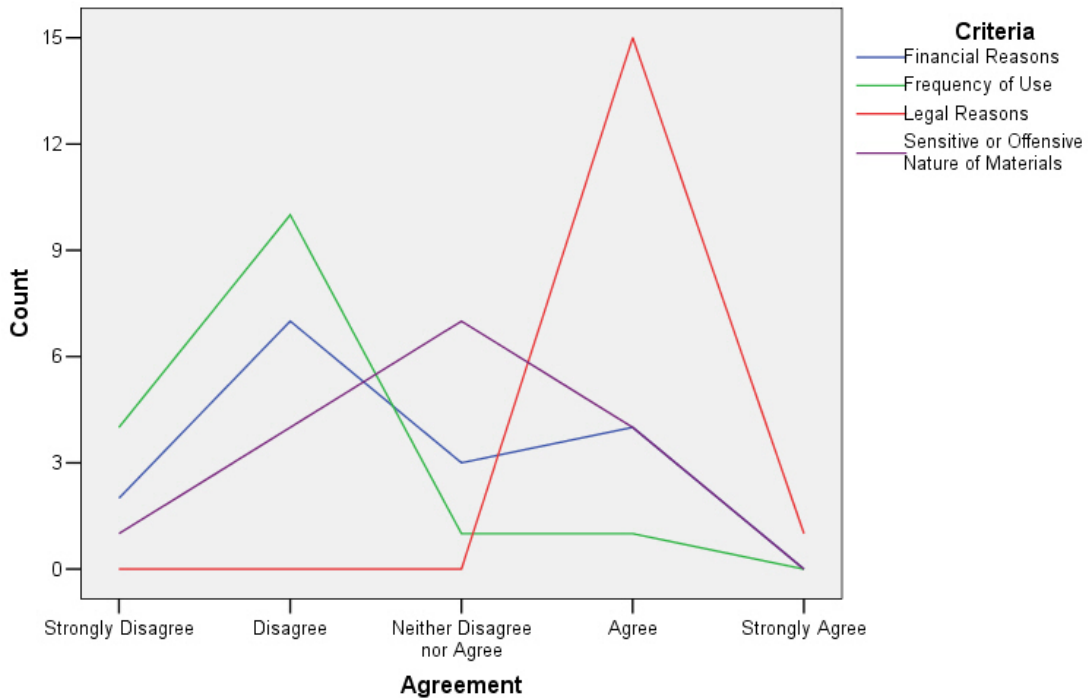


Figure 9 - End User Understanding of Deselection Criteria (Q35 & Q37-39)

When compared side-by-side, the criteria that respondents plan to use for deselection do not always correspond to the criteria that end users understand. (See Table 5.) This is particularly evident with copyright violations, which 93% of the respondents ($N=15$) plan to use as a reason for deselection (Q36). In contrast, only 6% ($N=16$) of respondents believe their end users generally understand how copyright protection applies to web-published materials.

Deselection Criteria	Curator Use	End User Understanding
Copyright violations	93%	6%
Legal reasons (such as fraud)	73%	100%
Usage	47%	6%
Financial reasons (such as storage costs)	47%	25%
Sensitive or offensive nature of material	40%	25%

Table 5 - Deselection Criteria and End User Understanding (Q33 & Q35-39)

3.5 Preservation of Web Collections

Section D of the survey addressed preservation needs and issues. The questions helped identify end user expectations and curator concerns that might impact web archive preservation activities.

3.5.1 Expectations

Questions 40 through 43 asked curators to indicate their agreement with statements regarding end user acceptance of preservation practices. (See Figure 10.) In general, respondents

indicated end users would expect a web archive to provide a unique persistent name for each object in the archive and to retain multiple versions of objects based on the degree of change to those objects.

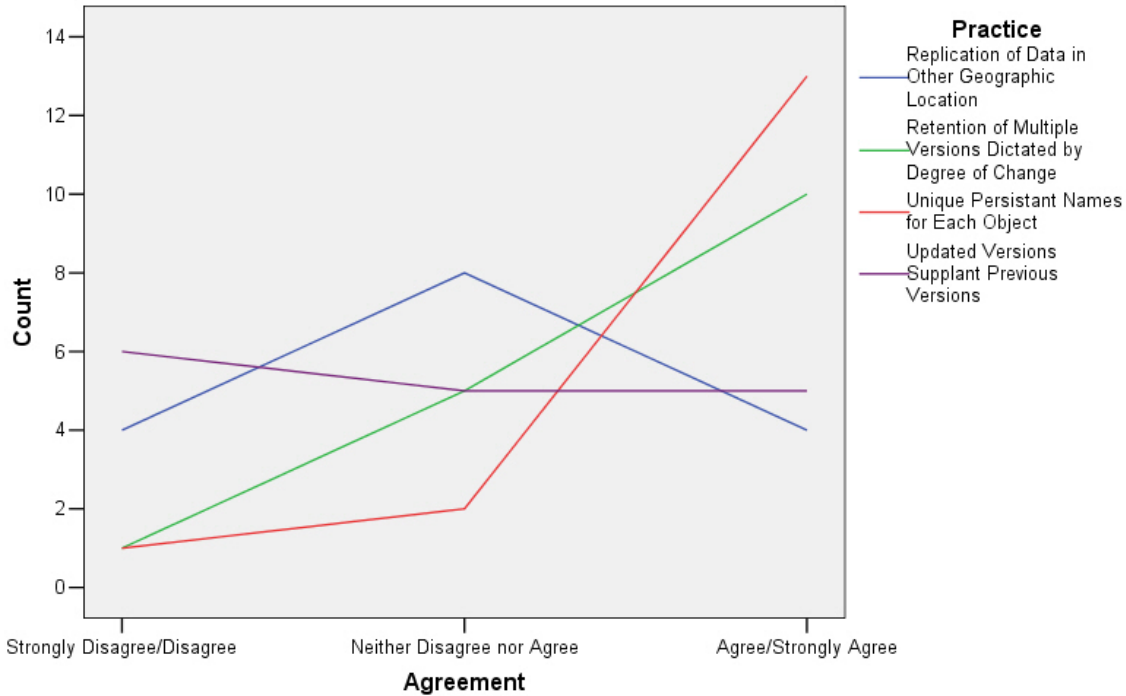


Figure 10 - User Expectations of Preservation Practices (Q40 - 43)

Respondents were fairly evenly divided regarding how accepting their end users would be if updated versions of web materials supplanted previous versions. Slightly more than one third (38%) of the respondents ($N=16$) indicated their end users would not be accepting of this practice and nearly one third (31%) indicated their end users would be accepting of this practice (Q40).

A majority (81%) of respondents ($N=16$) indicated their end users expected unique persistent names to identify each version, type, and format of materials in web archives (Q41). Likewise, a majority (63%) of respondents ($N=16$) indicated their end users generally find it acceptable that retention of multiple versions of web-published materials be dictated by the degree of change from version to version (Q42).

When respondents were asked if they agreed with the statement “It is important to end users that web archive content is replicated in another geographic location”, half of the respondents ($N=16$) neither agreed nor disagreed with this statement. One quarter (25%) agreed with the statement and one quarter (25%) disagreed (Q43).

3.5.2 Migration

Over time, preservation of archived web materials will likely require migration of those materials to new formats, versions, or platforms. Respondents were asked to evaluate the level of threat to the authenticity of materials for five different migration activities. The results are shown in Figure

11. The horizontal line at a count of seven represents half of the respondents who expressed an opinion¹³ (Q44).

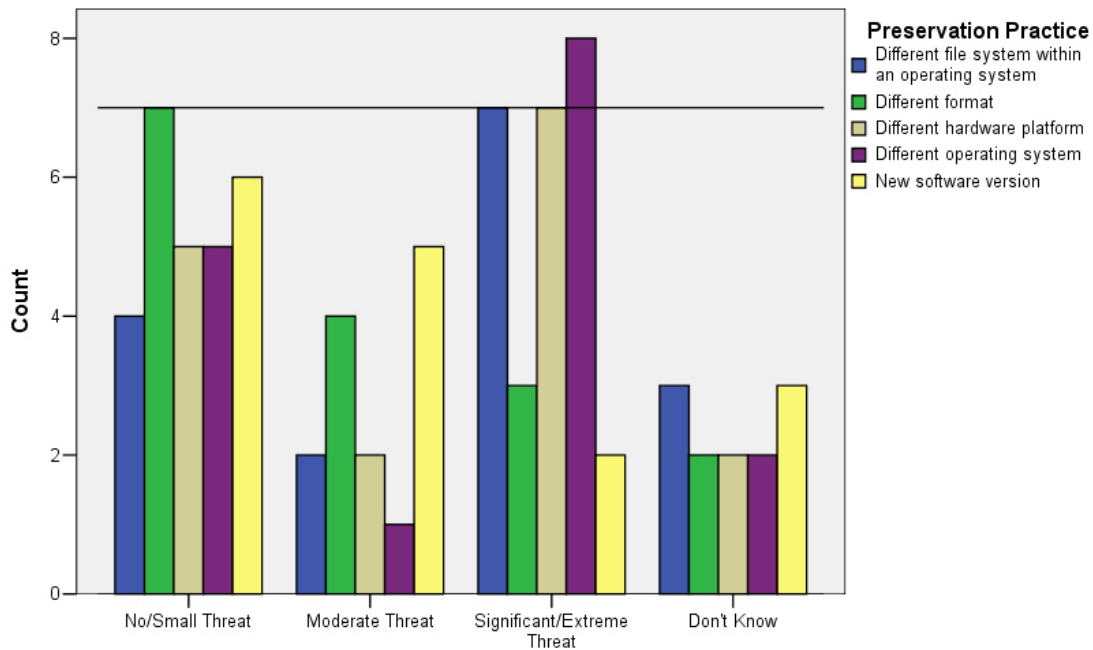


Figure 11 - Level of Threat to Authenticity by Migration Type (Q44)

Respondents who estimated the threat to the authenticity of archived materials ($n=14$) indicated the following:

- Migration of materials to a different operating system is a significant or extreme threat (57%).
- Migration to a different hardware platform or a different file system within an operating system is a significant or extreme threat (50%).
- Format migration is a small threat (50%).

