ARTICLES

Developing Collections of Web-Published Materials

Inga K. Hsieh
Kathleen R. Murray
Cathy Nelson Hartman

ABSTRACT. Librarians and archivists face challenges when adapting traditional collection development practices to meet the unique characteristics of Web-published materials. Likewise, preservation activities for Web-published materials must be undertaken at the outset of collection development lest they be lost forever. Standards and best practices for Web-collection development are still emerging, and librarians are struggling with the often daunting financial, staffing, and infrastructure challenges posed by collecting and preserving these materials. The results of a needs assessment with librarians, information providers, and academic researchers informed the identification of key collection development activities for Web-published materials. This research was conducted as part of the Web-at-Risk project, a collaborative effort of the California

Inga K. Hsieh is Research Assistant for the Web-at-Risk project (E-mail: ikh0003@unt.edu).

Kathleen R. Murray is Post-Doctoral Research Associate, University Libraries, University of North Texas, and Assessment Analyst for the Web-at-Risk project (E-mail: krmurray@unt.edu).

Cathy Nelson Hartman is Assistant Dean, Digital and Information Technologies, University Libraries, University of North Texas, and Project Manager for the Web-at-Risk project (E-mail: chartman@library.unt.edu).

Journal of Web Librarianship, Vol. 1(2) 2007
Available online at http://jwl.haworthpress.com

doi:10.1300/J502v01n02_02

Digital Library, the University of North Texas, and New York University.

doi:10.1300/J502v01n02_02 *[Article copies available for a fee from The Haworth Document Delivery Service: 1-800-HAWORTH. E-mail address: <docdelivery@haworthpress.com> Website: <http://www.HaworthPress.com> © 2007 by The Haworth Press, Inc. All rights reserved.]*

KEYWORDS. Web archives, digital archives, Web collections, collection development, Web preservation, digital preservation

INTRODUCTION

Developing collections of Web-published materials presents new challenges to the traditional practice of collection development. Web-published materials are those that are accessed and presented via the World Wide Web. These 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 and are increasingly used by a variety of end users, including scholars and researchers. Yet many of these materials are at risk of disappearing over time from their original locations. Web archives seek to reduce this risk by preserving Web-published materials for future access.

Standards and best practices for Web archives are emerging through the work of many national and international projects and collaborations.[1] Likewise, key considerations for the preservation of collections of Web-published materials in Web archives are emerging as librarians and libraries attempt to meet the information needs of their end users. Common issues and challenges identified by the Web-at-Risk project in a needs-assessment study conducted in 2005 informed the Web collection development considerations reported in this article.[2]

The Web-at-Risk project is a three-year collaborative research effort of the California Digital Library, the University of North Texas, and New York University. It was funded in 2004 by the Library of Congress as part of the National Digital Information Infrastructure and Preservation Program to begin building a national preservation network for "at-risk digital materials of significant cultural and historical value to the nation."[3] The project team works with a group of librarians who identify Web collections that they will build and manage via a Web-archiving service being prototyped at the California Digital Library. The needs-assessment activities of the project identified the needs and issues faced by librarians, curators, end users, and content providers regarding collection development for Web-published materials and Web archives.

Terminology

In the interest of effective discourse with the range of participants in this research, the project team defined key concepts at the outset. Listed below are the key definitions used in this study:

- An *Archive Service Provider* is an organization that offers repository services for other institutions or organizations that wish to create and preserve collections of Web-published materials but do not have the infrastructure to capture, store, and preserve the materials themselves.
- A *Digital Archive* is a collection of digital objects that may also exist in other forms. A digital archive preserves the digital formats for posterity and provides access to them.
- *Digital Objects* include interactive works such as video games, sensory presentations such as music, documents such as articles, and data such as statistical datasets. Two types of digital objects included in digital archives are born-digital objects and digital surrogates (e.g., digitized copies of print books or audio tapes).
- *Web-Published Materials* are materials accessed and presented via the Web. The materials include a range of material types from text documents to streaming video to interactive experiences. All Web-published materials are digital objects.
- A *Web Archive* contains Web-published materials for which an organization has accepted long-term responsibility for both preservation and access.
- A *Web Collection* typically consists of a group of related Websites but might also refer to a group of related Web-published materials. In the context of this article, Web collections are assumed to be preserved in a Web archive.
- A *Website* consists of one or more Web pages and other Web-published materials that are generally related in some way and are often within the same domain or sub-domain name space (e.g., unt.edu or library.unt.edu).