3.6 Web Archiving Service Requirements

In addition to identifying the needs and concerns of curators and end users, a second goal of the survey was to gather requirements for the Web Archiving Service. Questions in Section E of the survey gathered input from the curators regarding their expectations of the service, in particular of the curator interface, web crawler, and crawl analysis tools.

Many of the questions in this section asked respondents to gauge the importance¹⁴ of parameters or attributes of crawls, of source materials, or of captured materials. For these questions, respondents generally told us that all of the parameters or attributes were important, that is, few respondents rated any parameter or attribute as 'Not Important'. Additionally, for all questions, mean values for every parameter or attribute were near or above a mean value of 3.00 (range=1-5).

¹³ For each of the five preservation practices, either 2 or 3 of the respondents indicated they did not know if the practice represented a threat to authenticity.

¹⁴ Possible Response Values: 1.00 = 'Not Important', 2.00 = 'A Little Important', 3.00 = 'Somewhat Important', 4.00 = 'Very Important', and 5.00 = 'Extremely Important'

In an effort to make the results more meaningful to those creating functional requirements for the project's tools, mean responses were calculated for each parameter and attribute. These means were used to rank-order the results for analysis. For most questions, all respondents (N=16) rated the importance of each parameter or attribute. Exceptions are noted on a per question basis.

3.6.1 Selection: Crawl Attributes

Respondents were asked the importance of a variety of crawl attributes to their selection decisions (Q45). Table 6 rank-orders the attributes by importance to curators. Appendix F - Q46 identifies additional crawl attributes suggested by respondents.

Crawl Attribute	<i>M</i>	<i>SD</i>	Overall <i>M</i>
			3.30
Content object format	3.75	1.07	
Number of broken links	3.56	0.73	
Content object type	3.56	1.03	
Content URLs	3.44	0.85	
Number of failures by error code and type	3.38	0.96	
Total crawl size	3.00	1.03	
Number of external links	2.88	0.89	
Total size by file type	2.81	0.75	

Table 6 - Importance of Crawl Attributes in Selection Decisions (Q45)

3.6.2 Acquisition: Crawl Parameters

Respondents were asked to rate how important it would be for them to have the ability to specify various crawl parameters when initiating a crawl (Q47). Table 7 reports the mean values for each parameter in rank order. Additional parameters suggested by respondents are in Appendix F - Q48.

It is worth noting that a majority (67%) of respondents (N=15) indicated that it is very important for them to have the ability to exclude materials from the capture process based on specific parameters (Q49).

Crawl Parameter	<i>M</i>	<i>SD</i>	Overall <i>M</i>
			3.91
Number of levels within targeted or entry-point URLs to capture	4.44	0.63	
Content object formats to capture	4.00	0.82	
Content object types to capture	3.94	0.85	
Frequency of crawl	3.88	1.20	
Time period over which to repeat the crawl	3.81	1.22	
Depth of external links to capture	3.69	0.87	
Compliance with robot exclusions	3.63	0.96	

Table 7 - Importance of Crawl Definition Parameters (Q47)

3.6.3 Acquisition: Real-time Data Monitoring

Respondents were asked about the importance of being able to monitor real-time data for an in-progress crawl (Q50).¹⁵ Table 8 reports the percentages of responses that were either ‘Very Important’ or ‘Extremely Important’.¹⁶ The data most often indicated as ‘Very’ or ‘Extremely’ important were ‘Crawl completion status by EPU’ (60%, *n*=15) and ‘Content object formats captured’ (57%, *n*=14). Other real-time data attributes considered important by respondents are listed in Appendix F - Q51.

Datum	Very/Extremely Important
Completion status by EPU	60%
Content object formats captured	57%
Errors by error code	47%
Content object types captured	40%
Total size captured	33%
Total size captured by object type and format	33%

Table 8 - Importance of Realtime Data Reporting During Crawls (Q50)

3.6.4 Curation: Collection-Level Attributes

Respondents were asked to identify the importance of knowing the values of collection-level attributes in their ongoing collection management process (Q52). Table 9 reports the mean

¹⁵ This analysis does not include the one respondent who indicated ‘Don’t Know’ to all data attributes in question 50.

¹⁶ Although ‘Content object formats captured’ and ‘Content object types captured’ were indicated as ‘Very’ or ‘Extremely’ important by a substantial number of respondents (57%, *n*=14 and 40%, *n*=15 respectively), approximately one third of the respondents (36%, *n*=14 and 33%, *n*=15 respectively) found these data to be of little importance. This caused their mean scores to be skewed and misleading. As an alternative to mean scores, percentages of responses that were either ‘Very’ or ‘Extremely’ important were calculated.

values for each attribute in rank order. Additional attributes suggested by respondents are in Appendix F - Q53.

Collection-Level Attribute	<i>M</i>	<i>SD</i>	Overall <i>M</i>
			3.90
EPU for crawl	4.56	0.51	
Crawl complete date	4.44	0.89	
Measurement of content change over time ^a	4.27	0.88	
Crawl parameters	4.25	0.58	
Content URLs for crawl	3.94	0.85	
Errors encountered by error code	3.56	0.63	
Crawl curator	3.56	1.09	
Collection size by type and format	3.44	0.63	
Total crawl size	3.06	0.68	

^a N=15 for this attribute

Table 9 - Importance of Collection-Level Attributes for Collection Management (Q52)

3.6.5 Curation: Object-Level Attributes

Respondents were asked to identify the importance of knowing the values of certain object-level attributes in their ongoing collection management process (Q54). Table 10 reports the mean values for each attribute in rank order. Additional attributes suggested by respondents are in Appendix F - Q55.

Object-Level Attribute	<i>M</i>	<i>SD</i>	Overall <i>M</i>
			4.16
URL	4.81	0.40	
Title	4.63	0.50	
Archive Date	4.56	0.63	
Author	4.56	0.73	
Creation Date	4.50	0.52	
Format	4.25	1.07	
Subject	4.19	0.98	
Name	4.06	1.00	
Description	4.06	1.06	
Frequency of Change	4.00	0.73	
Type	3.94	1.06	
Language	3.31	0.87	
Size	3.25	0.93	

Table 10 - Importance of Object-Level Attributes for Collection Management (Q54)

3.6.6 Curation: Description

Respondents were asked to indicate the level(s) of descriptive metadata that is critical for materials in their planned collections (Q56). As shown in Table 11, it is important to curators (*N*=15) to be able to apply metadata to archive materials at several levels, with the website level being the most critical level.

Level	% Respondents
Website level	87%
Web page level	67%
Logical document level	67%
Object level	67%

Table 11 - Desired Level of Descriptive Metadata (Q56)

3.6.7 Curation: Organization

Respondents were asked to identify the importance of certain attributes of archived materials for use as end user access points or search criteria (Q57). Table 12 reports the mean values for each attribute in rank order. The three most important attributes are author, title, and URL. Additional attributes suggested by respondents are in Appendix F - Q58.

Captured Material Attribute	<i>M</i>	<i>SD</i>	Overall <i>M</i>
			3.90
Author	4.63	1.03	
Title	4.56	1.03	
URL	4.44	1.21	
Object Format/Type	3.94	0.85	
Date/Time of Capture ^a	3.73	1.49	
File Name	3.56	1.26	
Language	3.25	1.13	
File Size	3.13	1.26	

^a N=15 for this attribute

Table 12 - Importance of Attributes as End User Access Points (Q57)

4 Discussion

4.1 Building Web Collections

With regard to the Web-at-Risk project, the most significant challenges anticipated by the respondents are as follows:

Financial Challenges

- Cataloging
- Preservation
- IT Support
- Staff Training

Technical Challenges

- Metadata Creation
- Dynamic Source Materials
- Encrypted Source Materials
- Password Protected Source Materials

Some of these challenges, such as cataloging, preservation, IT support, and metadata creation, may be mitigated by the project's Web Archiving Service. Others, especially encrypted source materials and password protected source materials, will likely remain challenges that this project will not overcome.

Curators anticipate that end users will not be very tolerant of technical roadblocks, suggesting a conservative web collection development practice of first assessing the nature of desired source materials and then offering end users a realistic judgment of the likelihood that the materials can be included in an archived web collection.

4.1.1 Collection Policies & Material Selection

Although material selectors in libraries generally include web sites in subject lists, creating collections of web-published materials is a relatively new practice. Web materials in collections are frequently drawn from constantly changing sources. This highlights important considerations for collection-building practices, including:

- Assessing the change rate of source materials
- Establishing the interval at which collection materials will be captured
- Articulating criteria for retention of earlier versions

In spite of the fact that material and format types are not often explicitly specified in collection policies and practices, existing policies and practices do appear to either directly or indirectly support the collection of a wide range of digital materials and formats. These two attributes of web-published materials are certainly important to the project's curators, who ranked material type and format type high on both the list of configurable crawl parameters and the list of attributes that impact selection decisions and collection management decisions.

In developing a specific collection plan, curators should review the material formats and types in the source materials in advance. This could be done through crawling the targeted web sites and reviewing a report of key attributes of the materials, like format and type. Organizational policies should reference the types of materials and formats the organization supports technically. Material selection decisions for a web collection need to be reviewed in relation to the supported types. Additionally, curators need to understand the implications of archiving materials that the organization may not be able to present to end users.

4.1.2 Intellectual Property

Every single respondent mentioned copyright as an intellectual property consideration. Copyright violation was also mentioned by a large number of respondents as a deselection criterion for materials in an archive.

There are a variety of sources from which respondents plan to collect materials. Most of the respondents plan to collect from domestic non-commercial sources; however, one quarter plan to collect from foreign sources and over two-thirds concede that there are situations in which they might collect from commercial sources, particularly when those sources support the goals of the collection. This range of sources poses copyright challenges for which curators would like clear approaches.

Although a few respondents thought they would not need to obtain permission from content providers because of the fair use provisions of copyright law, fully half expressed confusion about whether permissions might need to be obtained before creating web collections for an archive and if so, how permissions should be acquired. The curators' confusion has a corollary in end users — namely that curators perceive that end users lack an understanding of how copyright law applies to web-published materials. It may be that end users will be intolerant of curatorial decisions based on copyright law compliance.

The Web-at-Risk project's Rights Management Protocol should help curators manage copyright issues, but they may need additional guidance when making decisions about whether or not advance permission of any kind is required before collecting materials in non-commercial, commercial, and international settings. Additionally, both curators and end users would benefit from educational materials targeted at how copyright law applies to web-published materials. Curators may find it helpful to be able to customize this training to their specific collections.

4.1.3 Material Organization and Presentation

It is interesting that 50% of the survey respondents intend to build collections of web-published materials at other than organizational or website levels. Specifically, they intend to collect at the logical document, webpage, or object level. This suggests the following implications for the organization and presentation of materials in these archived collections:

- Metadata creation at the website level will not suffice.
- The interface for 'original cataloging' will need the flexibility to address metadata specific to the level of selection.
- Presentation of materials might require a customized interface

4.1.4 End User Expectations

Respondents told us the following about what end users will expect from an archive of web collections:

- End users will want to search the archive using both full-text and subject categories
- End users will not necessarily need the archive to mirror the source materials except in the case where the level of collection is at the website or organizational level.
- The biggest threat to the authenticity of archived materials is the retention of multiple formats or versions of a document.
- Guarantee of the authenticity of the archived materials is primarily the responsibility of the creator of the source materials and, secondarily, the responsibility of the archive creator or owner.
- Links that point outside the archive should be selectable by the user and should present browser or archive generated messages as appropriate when the link is broken.
- Unique, persistent names are required to identify each archived object.
- End users will continue to want access to archived materials into the foreseeable future and may be intolerant of material deselection.

4.2 Web Archiving Service Requirements

The Web Archiving Service being developed by the Web-at-Risk project will include a Curator User Interface (CUI) to various tools that will allow curators to select, curate, and preserve their web collections. With regard to both the CUI and the planned tools, curators require features and flexibility in several areas. These include:

- Level of selection
- Frequency of reacquisition
- Specification of crawl configuration parameters
- Application of metadata
- Migration
- Validation

4.2.1 Level of Selection

Although many respondents plan to select their source materials at the website level, half plan to build collections at a more granular level, such as the web page or object level. Over half of the respondents also want to include materials from the first level of external links in their targeted source materials.

4.2.2 Frequency of Reacquisition

Planned source materials change frequently, so curators also need flexibility in scheduling material reacquisition. At times, individual web pages or documents are relevant to a collection, whereas the website to which they belong as a whole is not. As a result, curators would like the flexibility of scheduling, on an ad-hoc basis, crawls that capture only a few documents at a time; perhaps even for one time only.

For some materials, currency is essential. A crawler that can recognize when new materials appear at a source would be a valuable asset for curators who manage collections with these types of materials.

4.2.3 Specification of Crawl Configuration Parameters

The tools should offer advanced configuration capabilities since curators want as much control as possible over how their crawls are configured. They also want as much information as possible about the captured materials. Detailed real-time data monitoring is not critical for tools; however, the ability to view the completion status of a crawl and the object formats captured on a real-time basis is important.

4.2.4 Application of Metadata

It is clear that respondents are concerned about the challenges of applying metadata to captured materials. Generally, respondents were not concerned that embedding metadata would threaten the integrity of materials, so the automatic generation and application of as much metadata as possible by the Web Archiving Service's tools would likely be of significant benefit. This practice would be beneficial at all levels of selection (e.g. object level, website level). However, since a few respondents were concerned about this practice, the ability to disable it would also be a desirable feature.

4.2.5 Migration

Digital preservation may require that materials be migrated¹⁷ over time to maintain their accessibility. Respondents viewed the migration of materials to different formats or software

¹⁷ A method of preserving digital materials and access to those materials by copying or reformatting the materials while preserving their intellectual content.

versions to be relatively safe, but were concerned about the migration of materials to a new operating system or a new hardware platform.

4.2.6 Validation

A significant danger when dealing with the capture of digital materials is that of data corruption. Corrupt data is of no value to the end user. For this reason, the crawler tool should have the capability of validating the content subsequent to its capture.

4.3 Closing

Preservation of digital materials imposes technical requirements that make it important to address preservation issues as part of the material creation process. The more time that passes between the creation and preservation of digital objects, the more likely it is that data and information contained within the objects will become inaccessible. As a result, the usual temporal separation of collection building and archive creation fails to meet the needs of digital archives.

In order to effectively address this issue, the emerging work of web archiving must adopt practices from both the library community's collection development tradition and the archive community's preservation tradition. This introduces seeming paradoxes, such as the notion of 'deselection' from an 'archive'. These perceived paradoxes must be identified and explained in order to prevent misconceptions and failed expectations in web preservation efforts.

Many of the Web-at-Risk project's curators are already creating digital collections and some even have experience with creating digital archives; however, curators are encountering difficulties with funding, hiring the appropriate staff, technical issues, and metadata creation. Curators need assistance in each of these areas. A Web Archiving Service may help in the following ways:

- Optimizing the expertise of librarians to create web collections while decreasing the technical expertise required
- Reducing the number of technical support staff needed at the local library level
- Reducing the hardware and software investment required at the local library level
- Providing the hardware to collect, manage, and preserve web collections in archives
- Providing the technical software tools to select and manage materials in web collections
- Enabling metadata application at multiple levels (i.e., website, logical document, webpage, and object levels)
- Specifying a metadata standard at the website level
- Automatically generating basic metadata about websites and the objects comprising them

Copyright issues and metadata creation stand out as major challenges in the creation of web collections. Robust, full-featured tools are necessary to provide the flexibility curators need as well as to alleviate some of the financial and technical challenges they face today. In addition to support from the Web Archiving Service tools, educational materials could provide guidance to both curators and their end users regarding how copyright law applies to web collection and archive development.

In later phases of the Web-at-Risk project, curators will create case studies describing their web collection development processes and will also evaluate the Web Archiving Service's interface and tools. The degree to which the project's Web Archiving Service addresses curators' anticipated financial and technical challenges can then be assessed.

Appendix A. Collection Development Framework for Web Archives

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

Appendix B. Survey Participants

Public Policy and Political Movements	
Gabriella Gray	Curator Online Campaign Literature Archive Young Research Library UCLA
Ronald J. Heckart (collaborating with Nick Robinson)	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 (collaborating with Michael Nash)	Tamiment Archivist Tamiment Library New York University
Michael Nash (collaborating with Peter Filardo)	Head Tamiment Library & Robert F. Wagner Labor Archives New York University
Nick Robinson (collaborating with Ronald J. Heckart)	Librarian Institute of Governmental Studies Library Institute of Governmental Studies UC Berkeley

Local, State, Federal, and International Government Information	
Sherry DeDecker (collaborating with Janet Martorana)	Head Government Information Center Davidson Library UC Santa Barbara
Charles Eckman	Head Social Sciences Resource Center Green Library Stanford University
Valerie Glenn (collaborating with Arlene Weible)	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

Local, State, Federal, and International Government Information	
Amy Kautzman (handed over to Jim Church)	Head, Research Reference and Collections Doe/Moffitt Libraries UC Berkeley
Jim Church (in lieu of Amy Kautzman)	International Documents Librarian Doe/Moffitt Libraries UC Berkeley
Linda Kennedy (collaborating with Juri Stratford)	Head Government Information and Maps Department Shields Library UC Davis
Ann Latta	State and Local Documents Bibliographer Social Sciences Resource Center Green Library Stanford University
Janet Martorana (collaborating with Sherry DeDecker)	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 (collaborating with Linda Kennedy)	Government Information Librarian Shields Library UC Davis
Yvonne Wilson	California and Orange County Government Information Librarian Langson Library UC Irvine
Arlene Weible (collaborating with Valerie Glenn)	Head of the Government Documents Department University of North Texas Libraries

Appendix C. Survey Instrument

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	
_____	_____

The Web at Risk: Needs Assessment Survey Report

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)				

The Web at Risk: Needs Assessment Survey Report

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	
	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

The Web at Risk: Needs Assessment Survey Report

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.

Appendix D. 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 E. Letter of Consent

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 F. Survey Results

Note: Although 16 Curators participated in the survey, some chose not to answer all of the questions. Unless otherwise noted, where a 'Total' of less than 16 is indicated, this is the total number of curators who responded to the question. If the 'Total' has another meaning, it is indicated in the response data.

Section A. About Your Collections

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

Response	Total = 16
Websites and printed ephemeral election materials produced for campaigns for local, state, and federal offices and ballot measures affecting the [city ¹] area.	
Labor and Industrial Relations, Organizational Behavior, Labor History, and trade union issues, with a strong focus on the post World War II era	
U.S. labor and radical politics, with a [city ¹] concentration.	
American public affairs and public policy, with an emphasis on [the state ¹].	
U.S. and [state ¹] publications issued by governmental agencies, selected [... ¹] county and its cities	
Collection Focus: General publications, journals, government documents, numeric datasets and archival resources supporting the research, teaching and learning needs of the [institution ¹] community. Specific Focus as [position ¹]: Publications, documents and archives of international governmental agencies.	
U.S. and [state ¹] government information	
[State ¹] and other key US state and local (city/county/regional) government information ; non-governmental organizations (which includes policy institutes, think tanks, research institutes, community-based organizations, nonprofits, etc.) I am also responsible for the Canadian Depository collection - but that is out of scope for this project.	
The International Documents Collection at [institution ¹] is strongest in the areas of international economic and social development, human rights, peace and conflict studies, public health, and international law.	
Material (in all formats) to support the research and teaching interests of the campus. A broad-based research collection, strong in the sciences, particularly environmental and agricultural sciences and social sciences, that also serves public users of the region.	
General publications, journals, government documents, numeric datasets and archival resources supporting the research, teaching, and learning needs of the [institution ¹] community. Specific focus: state and local government documents including regional agencies with emphasis on [state ¹] local agencies.	
Government documents at the Federal, state and regional, county and city level	
Permanently valuable records of the State of [state ¹], including state agency publications and records.	