BACKGROUND

In order to provide important, relevant, and desirable resources to patrons, librarians have long developed collections of print and other analog materials. With increasing frequency, resources that meet the needs of patrons are published on the Web–often with no print counterpart.

In 2001, Louis Pitschmann remarked that "a growing number [of Websites] contain information not readily available elsewhere."[4] In a later study to assess the feasibility of preserving Web-published materials, Michael Day reported: "In the short time since its invention, the World Wide Web has become a vital means of facilitating global communication and an important medium for scientific communication, publishing, e-commerce, and much else."[5] More recently, a 2006 report from the Digital Preservation Coalition recounts a growing dependence on digital information evidenced in part "by the increased use of online publications (many of which no longer exist in paper form)" and "by our growing reliance on the internet as a source of information."[6]

To continue to meet patrons' needs, librarians must take steps to include Web-published materials in their collections. As with materials published in print format, many Web-published materials are created by authoritative sources. Stephen G. Nichols and Abby Smith pointed out these Web-published materials "have the same claim to intrinsic value that any intellectual property created with analog technologies would have."[7]

The volume of material that is Web-published, however, is overwhelming. The ease with which materials can be published on the Web results in significant quantities of information published by sources with little or no generally recognized authority regarding the subject content. Likewise, information can be misleading or simply wrong.

When selecting print publications, librarians play a critical evaluative role. As subject specialists, librarians select print materials for inclusion in library collections based on the reputation of the author or publisher or on the authority of the content as established by a peer-review process. Membership in a library's collection often confers reassurance to patrons that the materials are authoritative and deemed of value by a librarian with subject expertise. Pitschmann stated the evaluative role librarians perform with respect to print materials can be extended to Web-published materials:

> The value-added services that libraries have traditionally provided for print formats need to be applied to free web-based resources as well. The selection and cataloging functions of a physical library assure users that titles found there have met predetermined quality criteria and have been described in a manner that facilitates their identification and retrieval. Moreover, the cataloging process provides the authoritative and consistent grouping of related materials so vital to browsing and to the winnowing-and-sifting process that characterizes learning and research in an academic setting.[8]

In the Web-at-Risk assessment, many librarians reported spending a large amount of time selecting and evaluating Web-published materials. Particularly in emerging academic disciplines, such as modern cultural studies, the available information is often Web-born or Web-published. Librarians who select materials for these disciplines are spending an increasing amount of their time identifying Web-published materials and linking to these resources from Web-based subject guides.

Preserving Web-Published Materials

All too frequently, librarians identify important Web-published materials and discover later the materials have disappeared from the Web. Researchers warn that if efforts to capture and preserve Web-published materials are not undertaken, the materials will disappear.[9,10] After capture, however, Web-published materials have unique preservation requirements that present new challenges to libraries. "Digital preservation differs from 'traditional preservation' in that digital information needs to be actively and continuously managed," the Digital Preservation Committee claimed. "It cannot be simply stored and left on a shelf."[11]

In 2004, the National Information Standards Organization Framework Advisory Group created *A Framework of Guidance for Building Good Digital Collections*.[12] Although much of the framework is presented in terms of building collections of digitized materials (i.e., digital surrogates for analog materials), it is "intended to be flexible enough to accommodate new principles, considerations, and resources."[13] Web collections are a type of digital collection, and as a result, the principles included in the NISO framework readily transfers to Web collections. Because Web collections typically include materials that are born digitally and have no analog counterpart, the NISO framework's Principle 3, which states that a "collection should be sustainable over time,"[14] is particularly important for Web collections.

Although many librarians are aware of the need to identify and preserve Web-published materials, they are unclear how to begin. A respondent to one of the Digital Preservation Coalition's surveys in 2005 summed it up this way: "There is a high-level feeling that [something] should be done, but with no practical action. The lack of action seems to be due mostly to not knowing how to proceed."[15] The collection planning guidelines presented here provide librarians and others interested in collecting Web-published materials with a place to begin.