Response	Total = 16
United States federal government publications; [...] state government publications; United Nations, OECD, and other international governmental organizations' publications	
[...] State Documents, County of [county ¹] and the cities of [county ¹], [state ¹]	
[Institution ¹] library is a 71% federal depository, full depository for [...] state documents, and [...] city and county documents. The documents collection includes a large and diverse social sciences data collection.	

1 - Information removed to maintain anonymity.

2. Who are the end users of your collections?

Response	Total = 16
[Institution ¹] faculty, students and staff. Scholars and students across the world.	
[Institution ¹] faculty, staff and graduate students; citizen researchers interested in labor history; arbitrators, lawyeres and labor activists; researchers who visit from other universities with specific goals	
Students (grad & undergrad) and faculty in history and related disciplines; independent researchers, progressive and labor activists and labor union staff, documentarians and writers. About half are from outside the [institution ¹] community.	
Institute of Governmental Studies scholars, [institution ¹] library users, general public	
[Institution ¹] faculty, students, staff; [city ¹] community	
Users: [institution ¹] faculty, students, staff, affiliated researchers and members of the general public.	
undergraduate and graduate students, faculty, researchers, librarians, government officials, the general public	
[Institution ¹] Faculty, students, staff; other university/college students; the general community; local agencies and organizations.	
Students and faculty from a wide range of programs and departments, including but not limited to political science, economics, demography, public health, geography, public policy, and multiple area studies programs, e.g. Africa and South/Southeast Asia.	
Students. faculty and staff; local community users, regional users, remote users.	
[Institution ¹] faculty, students, staff, affiliated researchers and members of the general public.	
Academic researchers; upper division undergraduates and graduate students, faculty, and general public/community users from local area	
The Legislature, state agencies, and the general public.	
Faculty, graduate students, undergraduate students, and general public (ranging from high school students to scholars)	
Faculty, staff and students of [institution ¹]; community users, consulting firms, non-profit organizations, governmental agencies.	

Response	Total = 16
[Institution ¹] affiliates (faculty, students, researchers, staff), affiliates from other local colleges and universities, [... ¹] community members	

1 - Information removed to maintain anonymity.

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

(Note: Answers to Question 3 are in Appendix G)

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

Material Type	Total	0%	< 25%	25-50%	51-75%	> 75%
Journals & Periodicals	9					
#		3	1	0	2	3
%		33.3	11.1	0.0	22.2	33.3
Books & Brochures	9					
#		2	4	0	1	2
%		22.2	44.4	0.0	11.1	22.2
Databases	8					
#		3	2	0	0	3
%		37.5	25.0	0.0	0.0	37.5
Newspapers	8					
#		4	1	1	0	2
%		50.0	12.5	12.5	0.0	25.0
Videos	8					
#		2	6	0	0	0
%		25.0	75.0	0.0	0.0	0.0
Audio files	9					
#		3	4	1	1	0
%		33.3	44.4	11.1	11.1	0.0
Image files	9					
#		1	2	3	1	2
%		11.1	22.2	33.3	11.1	22.2
Technical & Research Reports	8					
#		2	1	1	2	2
%		25.0	12.5	12.5	25.0	25.0

The Web at Risk: Needs Assessment Survey Report

Material Type	Total	0%	< 25%	25-50%	51-75%	> 75%
Proceedings of Meetings & Symposia	8					
#		3	1	0	1	3
%		37.5	12.5	0.0	12.5	37.5
Doctoral Dissertations & Master's Theses	8					
#		5	1	0	2	0
%		62.5	12.5	0.0	25.0	0.0
Government Records	9					
#		1	2	1	1	4
%		11.1	22.2	11.1	11.1	44.4
Unpublished Works & Publications of Limited Circulation	9					
#		2	4	1	1	1
%		22.2	44.4	11.1	11.1	11.1
Other:	3					
Websites	1					
#		0	0	0	0	1
Static html	1					
#		0	0	0	0	1
Not Specified	1					
#		0	0	1	0	0

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.)

Response	Total	#	%
	14		
a. Yes		5	35.7
b. No		9	64.3

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

Respondents	Total	#	Total	%
			5	31.3
Responses	Total	#		
	7			
Web / HTML interface to mirrored websites		1		20.0
Content Management System (CMS)		0		0.0
Institutional Repository Software (e.g., DSpace, Eprints, Fedora)		3		60.0
The UCLA Digital Library Program has developed an in-house system for the management and delivery of digital content.			1	
eScholarship Repository, CDL			1	
ContentDM			1	
Other		3		60.0
Peer-to-Peer System: LOCKSS			2	
ISL/UIUC SafetyNet Software, ad hoc storage on file servers, simple database interfaces to documents on webservers.			1	

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

Response	Total = 6
Technical limitations of capture software.	
Metadata: Choice of standard (Dublin Core) and DC elements to include. Ongoing creation/cataloging of objects.	
Attracting faculty and sustaining faculty interest	
Obtaining FTE support for ongoing digital file conversion (contracts, PDF generation, scanning, etc)	
Internal resources for capture and description	
Internal resources for access and display	
Lack of technical support staff, server space, and software.	
Long waiting period for digital projects, taken up by other libraries and special collections.	
software development	
Cost	

Response	Total = 6
Not enough staff with computer/IT skills necessary to work with digital documents.	
Major shifts in workflow and responsibilities that cause disruptions in established policies and procedures.	

Section B. Selection: Policy, Identification, & Acquisition

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

Material Type	Total	Include (√)	Exclude (√)	Not Specified (√)
Journals & Periodicals	15			
#		10	0	5
%		66.7	0.0	33.3
Books & Brochures	15			
#		10	0	5
%		66.7	0.0	33.3
Databases	15			
#		8	0	7
%		53.3	0.0	46.7
Newspapers	15			
#		7	0	8
%		46.7	0.0	53.3
Videos	15			
#		5	1	9
%		33.3	6.7	60.0
Audio files	15			
#		4	1	10
%		26.7	6.7	66.7
Image files	15			
#		7	0	8
%		46.7	0.0	53.3
Technical & Research Reports	15			
#		9	1	5
%		60.0	6.7	33.3

The Web at Risk: Needs Assessment Survey Report

Material Type	Total	Include (√)	Exclude (√)	Not Specified (√)
Proceedings of Meetings & Symposia	15			
#		6	0	9
%		40.0	0.0	60.0
Doctoral Dissertations & Master's Theses	14			
#		4	1	9
%		28.6	7.1	64.3
Government Records	15			
#		10	0	5
%		66.7	0.0	33.3
Unpublished Works & Publications of Limited Circulation	14			
#		5	0	9
%		35.7	0.0	64.3
Other:	3			
web pages	1			
#		1	0	0
websites	1			
#		1	0	0
copyrighted material	1			
#		0	1	0

Additional Comments:

Response	Total = 5
<p>NOTE: again, [our library¹] does not have an overall digital collection policy. We do plan to collect, under this particular project, political websites relating to [state¹] labor and radical politics. NOTE: again, section 9 is N/A, for the same reasons.</p>	
<p>Our collection policy for government information generally states that we collect "materials in all formats". We don't have a specific collection policy for digital collections. So, I indicated this as "include" above; if there was a particular material type that we don't specifically cover in the government information policy, i left it not specified. I also answered not applicable for #9 because we don't have a specific policy.</p>	

Response	Total = 5
<p>I am not sure what you mean by support. If this means providing access, the libraries do that for selected web resources via our web pages and the library catalog. At this point this is primarily for serial titles and databases. I have not even attempted to submit requests to catalog digital monographs.</p>	
<p>Little support is provided to check existing links in the catalog, and certainly no effort is underway to digitally preserve international government information.</p>	
<p>We do not currently have any digital projects; Our collection is also limited, except for reference materials, to works published by government agencies. Our answer to this question reflects general collection policy for the Library,</p>	
<p>Our collections can include any and all formats and genres listed above. Items are selected primarily on the basis of their provenance, not format or genre.</p>	

1 - Information removed to maintain anonymity.

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	Total	Acceptable (√)	Acceptable within Limits (√)	Not Acceptable (√)	Not Applicable (√)
Adobe Portable Document Format (pdf)	15				
#		12	1	0	2
%		80.0	6.7	0.0	13.3
Adobe PostScript (ps)	14				
#		5	2	2	5
%		35.7	14.3	14.3	35.7
Lotus 1-2-3 (wk1, wk2, wk3, wk4, wk5, wki, wks, wku)	15				
#		8	2	2	3
%		53.3	13.3	13.3	20.0
Lotus WordPro (lwp)	15				
#		5	2	4	4
%		33.3	13.3	26.7	26.7
MacWrite (mw)	15				
#		5	2	5	3
%		33.3	13.3	33.3	20.0
Microsoft Excel (xls)	15				
#		10	2	0	3
%		66.7	13.3	0.0	20.0

The Web at Risk: Needs Assessment Survey Report

Digital Format	Total	Acceptable (√)	Acceptable within Limits (√)	Not Acceptable (√)	Not Applicable (√)
Microsoft PowerPoint (ppt)	15				
#		8	2	1	4
%		53.3	13.3	6.7	26.7
Microsoft Word (doc)	15				
#		10	2	0	3
%		66.7	13.3	0.0	20.0
Microsoft Works (wks, wps, wdb)	14				
#		4	3	2	5
%		28.6	21.4	14.3	35.7
Microsoft Write (wri)	14				
#		4	4	1	5
%		28.6	28.6	7.1	35.7
Rich Text Format (rtf)	15				
#		11	1	0	3
%		73.3	6.7	0.0	20.0
Shockwave Flash (swf)	15				
#		7	1	3	4
%		46.7	6.7	20.0	26.7
Audio (mp3, wav, midi, ra)	14				
#		9	2	0	3
%		64.3	14.3	0.0	21.4
Images (jpeg, jpg, gif, png, tif)	15				
#		11	1	0	3
%		73.3	6.7	0.0	20.0
Text (ans, txt)	15				
#		11	0	0	4
%		73.3	0.0	0.0	26.7
Video (mpeg, ra, mov, rm)	15				
#		8	4	0	3
%		53.3	26.7	0.0	20.0
Web Pages (htm, html, asp, jsp, php)	15				
#		11	1	0	3
%		73.3	6.7	0.0	20.0

The Web at Risk: Needs Assessment Survey Report

Digital Format	Total	Acceptable (√)	Acceptable within Limits (√)	Not Acceptable (√)	Not Applicable (√)
Supporting Code (css, js)	14				
#		7	3	1	3
%		50.0	21.4	7.1	21.4
Other:	2				
MS Access	1				
#		1	0	0	0
databases	1				
#		1	0	0	0

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

Response	Total	#	%
	16		
a. Yes		9	56.3
b. No		7	43.8

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

Total	None at All	Very Little	Some	A Fair Amount	A Large Amount
16					
#	1	2	5	5	3
%	6.3	12.5	31.3	31.3	18.8

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.

Total	Not Accepting	A Little Accepting	Somewhat Accepting	Very Accepting	Extremely Accepting	Don't Know
16						
#	3	1	5	3	0	4
%	18.8	6.3	31.3	18.8	0.0	25.0

Note: The following was received via email from one participant:

“In our opinion, the public will expect to find the information they are looking for without concern for privacy issues (although they might take exception if it is their privacy being violated.)”

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.

Total	Not Accepting	A Little Accepting	Somewhat Accepting	Very Accepting	Extremely Accepting	Don't Know
16						
#	5	4	5	0	0	2
%	31.3	25.0	31.3	0.0	0.0	12.5

Note: The following was received via email from one participant:

“Similar concern as above [additional emailed response to question 12]. The public will expect us to find a way to capture the materials. They won't accept excuses of, "It was password protected." They'd expect us to find a way around the technical barriers. At the same time, the public often doesn't know what's not been captured, so the issue may never arise.”

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

Response	Total	#	%
	16		
Object level (Example: images or movies)		2	12.5
Web page level (Example: .html, .xml, etc.)		3	18.8
Logical document level (Example: article spanning multiple .html files)		3	18.8
Website level (Example: all content within a URL)		7	43.8
Organizational level (Example: websites within an agency's top-level URL)		1	6.3

Note: The following was received via email from one participant:

“Our principal concern is to capture documents, but with metadata to put them in context of creation (provenance, related documents roughly equivalent to a series). We are currently using software that captures websites en masse, but we find access to the documents through archived websites problematic and cumbersome.”

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

Response	Total	#	%
	16		
a. Yes		2	12.5
b. No		14	87.5

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

Information Source (Total = 2)	URL
publications from the Institute for Local Government	http://www.westerncity.com/index.jsp?zone=ilsg
Public Policy Institute of CA	http://ppic.org/main/home.asp
International Institute for Environment and Development	http://www.iied.org/
Population Reference Bureau	http://www.prb.org/
San Diego Assn of Governments	http://www.sandag.org/
Human Rights Watch	http://www.hrw.org/

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

Response	Total = 11 ¹
When a website contains the content of articles/items originally published by commercial sources.	
There is a potential for union newspapers and the news media to play a role in the collections we build	
Relevant content, e.g., mass media in our overall collecting scope	
Educational (.edu) and nonprofit (.org) organizations web-publish documents that they also will supply in print for a fee.	
Material produced under the auspices of an international organization but co-published by a commercial entity.	
If they were in danger of disappearing and deemed necessary to complement a site we had harvested.	
If an agency has outsourced the publication of its material.	
If the commercial website or organization re-publishes governmental information, or takes over publishing information previously supplied by a .gov source	
We would not actively collect from commercial sources, although some commercial materials may find their way into the collections For example, legislative study committee reports might include clippings from a commercial newspaper.	

Response	Total = 11¹
We might want to collect some public policy reports from non-profit organizations or think tanks. This may not be possible in many cases due to copyright restrictions.	
I am defining "commercial" as non-governmental. This could include truly commercial enterprises, 501(c)3 non-profits as well as other IGOs/NGOs. These organizations produce documents of interest, both separate from the authority of a specific govt agency and in collaboration with or via contract from govt agencies.	

1 - Total includes the two respondents from question 15 above who are definitely planning to collect from commercial sources.

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

Response	Total	#	%
	16		
a. Yes		4	25.0
b. No		12	75.0

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

Information Source (Total = 3)	Is Commercial?	URL
World Intellectual Property Organization	N	www.wipo.org
UNESCO	N	www.unesco.org
World Trade Organization	N	www.wto.org
UN Conference on Trade and Development	N	www.unctad.org
International Monetary Fund	N	www.imf.org
International Institute for Environment and Development	Y	http://www.iied.org/
International Crisis Group	Y	http://www.crisisgroup.org
International Labour Organization	Y	http://www.ilo.org
Joint United Nations Programme on HIV/AIDS	N	http://www.unaids.org/en/resources/publications.asp
United Nations Children's Fund (UNICEF) Innocenti Research Centre	N	http://www.unicef-icdc.org/publications/
United Nations Research Institute for Social Development	N	http://www.unrisd.org/
United Nations Economic and Social Commission for Asia & the Pacific	N	http://www.unescap.org/publications/txtonline.asp

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.

Response	Total = 11
Web sites of non governmental organizations and non profits that do advocacy work, both within the US and beyond (examples without URLs: southeast asian trade union associations)	
Nonprofit public policy research institutes, California Councils of Government, academic research institutes, California governmental agencies.	
none other than identified government sources	
Regional IGO's (e.g., OAS, ASEAN, IADB, etc.)	
Regional/quasi-government sources, e.g. SCAG, ABAG, SANDAG, SACG; Nonprofits and policy institutes, e.g. losangeleslivingwagestudy.org and Public Policy Institute of CA.	
<p>I could include other International Governmental Organization that publish on the web, whose content is of interest to [institution¹] and may be at risk. These include:</p> <p>United Nations Population Fund http://www.unfpa.org/publications/index.cfm</p> <p>United Nations High Commissioner for Refugees http://www.unhcr.ch/cgi-bin/texis/vtx/publ</p>	
university web sites, e.g. University of California at Merced	
If given the opportunity, we would gladly preserve websites from local governments in [county ¹] and the [... ¹] region. We would focus on sites related to environment (primarily water, pollution and land use/development) at the city, county, regional, state and federal level.	
Local governments, including cities, counties, and special districts.	
Collect webpages from quasi-governments such as the Southern California Association of Governments and non-profit organizations such as Health Care Council of Orange County, California	
<p>I would like to focus on local and regional organizations and community groups on both sides of the border, especially those that are partners in the Regional Workbench Consortium (http://regionalworkbench.org), a collaborative research organization headed by the UCSD Urban Studies Program, Center for US-Mexican Studies, Scripps, the San Diego Supercomputer Center and others. This type of digital information, of great importance for local researchers, is particularly in danger of being lost. I am hoping that the tools that CDL creates will allow for automatic crawls/retrieval as well as ad hoc retrievals when individual documents are found.</p>	

1 - Information removed to maintain anonymity.

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

Response	Total = 16
Some of the digitized printed ephemera and many, if not all, of the websites in the collection and elements incorporated into the websites are protected by the U.S. Copyright Law. The Archive makes this material available to researchers on the basis of Fair Use.	

Response	Total = 16
<p>1. Permission from US trade union international and local office (each organizational level runs their own Webs) 2. Permission from non profit publications or newspapers (unions primarily)</p>	
<p>Copyright; Privacy</p>	
<p>Ownership of web-published reports from educational and nonprofit sources.</p>	
<p>none - will focus on non-copyrighted sources</p>	
<p>Are there restrictions indicated on the IGO's page regarding the right of reproduction and redistribution.</p>	
<p>We don't anticipate any - we collect only non-copyrighted material.</p>	
<p>I am currently unclear of the copyright issues associated with materials published by the sources in #18, and if there would even be any issue with capturing and archiving these materials (esp. in light of the recent law suit against the Internet Archive). Also, many local and some state agencies commission private consulting firms to write reports for them (e.g. EIR's or the King Drew Medical Center Navigant Report). Again, I am not clear what kind of intellectual property rights issues I need to address.</p>	
<p>The publications and sites listed are not copyrighted, to the best of my knowledge, however it does not necessarily follow that we have license to grab and store site content without the permission of the international government organizations in question. Obtaining this permission may involve detailed discussion and negotiation. However, in a survey of IGOS [...] in 2001, the overwhelming majority of those asked were willing to allow libraries to archive publicly accessible digital content.</p>	
<p>The copyright status of [...] state and documents and publications of local jurisdictions is not clear. Publications produced under contract to a government agency often are copyrighted. Images on a web page may be copyrighted.</p>	
<p>copyright restrictions</p>	
<p>Possible copyright restrictions for sites which are quasi governmental (for example: http://www.santacruzlafco.org/), or sites which re-package government information (for example: Santa Cruz Municipal Codes online, a lexis nexis site at http://municipalcodes.lexisnexis.com/codes/santacruzco/)</p>	
<p>Copyright: As a state agency we can collect and reproduce documents covered by the state's copyright. (Under [state¹] law, state documents are not automatically part of the public domain as are federal documents; however, agencies seldom enforce copyright and are probably unaware that they hold copyright.)</p> <p>Privacy: Although access to most materials in our collections are accessible under public records laws, some records contain personal information (such as social security numbers) that has been, in effect, protected by the legal concept of practical obscurity. Making these materials available on the web countervenes practical obscurity, forcing us to consider whether we should redact such information from the web version. This problem is not likely to rise with 'publications' but may with many records.</p> <p>Commercial use: [State¹] law requires individuals who use information in public records for a commercial purpose to pay the state for such commercial use.</p> <p>Cultural sensitivity materials: Some materials, especially older works, may contain images or</p>	

Response	Total = 16
information that Native Americans or other groups consider esoteric, ceremonial, or offensive. For example, images of ceremonial dances, works that describe esoteric knowledge to be used only by individuals with appropriate initiation and station within the culture, or images of human remains.	
May need permissions for technical reports created by private contractors under government contract, publications of non-profit organizations, and [... ¹] state government publications	
Documents on webpages may be copyrighted	
Copyright considerations and limitations on access to digital resources. Securing appropriate rights from donors and licenses as necessary to meet access and use objectives.	