WEB-COLLECTION PLANNING GUIDELINES

Not unlike the role collection plans often serve for traditional collection development within a library, a Web-collection plan can guide the librarian's work by articulating the role a Web collection has within the organization, describing its characteristics, and identifying the organization's commitment to its preservation. Collection development activities for Web-published materials typically fall into three main phases–selection, curation, and preservation. Table 1 identifies the main activities involved in each phase. These activities are not always conducted in a linear fashion, and requirements in one activity may influence tasks in other activities. For example, preservation responsibilities will often dictate tasks in both the selection and curation phases.

Web-collection plans address each activity identified in Table 1. While these activities conceptually parallel activities for print materials, most librarians find more time and effort is required to complete them.[16] Additionally, there are unique collection development considerations that must be addressed for Web-published materials. These considerations are highlighted in the following sections, which correspond to the recommended sections to include in a Web-collection plan:

- Mission and scope
- Selection and acquisition
- Description and organization
- Presentation and access requirements
- Maintenance and weeding
- Preservation
- Supporting appendices

Section 1: Mission and Scope

As with traditional collections, a Web collection is based on a clearly articulated mission and scope that guide collection development. A

TABLE 1. Web Collection Development Activities

Selection Phase	Curation Phase	Preservation Phase
• Selection • Acquisition	• Description • Organization • Presentation • Maintenance • Deselection	• Preservation

collection plan should begin by articulating the mission of the institution and clearly identifying the group or groups served by the collection. Understanding the unique characteristics and needs of each user group will influence collection development activities in all phases, including what to collect and the metadata required for information discovery. Assessing how targeted end users employ Web-published materials to carry out their organizational or professional responsibilities can help both to identify gaps in existing collections and to prioritize new materials for inclusion in collections.

Websites in a collection may be related by a common subject area or theme, such as local genealogy resources, or related to a common event, such as a national election. A scope statement included in the collection plan identifies the subject area, theme, or event that unites the materials in the Web collection and describes how the collection supports the mission of the library, organization, or institution.

Section 2: Selection and Acquisition

Selection is a critical part of Web-collection development and perhaps the most difficult. According to Day, "While the Web contains much that would definitely be considered to be of continuing value (e.g., the outputs of scholarly and scientific research, the Websites of political parties, etc.), there is much content that is of low-quality (or worse)."[17] This statement highlights the value librarians can add to the selection of Web-published materials through identifying reputable and authoritative materials of relevance to their end users for inclusion in archived Web collections.

Print and other analog materials usually have clear boundaries–the item begins and ends with the container (e.g., the cover of a book). Selection and acquisition of Web-published materials is complicated by the lack of clear boundaries. As Peter Lyman pointed out, "Information on the Web is not discrete; it is linked. Consequently, the boundaries of the object to be preserved are ambiguous."[18]

For academic librarians, the unit of selection and acquisition for Web-published materials (e.g., individual images, discrete Web pages, or entire Websites) depends heavily on the discipline and the purpose of a collection. For certain disciplines, such as anthropology and history, source material context is often critically important for research, and therefore, the Website is the appropriate unit of selection and acquisition. For other research fields, the original Web context of the source materials is not always critical, and end users are better served by the ability to interact directly with individual objects such as images or documents.

Selection of print materials is commonly assisted by tools such as approval plans, depository agreements, and vendor or distributor catalogs. In addition, librarians often assess the authority of print materials based upon the known reputation of the publisher or author as well as on common publishing practices, such as the peer-review process employed by professional journals. Selection of Web-published materials is complicated by the lack of comparable selection tools and practices. Similarly, techniques for assessing the authenticity of print materials do not readily transfer to Web-published materials. These factors add to the difficult and time-consuming nature of the selection process.

A clear understanding of Website structure and Web-crawler behavior increases the efficacy of site selection. The physical computers that store and serve Web pages, as well as other Web-published materials identified by URLs, are called *hosts*. Although a Website may be wholly contained on a single host, some Websites consist of materials that are stored on and served by two or more hosts. In this case, it is important during the selection process to identify each host so the entire Website is captured during the acquisition process.

Archive service providers are likely to have established capture configurations that include specifications for both required and optional capture parameters, such as the URL for each Website to be captured or the maximum size of a capture. Broadly understanding the capture requirements relevant to a specific collection will help in choosing the most appropriate service provider.

The capture of Websites for a collection will likely be based on a *seed list* of URLs, which includes one or more URLs from which a Web crawler begins capturing Web-published materials. Web crawlers extract additional candidate URLs for capture from the Web pages in the seed list. Candidate URLs may reside on the same host as a seed URL (i.e., the *local host*) or a secondary host to a seed URL (i.e., an *external host*). From a Web-crawler perspective, *depth* refers to the number of linked URLs away from a seed URL that a crawler should capture content. Candidate URLs are evaluated by a crawler based on predefined settings such as whether to capture materials from external hosts and the desired depth of a capture. The crawler adds URLs that meet the predefined settings to its list of captured URLs.

A Web-collection plan should list the Websites to be included in the collection and describe each site. Keep in mind that Website structure will directly affect the URLs that should be included in a seed list. If a Website is served by more than one host the URL for each host should be specified in the seed list.

A Web-collection plan should also identify both when and how often each URL on the seed list should be captured. Possible capture frequencies might include one time only, every "x" number of days, quarterly on a specific date, whenever content changes, or upon request from the content provider. Evaluation of the frequency at which targeted materials experience significant change will help determine the appropriate capture frequency for a URL.

Specific types or formats of Web-published materials that should not be captured during crawls of seed URLs should also be identified in the Web-collection plan. Material types include generic categories such as text, images, audio, and video, as well as application-specific data types. Formats refer to specific encoding schemes such as html, jpeg, gif, PDF, etc. A Web-published file's type and format are identified by MIME types (e.g., text/html and image/gif). There may be policy or technical reasons that mandate that certain material types and formats will not be included in a collection.

Copyright considerations pose another challenge to the collection of Web-published materials. Lyman pointed out: "Although the Web is popularly regarded as a public domain resource, it is copyrighted; thus, archivists have no legal right to copy the Web."[19] For each seed URL, determine the rights that will govern the capture of its content. Also, as appropriate, determine the rights of sourced or embedded objects contained in the Websites. The Web-collection plan should include rights metadata for each seed URL. At a minimum, this might include contact information for the publisher, creator, or owner, contact history, and the date copyright permission was granted.

Because Websites with interactive and dynamic content are often not effectively captured by Web crawlers, one must evaluate the Websites in the seed list and identify and describe their reliance on this type of content. Consider the following: Is a site password protected? Are e-mail links and comment forms included? Does the Website rely on databases to generate Web pages? Does the Website create pages on-the-fly, possibly combining style sheets with server-side scripts or code?

The Web-archive service provider may allow preliminary or "test" crawls of Websites in the seed list. Further, the service provider might supply tools that can assist with an evaluation of Website interactivity based on the materials captured during test crawls. It may be possible to extrapolate from these evaluations to characterize the content of entire Websites.

Typically, a Web archive acquires Web-published materials by capturing content from Websites using a Web crawler. One major exception to this might be databases, which are usually neither accessible by nor friendly to a Web crawler. In such cases, it might be preferable for a content provider to make arrangements to submit the database to the Web-archive provider via an alternate method.

Section 3: Description and Organization

Cataloging and classification of print materials have well-established procedures, guidelines, and tools. Web-published materials have few established guidelines for descriptive cataloging, and the ambiguous boundaries that make them difficult to select and acquire also make them difficult to describe and organize.

Librarians view cataloging and metadata creation as one of the greatest challenges to creating and maintaining a collection of Web-published materials. Because metadata is strongly related to end-user information discovery, understanding the needs and salient characteristics of a collection's end users is critical. However, as with selection and acquisition activities for Web-published materials, descriptive cataloging is quite labor-intensive. Most librarians who participated in the Web-at-Risk needs assessment thought automated metadata generation is needed for captured materials, including subject or topic classification.