1 - Information removed to maintain anonymity.

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

Total	Not at All	A Little	Somewhat	Quite Often	At Least Daily	Don't Know
16						
#	0	3	6	6	1	0
%	0.0	18.8	37.5	37.5	6.3	0.0

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.)

Response	Total	#	%
	16		
a. Yes		13	81.3
b. No		3	18.8

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

Response	Total = 12
quarterly or semi-annually would suffice	
varies as to content	
Quarterly	
Ideally we would not have to re-acquire the materials. however if the site has changed after the initial harvest we would re-acquire the materials.	
Every 3 mos; Is it possible after evaluating some results to change this? it is almost an agency by agency case; some info i know gets updated daily; some once a month;some every 6 months or never.	
Monthly crawls would probably be best.	

Response	Total = 12
Not determined at this time.	
intervals vary depending on the site, 6 mos. to 2 years	
at least twice a year	
We adjust the frequency of capture to the relative importance of the source of the materials. The most important agencies and officials' sites may be captured monthly, others quarterly, and some (Continued via email:) semi-annually. We may adjust this in the near future to weekly for the most important and monthly for all the others.	
Unsure if we would continue to collect after the grant period	
Depends on the material and digital archiving policy and practice. Perhaps annually or semi-annually	

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.)

Response	Total	#	%
	16		
a. Yes		9	56.3
b. No		7	43.8

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.)

Response	Total	#	%
	16		
Allow selection and let browser provide standard messages for broken links		7	43.8
Allow selection but provide custom messages for broken links		7	43.8
Deny selection but leave text with no notification of broken links		0	0.0
Deny selection but leave text with notification of broken links		1	6.3
Other		1	6.3

If other, please explain

Response	Total = 1
Links to materials not archived are included but not active. (People will know where the link pointed to, but will have to take an extra step to cut and paste the link to help underscore they are leaving the site.) Links to materials in the archives are mangled so that they continue to point within the archives, not to live materials outside the site, so these links shouldn't break over time.	

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

Response	Total	#	%
	16		
a. Yes		2	12.5
b. No		12	75.0
c. Don't Know		2	12.5

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

Respondents	Total	%
	15	93.8
Responses	Total	#
	22	
Multiple versions captured at different points in time		10
Addition of enhanced metadata to captured materials		3
Multiple formats of the same object (e.g., .txt and .pdf)		9

Note: The following was received via email from one participant:

“We weren't sure about this question. If a website has both a txt and pdf version of the same file, they may both be authentic. The presense of two versions does not raise risks of authenticity so much as reliability; if the content differs, someone might argue (in litigation) that they relied on one, not the other.

If you're talking about keeping two versions of a document, with one being created by the archival repository, the situation may be slightly different. Authenticity may be a problem unless you can distinguish which was taken from the website and which was created by the repository. If you can demonstrate the process by which the copy was made and when the transformation was done, I don't think there would be much, if any, threat to authenticity.”

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

Response	Total	#	%
	16		
Content provider		8	50.0
Web archive creator or curator		5	31.3
End users		0	0.0
Other		3	18.8

If other, please explain.

Response	Total = 3
A combination of all of the above, especially the level of trusted methods used by the repository and curator, and end user opinion...	
I am not sure about this. This point would need to be discussed with the agencies.	
We can only certify the authenticity of the document as something distributed on an agency's website. Someone else would have to ensure that what was on the website when we captured it was authentic.	

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

	Total	Not Challenging	A Little Challenging	Somewhat Challenging	Very Challenging	Extremely Challenging
Needs assessment	15					
#		3	7	1	4	0
%		20.0	46.7	6.7	26.7	0.0
Contract negotiation	14					
#		1	5	2	6	0
%		7.1	35.7	14.3	42.9	0.0
Copyright/intellectual property issues	16					
#		1	4	5	5	1
%		6.3	25.0	31.3	31.3	6.3
Initial hardware & software implementation	16					
#		2	0	8	5	1
%		12.5	0.0	50.0	31.3	6.3

The Web at Risk: Needs Assessment Survey Report

	Total	Not Challenging	A Little Challenging	Somewhat Challenging	Very Challenging	Extremely Challenging
Harvest	16					
#		3	4	6	2	1
%		18.8	25.0	37.5	12.5	6.3
Network access	16					
#		5	4	4	2	1
%		31.3	25.0	25.0	12.5	6.3
Storage	16					
#		2	5	4	2	3
%		12.5	31.3	25.0	12.5	18.8
Cataloging	16					
#		0	2	2	7	5
%		0.0	12.5	12.5	43.8	31.3
Presentation	16					
#		0	4	7	4	1
%		0.0	25.0	43.8	25.0	6.3
Re-harvest	16					
#		1	6	5	3	1
%		6.3	37.5	31.3	18.8	6.3
Management & deselection	16					
#		0	4	7	5	0
%		0.0	25.0	43.8	31.3	0.0
Preservation	16					
#		1	0	5	3	7
%		6.3	0.0	31.3	18.8	43.8
IT Support	15					
#		1	0	5	4	5
%		6.7	0.0	33.3	26.7	33.3
Staff Training	16					
#		1	0	7	7	1
%		6.3	0.0	43.8	43.8	6.3

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

- a. Hardware and software maintenance
- b. Unclear collection boundaries in the web environment
- c. Maintenance of look and feel of original material
- d. Metadata creation
- e. Password protected source material
- f. Encrypted source material
- g. Authenticity
- h. Persistent naming
- i. Dynamic nature of some web materials
- j. Frequency of change
- k. Real-time content changes during capture

	Total	Not Challenging	A Little Challenging	Somewhat Challenging	Very Challenging	Extremely Challenging	Don't Know
a.	16						
#		1	3	4	5	2	1
%		6.3	18.8	25.0	31.3	12.5	6.3
b.	16						
#		0	4	8	2	1	1
%		0.0	25.0	50.0	12.5	6.3	6.3
c.	15						
#		0	2	5	2	4	2
%		0.0	13.3	33.3	13.3	26.7	13.3
d.	16						
#		1	1	1	10	3	0
%		6.3	6.3	6.3	62.5	18.8	0.0
e.	15						
#		1	0	3	2	9	0
%		6.7	0.0	20.0	13.3	60.0	0.0
f.	16						
#		1	0	4	2	9	0
%		6.3	0.0	25.0	12.5	56.3	0.0
g.	16						
#		2	0	8	4	2	0
%		12.5	0.0	50.0	25.0	12.5	0.0
h.	16						
#		0	2	6	5	2	1
%		0.0	12.5	37.5	31.3	12.5	6.3

The Web at Risk: Needs Assessment Survey Report

	Total	Not Challenging	A Little Challenging	Somewhat Challenging	Very Challenging	Extremely Challenging	Don't Know
i.	16						
#		0	4	0	4	8	0
%		0.0	25.0	0.0	25.0	50.0	0.0
j.	16						
#		0	3	7	4	2	0
%		0.0	18.8	43.8	25.0	12.5	0.0
k.	16						
#		1	5	1	3	1	5
%		6.3	31.3	6.3	18.8	6.3	31.3

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

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

Total	Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
16					
#	0	0	0	3	13
%	0.0	0.0	0.0	18.8	81.3

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

Total	Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
16					
#	0	0	1	9	6
%	0.0	0.0	6.3	56.3	37.5

The Web at Risk: Needs Assessment Survey Report

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.

Total	Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
16					
#	0	1	7	5	3
%	0.0	6.3	43.8	31.3	18.8

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

Total	Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
16					
#	0	0	0	3	13
%	0.0	0.0	0.0	18.8	81.3

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

Respondents	Total	%
	15	93.8%
Responses	Total	#
	45	
Usage data thresholds		7
Sensitive or offensive material		6
Copyright violation		14
Fraud		11
Storage costs		7

34. What additional deselection criteria will you use?

Response	Total = 9
Takedown request by copyright owners.	
Enduring value in relation to the universe of documentation	
Available to our end users elsewhere / redundancy with other digital collections.	
incomplete/corrupted files	

Response	Total = 9
Corrupted files - if a file or document can no longer be opened/used; Dated or Superseded material - this mirrors the way we handle our print collections; Also, I would use a combination of the above criteria - if storage cost becomes too high, then I'd look at low use materials to determine deselection. As for fraudulent material, I'm not sure I'd want it removed; I'd want to find a way to note the issue for take down if I had to remove it.	
Once archived, it would be my preference not to deselect material without an extraordinarily good reason, for example express written direction of the issuing agency.	
lack of relevancy of particular data captures to local research needs; duplicative capture (no change to site during selected timeframe for capture)	
Does the material continue to support the university's academic mission (programs, degrees, etc.)	
Agency publications will not be deselected. Archival materials' value may be reassessed over time to ensure that previous appraisal decisions that materials are of permanent value remain valid.	

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

Total	Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
16					
#	4	10	1	1	0
%	25.0	62.5	6.3	6.3	0.0

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

Total	Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
16					
#	6	9	0	1	0
%	37.5	56.3	0.0	6.3	0.0

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.

Total	Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
16					
#	1	4	7	4	0
%	6.3	25.0	43.8	25.0	0.0

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

Total	Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
16					
#	0	0	0	15	1
%	0.0	0.0	0.0	93.8	6.3

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

Total	Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
16					
#	2	7	3	4	0
%	12.5	43.8	18.8	25.0	0.0

Section D. Preservation

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

Total	Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
16					
#	2	4	5	5	0
%	12.5	25.0	31.3	31.3	0.0

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

Total	Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
16					
#	0	1	2	11	2
%	0.0	6.3	12.5	68.8	12.5

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.