Because descriptive metadata updates and changes are costly, Alexa McCray and Marie Gallagher believe it is important "to decide on the nature and number of metadata elements early in a project."[20] Further, they stated decisions "on the basic conceptual units, or objects, the system will include" are essential in determining the level at which metadata will be assigned.[21] Web-collection plans should identify the basic units of description (e.g., Website, seed URL, Web page) and incorporate decisions regarding metadata schema and encoding method, content and input rules, and instruction regarding which extensions and qualifiers are allowed. Controlled vocabularies specific to a collection and meaningful to a collection's intended user group(s) may exist or can be developed.

Section 4: Presentation and Access Requirements

As with the contextual issues involved in Website selection, the importance of replicating the original look and feel of captured Web-published materials when they are viewed from an archive depends on the targeted user groups. Historians may consider the original Website context of the captured materials essential, while other researchers may consider the original context of materials such as archived datasets or documents superfluous.

Intellectual property and copyright considerations also affect the presentation of archived materials. To some extent, copyright protection for print materials is promoted by the manner by which they are accessed. Print materials must be obtained physically and can generally only be used by one person at a time; photocopying the material for use can be expensive and cumbersome. Web-published materials, on the other hand, are generally are easy to access, accessible by multiple users simultaneously, and extremely easy to copy. Thus, Web-published materials do not have the same inherent safeguards as print materials.

It is important to anticipate how user groups will want to interact with a Web collection for discovery and evaluation of the collection's materials. A Web-collection plan should identify the search methods end users will require. Consideration should be given to such discovery tools as simple keyword searches, advanced search screens, and browse capability. Collection plans should also identify the information end users will require in both basic and detailed search results. Finally, the retrieval method for end users to obtain desired materials once they have been identified should be considered. Will users be able to directly follow hyperlinks in search results to retrieve materials they want? Will Web materials be available for export from the archive?

In certain cases, librarians may designate Web collections as either "visible" or "dark"–that is, as accessible or not accessible to users. A variation on a dark archive might be a designation that a collection will become visible only at some future point in time. This might be done to protect personal privacy or to preserve a competitive market position. For example, public access to archived collections might be delayed until public access no longer has the potential to cause economic disadvantage or damage to the content producer.

Additionally, an archive might restrict access to its stored content based on agreements with content producers. Alternatively, an archive might employ a model of the fair-use doctrine and require archive users to formally agree to restrict use of the information to designated applications such as personal use.

In practice, most archived Web collections composed of captured Websites are presented as mirror experiences of the originally published sites. If Web-published materials are selected and captured at a more granular level than the Website–for example, a collection of videos of volcanic activity selected from a range of Websites–then a special user interface may be necessary to present them to end users. The Web-collection plan should identify and include requirements for special user interfaces such as these.

Librarians should evaluate the importance of retaining the look and feel of Websites in a Web collection and state the importance of this for the collection's user groups. Web-collection plans should also identify the appropriate handling of information content removed from archived Web pages for policy or legal reasons. Should users be alerted to this alteration? If yes, how should users be alerted?

When archived Web pages retain the look and feel of the original sites, librarians should address functionality issues associated with dynamic content. It is important to identify how links to non-archived materials will be handled. Will users be allowed to access hyperlinked materials and Websites that are not located within the Web archive? If so, will users be alerted that they are leaving the archive? If not, will links simply be disabled or will information about links such as the specified URL be presented along with an informative message? What about preservation of e-mail links? How will forms be addressed within the Web archive? For example, will the "Submit" button be disabled, or will an annotated static screen shot of the original form be available?

Websites often publish multiple types and formats of the same content. For example, a Website might publish video, audio, and transcript files of a single event, such as a public address by a government official. Each of these material types might be published in different formats–for example, the text transcript might be available in both DOC and PDF formats. When multiple types and formats of information objects contained in Websites are captured, will all the types and formats be discoverable and made accessible to users? A Web-collection plan should identify the types and formats of information objects users are allowed to access. This might vary according to a user's access location; a user at the institution's library, for example, might be allowed more extensive access than a user at home.