Total	Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
16					
#	0	1	5	9	1
%	0.0	6.3	31.3	56.3	6.3

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

Total	Strongly Disagree	Disagree	Neither Disagree nor Agree	Agree	Strongly Agree
16					
#	0	4	8	4	0
%	0.0	25.0	50.0	25.0	0.0

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.

- a. Migration to new version of same software (e.g., from version 2 to 6 of Microsoft Word)
- b. Migration to different format (e.g., text to pdf)
- c. Migration to different hardware platforms
- d. Migration to different operating system environments
- e. Migration to different file system within an operating system environment

	Total	No Threat	Small Threat	Moderate Threat	Significant Threat	Extreme Threat	Don't Know
a.	16						
#		2	4	5	1	1	3
%		12.5	25.0	31.3	6.3	6.3	18.8
b.	16						
#		0	7	4	3	0	2
%		0.0	43.8	25.0	18.8	0.0	12.5
c.	16						
#		1	4	2	5	2	2
%		6.3	25.0	12.5	31.3	12.5	12.5
d.	16						
#		0	5	1	5	3	2
%		0.0	31.3	6.3	31.3	18.8	12.5

	Total	No Threat	Small Threat	Moderate Threat	Significant Threat	Extreme Threat	Don't Know
e.	16						
#		0	4	2	6	1	3
%		0.0	25.0	12.5	37.5	6.3	18.8

Section E. Curator User Interface

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?

- a. Total crawl size
- b. Content object types (image, audio, video, etc.)
- c. Content object formats (html, jpeg, gif, pdf, etc.)
- d. Total file size by type
- e. # Links to external URLs
- f. Content URLs within the targeted or entry-point URLs
- g. # Broken Links within targeted or entry-point URLs
- h. Failures by # and Type (timeouts, server errors, unsupported schemes such as 'mailto')

	Total	Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
a.	16					
#		1	4	6	4	1
%		6.3	25.0	37.5	25.0	6.3
b.	16					
#		1	1	4	8	2
%		6.3	6.3	25.0	50.0	12.5
c.	16					
#		1	0	5	6	4
%		6.3	0.0	31.3	37.5	25.0
d.	16					
#		0	6	7	3	0
%		0.0	37.5	43.8	18.8	0.0
e.	16					
#		1	4	7	4	0
%		6.3	25.0	43.8	25.0	0.0

The Web at Risk: Needs Assessment Survey Report

	Total	Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
f.	16					
#		1	1	6	7	1
%		6.3	6.3	37.5	43.8	6.3
g.	16					
#		0	1	6	8	1
%		0.0	6.3	37.5	50.0	6.3
h.	16					
#		0	2	6	7	1
%		0.0	12.5	37.5	43.8	6.3

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

Response	Total = 6
The website's navigational method, e.g. Javascript; Flash Multiple domains, e.g. sites to spread across multiple domains.	
Contents!	
Identification of new files of specified type added since last crawl. Extremely important.	
completeness - confirmation that all files on the site have been captured	
Location within the logical file system is extremely important. We want to be able to include or exclude materials in different directories or subdirectories on the website.	
I can't think of any other attributes at this time. However, I think it would be important to be able to change attributes over time – i.e., when new file formats come into existence... It's also important to be able to review the robots.txt file or each top level url to make sure the crawl is compliant with the site administrator's wishes.	

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

- a. Frequency of the crawl (daily, weekly, monthly, etc.)
- b. Time period over which to repeat crawl (1 month or 6 months at specified frequency)
- c. # Levels within targeted or entry-point URLs to capture
- d. Depth of links to external URLs to capture
- e. Compliance with robot exclusions (obey or ignore)
- f. Content object types to capture (image, audio, video, etc.)
- g. Content object formats to capture (html, jpeg, gif, pdf, etc.)

The Web at Risk: Needs Assessment Survey Report

	Total	Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
a.	16					
#		1	1	3	5	6
%		6.3	6.3	18.8	31.3	37.5
b.	16					
#		1	1	4	4	6
%		6.3	6.3	25.0	25.0	37.5
c.	16					
#		0	0	1	7	8
%		0.0	0.0	6.3	43.8	50.0
d.	16					
#		0	2	5	6	3
%		0.0	12.5	31.3	37.5	18.8
e.	16					
#		0	0	9	3	4
%		0.0	0.0	56.3	18.8	25.0
f.	16					
#		0	0	6	5	5
%		0.0	0.0	37.5	31.3	31.3
g.	16					
#		0	0	5	6	5
%		0.0	0.0	31.3	37.5	31.3

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

Response	Total = 4
The date of the crawl (for one time crawls)	
Languages; Has the site changed?	
Projected length of crawl, which may be dynamically revised during its run.	
Keywords, stop criteria, search in PDF files, language, number of external links linking to the document.	

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?

Total	Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
15					
#	0	2	3	10	0
%	0.0	13.3	20.0	66.7	0.0

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?

- a. Crawl completion status by targeted or entry-point URL
- b. Total size captured
- c. Content object types captured (image, audio, video, etc.)
- d. Content object formats captured (html, jpeg, gif, pdf, etc.)
- e. Total file size by object type & format
- f. Errors encountered by error code (200, 300, 400, 404, 500, etc.)

	Total	Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important	Don't Know
a.	16						
#		1	1	4	4	5	1
%		6.3	6.3	25.0	25.0	31.3	6.3
b.	16						
#		0	2	8	3	2	1
%		0.0	12.5	50.0	18.8	12.5	6.3
c.	16						
#		1	4	4	2	4	1
%		6.3	25.0	25.0	12.5	25.0	6.3
d.	15						
#		1	4	1	4	4	1
%		6.7	26.7	6.7	26.7	26.7	6.7
e.	16						
#		2	1	7	4	1	1
%		12.5	6.3	43.8	25.0	6.3	6.3
f.	16						
#		0	2	6	4	3	1
%		0.0	12.5	37.5	25.0	18.8	6.3

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

Response	Total = 4
Answers in 50 based on the meaning of "real time"--it would be acceptable to receive a report on the crawl at the end of the process and analyze it then	
per-cent completion per targeted entry-point	
Identification of new files of specified type since last crawl. Extremely important.	
crawler must be able to validate content	

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.

- a. Curator for each crawl in the collection
- b. Crawl completion date(s)
- c. Targeted or entry-point URLs for each crawl
- d. Content URLs within targeted or entry-point URLs for each crawl
- e. Parameters of each crawl
- f. Total size of each crawl
- g. Total collection size by type & format
- h. # Errors encountered for each crawl by error code (200, 300, 400, 404, 500, etc.)
- i. Measurement of content change over time

	Total	Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
a.	16					
#		0	3	5	4	4
%		0.0	18.8	31.3	25.0	25.0
b.	16					
#		0	1	1	4	10
%		0.0	6.3	6.3	25.0	62.5
c.	16					
#		0	0	0	7	9
%		0.0	0.0	0.0	43.8	56.3
d.	16					
#		0	1	3	8	4
%		0.0	6.3	18.8	50.0	25.0

The Web at Risk: Needs Assessment Survey Report

	Total	Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
e.	16					
#		0	0	1	10	5
%		0.0	0.0	6.3	62.5	31.3
f.	16					
#		0	3	9	4	0
%		0.0	18.8	56.3	25.0	0.0
g.	16					
#		0	1	7	8	0
%		0.0	6.3	43.8	50.0	0
h.	16					
#		0	0	8	7	1
%		0.0	0.0	50.0	43.8	6.3
i.	15					
#		0	1	1	6	7
%		0.0	6.7	6.7	40.0	46.7

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

Response	Total = 2
Identification of new files of specified type since last crawl.	
Language	

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.

	Total	Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
URL	16					
#		0	0	0	3	13
%		0.0	0.0	0.0	18.8	81.3
Size	16					
#		0	3	8	3	2
%		0.0	18.8	50.0	18.8	12.5

The Web at Risk: Needs Assessment Survey Report

	Total	Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
Type (image, audio, video, etc.)	16					
#		0	2	3	5	6
%		0.0	12.5	18.8	31.3	37.5
Format (html, jpeg, gif, pdf, etc.)	16					
#		0	2	1	4	9
%		0.0	12.5	6.3	25.0	56.3
Title	16					
#		0	0	0	6	10
%		0.0	0.0	0.0	37.5	62.5
Author	16					
#		0	0	2	3	11
%		0.0	0.0	12.5	18.8	68.8
Subject	16					
#		0	1	3	4	8
%		0.0	6.3	18.8	25.0	50.0
Description	16					
#		0	2	2	5	7
%		0.0	12.5	12.5	31.3	43.8
Creation Date	16					
#		0	0	0	8	8
%		0.0	0.0	0.0	50.0	50.0
Object name (e.g. filename)	16					
#		0	1	4	4	7
%		0.0	6.3	25.0	25.0	43.8
Language	16					
#		0	3	6	6	1
%		0.0	18.8	37.5	37.5	6.3
Archived Date	16					
#		0	0	1	5	10
%		0.0	0.0	6.3	31.3	62.5

The Web at Risk: Needs Assessment Survey Report

	Total	Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
Measurement of change over time	16					
#		0	0	4	8	4
%		0.0	0.0	25.0	50.0	25.0

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

Response	Total = 2
Structure/levels, change therein	
# of libraries who have cataloged the document, number of downloads of and number of links to the object (i.e., in google or other search engines)	

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

Respondents	Total		%
	15		93.8
Responses	Total	#	%
	44		
Object level (Example: images or movies)		10	66.7
Web page level (Example: .html or .xml files)		10	66.7
Logical document level (Example: article spanning multiple .html files)		10	66.7
Website level (Example: all content within a targeted or entry-point URL)		13	86.7
Other		1	6.7
Aggregates (information about collections and series to which documents belong. Hash value to demons			

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.

	Total	Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
URL	16					
#		1	1	0	2	12
%		6.3	6.3	0.0	12.5	75.0

The Web at Risk: Needs Assessment Survey Report

	Total	Not Important	A Little Important	Somewhat Important	Very Important	Extremely Important
Date/Time of Capture	15					
#		2	1	3	2	7
%		13.3	6.7	20.0	13.3	46.7
Object Format/Type	16					
#		0	1	3	8	4
%		0.0	6.3	18.8	50.0	25.0
Language	16					
#		1	3	5	5	2
%		6.3	18.8	31.3	31.3	12.5
File Size	16					
#		2	2	7	2	3
%		12.5	12.5	43.8	12.5	18.8
File Name	16					
#		1	2	5	3	5
%		6.3	12.5	31.3	18.8	31.3
Author	16					
#		1	0	0	2	13
%		6.3	0.0	0.0	12.5	81.3
Title	16					
#		1	0	0	3	12
%		6.3	0.0	0.0	18.8	75.0

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

Response	Total = 9
Controlled Vocabulary.	
Topical; Keyword	
Added entry for organizational author, publisher, date of publication.	
subjects or topics,e.g. housing, water, air pollution,etc -- will this be handled by the metadata?	
Subject, full-text, provenance (agency browse)	
subject designation	
keyword and boolean searching	
Agency of origin (publisher/provenance).	

Response	Total = 9
Agency or aegis, date of publication, subject, geographic area	

59. We welcome any additional comments you may have.

Response	Total = 6
I think it is important for the metadata to support cross-collection searching and integration with existing discovery systems.	
It will be interesting to evaluate my responses once I've interacted with the tools and had an opportunity to analyze the contents of a crawl. I had a very difficult time try to designate the re-acquire interval (question #21). Regarding some the questions about our end users -- I'm not sure how much users think about many of those access issues (at this point in time). In the web environment, it almost seems that they will take what they can get and they are not concerned with what is happening behind the scenes. That will probably change over time?	
I had to complete this surevy under some time pressure because I was informed that international government information could be included after the deadline. I apologize in advance for any inconsistences in my responses. If further information is required I would be happy to provide. [Name and Institution ¹]	
We have no current experience with large-scale harvesting of material from the web. Our answers might be quite a bit different once we have had some experience, but these responses reflect our best guesses.	
I should provide more context around my answers to #28 of this survey. Our campus is currently consolidating all IT functions across campus, so the level of technical support that will be available is unknown at this point. Currently, our Computer and Network Services unit is very helpful and responsive. My library is supportive this project and I am hopeful that the IT consolidation will not significantly diminish this level of support.	
The questions for which I have not replied indicate a "no" or "none". It would be very helpful to review the survey answers after the first couple of crawls to determine if we have the same or better understanding of the issues represented by the questions.	

1 - Information removed to maintain anonymity.

Appendix G. Respondents' Most Important Digital Collections

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

Collection Name	Collection Description (Total 9 Curators with 29 Collections)	Collection Location
UCLA Online Campaign Literature Archive	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.	http://digital.library.ucla.edu/campaign/
eScholarship Working Paper Repository	Faculty working paper series, labor union contracts, program materials	http://repositories.cdlib.org
Labor Research Portal	Guides to the Web, labor research and labor information	http://www.iir.berkeley.edu/library
Labor Contracts Database	mirror of eScholarship repository	http://www.iir.berkeley.edu/library
Numeric Social Science/Government Data		http://library.stanford.edu/services/social_sci_data_soft/data.html
Social Science Databases & Indexes		http://library.stanford.edu/catdb/ssi.html
Government Information Databases & Indexes		http://library.stanford.edu/catdb/govinfo.html
Digital Collections (Created Locally)		http://library.stanford.edu/depts/green/about/rooms/ssrc/digitalcollections.html
CyberCemetery	This site provides permanent public access to the Web sites and publications of defunct U.S. government agencies and commissions.	http://govinfo.library.unt.edu/

Collection Name	Collection Description (Total 9 Curators with 29 Collections)	Collection Location
Congressional Research Service Reports	The Congressional Research Service (CRS) does not provide direct public access to its reports, requiring citizens to request them from their Member of Congress. Some Members, as well as several non-profit groups, have posted the reports on their Web sites. This site aims to provide integrated, searchable access to many of the full-text CRS reports that have been available at a variety of different Web sites since 1990.	http://digital.library.unt.edu/govdocs/crs/
Texas Register Archive	The online edition of the Texas Secretary of State publication the Texas Register, with issues going back to June 1991.	http://texinfo.library.unt.edu/texasregister/default.htm
Gammel's Nineteenth Century Laws of Texas	H.P.N. Gammel's The Laws of Texas, 1822-1897 has long been one of the most important primary resources for the study of Texas' complex history during the Nineteenth Century. His monumental compilation charts Texas from the time of colonization through to statehood and reveals Texas' legal history during crucial times in its development. The Laws consist of documents not only covering each congressional and legislative session but comprise other documents of significance, including the constitutions, select journals from the constitutional conventions, and early colonization laws. The first ten volumes of The Laws of Texas are available from this site, along with the Analytical Index.	http://texinfo.library.unt.edu/lawssoftexas/default.htm
City of Los Angeles Community Plan Profiles [1995]	Digitized version of the LA Community plans, includes history of the various communities, land use and zoning maps. *Note this is really the only "digital collection" we have.	http://www.library.ucla.edu/libraries/yr/referenc/plans/laprofiles1995/index.html
United Nations Research Institute for Social Development (UNRISD)	Autonomous agency of the United Nations that conducts research on the social dimensions and problems affecting international development.	http://www.unrisd.org/
United Nations Economic and Social Commission for Asia & the Pacific	Regional arm of the United Nations Secretariat in the Asia & Pacific Region. Focuses on promoting economic and social development through regional and subregional cooperation and integration.	http://www.unescap.org/publications/txtonline.asp

Collection Name	Collection Description (Total 9 Curators with 29 Collections)	Collection Location
United Nations Children's Fund (UNICEF) Innocenti Research Centre	The Innocenti Research Centre is the main research arm of the United Nations Children's Fund. The centre is charged with monitoring the impact of social and economic policies on children and advocating to support the implementation of international standards on the rights of the child.	http://www.unicef-icdc.org/publications/
UNAIDS: Joint United Nations Research Programme on HIV/AIDS	The Joint United Nations Programme on HIV/AIDS is the main advocate for coordinated global action on the AIDS epidemic. It's mission is to lead, strengthen and support a response to HIV and AIDS that includes preventing transmission of HIV, providing care and support to those already living with the virus, reducing the vulnerability of individuals and communities to HIV, and alleviating the impact of the epidemic.	http://www.unaids.org/en/resources/publications.asp
Numeric Data	extensive numeric data collections including data from ICPSR and the Roper Center for Public Policy Research, government agencies at both the national and state levels and inter-governmental agencies.	http://library.stanford.edu/services/social_sci_data_soft/data.html
Social Science Databases	bibliographic and content databases supporting research in the social sciences.	http://library.stanford.edu/catdb/ssi/html
Digital Collections	social science data collections which were digitized by Stanford Libraries to support Stanford researchers in the social sciences.	http://library.stanford.edu/depts/green/about/rooms/ssrc/digitalcollections.html
Government Document Databases	bibliographic and full-text databases providing access to U.S. federal, international, and state and local documents.	http://library.stanford.edu/catdb/govinfo.html
Historical photographs	Archives' photographs focus upon the unique cultural heritage of the state and territory of Arizona, beginning in 1863. The collection includes about 90,000 images, including photographs, slides, negatives, glass plate negatives, tintypes, transparencies, postcards and others, of which about 30,000 are currently digitized.	http://photos.lib.az.us

Collection Name	Collection Description (Total 9 Curators with 29 Collections)	Collection Location
State agency publications	The Law and Research Library includes a depository program that preserves and provides access to all state agency publications. The collection dates from the Territorial period to the present, and includes annual and special reports, serials, and monographs created by or under contract to the State. Many reports previously printed are now available only on the web. Legislative Study Committee Reports have been scanned and made available online. Other reports are being captured into CONTENTdm, but are not yet available online.	http://www.lib.az.us/is/state/lsc/ (Study Committee Reports only; others not yet online)
Web SafetyNet Archives	The Law and Research Library is participating with the Illinois State Library and the University of Illinois at Urbana-Champaign to test software to capture state agency websites. Many items on these sites are properly part of the state agency depository collection or archival collections. However, this collection includes many items that would not normally be selected for preservation because of their limited value.	http://wap.lib.az.us/
Alt Fuels Court Records	Court records used in litigation resulting from lawsuits relating to state subsidies for cars to use fuels other than gasoline. The collection includes scans of government agency paper documents that are not yet – and may never – be in the collections.	Not yet online
Visual Arts Slide Images	Over 200,000 digitized art slides incorporated into ArtStor, but not Web accessible via UCSD.	
Mandeville Special Collections Digital exhibits		http://orpheus.ucsd.edu/speccoll/online.html
California explores the ocean	CEO provides access for the citizens of California to a diverse array of resources about the ocean and ocean exploration from the unique collections of the Scripps Institution of Oceanography Archives and Library and to a selection of ocean related photographs and oral histories from the collections of the San Diego Historical Society.	http://ceo.ucsd.edu/

Collection Name	Collection Description (Total 9 Curators with 29 Collections)	Collection Location
Social Sciences Data Collection	<p>SSDC is a collection of numeric social science data including economic, public opinion, survey research, administrative, election, and census data. This is a digital-only collection. The files consist of raw data that students and faculty analyze with statistical software. These are not printable tables of statistics, but raw data for analysis. The collection is used primarily by students and faculty in Economics, Sociology, Political Science, and Urban Studies.</p> <p>The current collection size is over 350 titles including just under one hundred gigabytes in more than seven thousand files. These include both data files and "metadata" files of various kinds ("codebooks" and other kinds of metadata that describe the contents of the data files).</p>	<p>http://ssdc.ucsd.edu</p>