Kenneth Thibodeau cautioned:

> Given that a digital information object is not something that is preserved as an inscription on a physical medium, but something that can only be constructed–or reconstructed–by using software to process stored inscriptions, it is necessary to have an explicit model or standard that is independent of the stored object and that

> provides a criterion, or at least a benchmark, for assessing the authenticity of the reconstructed object.[22]

Consideration should be given to the needs of the collection's user group(s) with respect to the authenticity of archived Web-published materials. For some user groups, the reputation of the library or archive provider might be sufficient. Other user groups might require a certification process such as that proposed by the Research Libraries Group.[23] Is it necessary for authentication of materials to result in some type of indicator that the materials in a Web collection are reliable copies of source materials? Should such an indicator be visible to users when they view a Website in an archived Web collection?[24]

Section 5: Maintenance and Weeding

Maintenance of print collections is straightforward, involving shelving and repair. Deselection of print materials is necessary and accepted because there is a finite amount of shelf space. Those print materials, recognized as rare or of enduring value, are ultimately archived to be kept indefinitely, and their curation is well understood. Although archived items are subject to deselection on occasion, this is the exception rather than the rule.

In many Web archives, deselection or weeding will never occur. In fact, weeding appears to belie the essential preservation role of an archive. Yet there may be circumstances in which weeding is desirable. These circumstances might be dictated by retention guidelines and/or mandated by economic constraints

A Web-collection plan should identify the anticipated maintenance activities for the collection. Areas to consider for maintenance include seed lists, capture specifications for seed lists, rights metadata, descriptive metadata, and criteria for collection membership. The Web-collection plan should also identify triggers for conducting the identified maintenance activities. An example trigger might be an annual review of the URLs in the seed list for continued conformance to the criteria for collection membership.

Circumstances in which Websites or information objects might be removed from an archive should be identified in a Web-collection plan. These circumstances might include removal of materials in response to requests from content providers or in accordance with a user group's judgment of a site's continuing value to the Web collection. The plan should describe what it means to deselect a Website or a Web collection from an archive: Does it mean the Website will never be captured again? Does it mean preservation activity will be discontinued? Does it mean that the content will be removed from the archive?

System-generated data that might assist with the evaluation of the Web collection and with weeding and other maintenance decisions should be identified in the Web-collection plan. Usage data is one example of system-generated data that can be useful in collection evaluation. While of dubious value when employed as the sole weeding criteria for a collection, material usage can point to important trends in users' information needs, which can help inform collection development. Likewise, unexpectedly low usage of some materials might indicate potential problems in regard to the metadata creation and indexing processes. The plan should also describe methods of obtaining feedback with regard to the usefulness of the Web collection from its identified user group(s).

Section 6: Preservation

"With digital information, whose life span can be as short as one software upgrade, the decision to preserve must be made almost simultaneously with its creation," Nicholls and Smith said. "This turns the traditional preservation paradigm on its head."[25] Unlike print materials, for which consideration of preservation can be postponed for many years, the preservation of Web-published materials captured and stored in a Web archive must be addressed as soon as they are captured. As a result, the tasks of maintenance and preservation for Web collections effectively merge.

As with all digital materials, the materials in Web collections must be refreshed as their storage medium becomes outdated. They also must be migrated to different formats on occasion as the software, designed to interpret and present the information contained in the digital files, itself becomes obsolete. Preservation activities also include the creation of preservation metadata, which helps ensure the fundamental integrity of materials in the archive over time by establishing integrity indicators and provenance for materials in the archive.

According to the Digital Preservation Coalition, "Technology obsolescence is generally regarded as the greatest technical threat to ensuring continued access to digital material."[26] Curators of Web collections must plan for indefinite access to the content of the captured materials by identifying ways of mitigating the effects of technological obsolescence.

Librarians and curators of Web collections have a role in deciding how access to the content of captured material will be preserved. Will the original look and feel of captured materials be preserved by having newer hardware and software emulate the original platform? Will access to captured content be preserved by migrating captured materials to newer file formats? Will original software be preserved along with the captured content to ensure future accessibility? In responding to these questions, librarians and curators represent the needs and concerns of user groups in the decision process. Curators must also be aware of the implications, with regard to authenticity and copyright, when originally captured materials are migrated to different formats because of the threat of technological obsolescence.

Specify in the plan any policies or practices that must be considered when dealing with hardware and software obsolescence. A Web-collection plan should estimate the importance of maintaining the original look and feel of collected materials, identify a process for determining acceptable methods of providing continued access to the materials, and identify a process for evaluating the impact of those methods on the authenticity of materials and their copyright protection.

The Open Archival Information System reference model recommends the creation of four categories of preservation metadata listed in Table 2.[27] Example metadata elements for each category illustrate the types of metadata expected to be necessary for preservation of materials in a Web archive.

A Web-collection plan should identify any preservation metadata elements necessary to preserve the collection and specify who has responsibility for creating and maintaining each element. Librarians may find it useful to examine the preservation metadata element sets proposed by OCLC's Preservation Metadata: Implementation Strategies (PREMIS) working group[28] and by the National Library of Australia.[29]

TABLE 2. OAIS Recommended Preservation Metadata Categories

Category	Example
Reference	• One unambiguous identifier • Other identifiers (e.g., URLs)
Context	• Why content was created • How it relates to other content
Provenance	• Origin and history of content • Who has owned/controlled it • What changes/migrations have been done on it
Fixity	• Information regarding verification/validation of data integrity of the content • Integrity indicator

Supporting Appendices

Appendices can include a range of materials that augment the Web-collection plan. These might include agreements with content providers and Web-archive providers as well as applicable institutional policies, practices, standards, and guidelines that affect the collection. Alternately, appendices might provide a reference list to these agreements and policy documents. What is included relates to the Web collection being built, the archive service provider, the source of the content, and the institution or organization planning the collection.

A content-provider agreement or submission agreement specifies in some detail the legal relationship between a content provider or information producer and a Web-archive service provider. Submission agreements need to identify what Web-published content or data will be submitted and what metadata will accompany the content and data.

The agreement should also specify any procedures or protocols for Website capture by the Web-archive service provider and alternately for data submission by the content provider. Additionally, procedures for verifying successful transmission and procedures for getting answers to questions about the content should be specified in the agreement.

A Web-archiving service agreement should be contracted between the Web-archive service provider and the institution or organization that builds the Web collection. Such an agreement would identify the parties to the agreement and describe their respective roles and responsibilities in regard to Web archiving. Additionally, the service terms and conditions should be described, including penalties for non-performance, notices of service or contract termination, verification of integrity of captured materials, and error handling procedures.[30]

If more than one institution is collaborating to build a Web collection, one or more of the institutions may require some type of collaboration agreement. The specific terms and conditions may be dictated by the institutions as well as predicated by the type and scope of the agreement.

DISCUSSION

Website Architecture

In order to build an effective Web collection, an understanding of both the structure of the Websites to be captured and the way in which the capture software works is critical. As previously noted, the boundaries of a single Website are ambiguous. Lyman reported: "The average Web page contains 15 links to other pages or objects and five sourced objects, such as sounds or images."[31] Further complicating the picture is that although a single Website may be wholly contained on a single host, some Websites consist of materials that are served by two or more hosts.

Librarians must develop an understanding of the structure of selected Websites. They must be able to estimate with some degree of accuracy how many links a Web crawler should follow when capturing Websites on a seed list and whether to include materials from external hosts. Because librarians engaged in building Web collections often retain their traditional collection responsibilities as well, it is unrealistic in many cases to suppose librarians will be able to gain this knowledge and expertise without automated tools specifically designed to assist them with these activities.

Intellectual Property Considerations

Preservation of born-digital materials is complicated by copyright law intended for print materials. To date, efforts to extend intellectual property protections to the digital world, such as the Digital Millennium Copyright Act, have further constrained digital preservation efforts by restricting digital copying. The complicated nature of intellectual property in conjunction with digital materials is evidenced by the Library of Congress' Section 108 Study Group, which is "charged with updating for the digital world the Copyright Act's balance between the rights of creators and copyright owners and the needs of libraries and archives."[32] Until these issues are worked out, it is advisable for librarians building Web collections to assess the copyright status of all Websites they intend to include in their collections and to document both their findings and any actions taken to obtain copyright clearances.

Organizational Support

Creation and preservation of collections in Web archives requires enormous effort and resources, spanning several departments within an organization. As a consequence, successful Web-archival programs within an organization will require managerial commitment and sustained funding.

Information-management professionals, whether librarians, curators, or archivists, have expertise in collecting and preserving materials, but they often do not have the technical expertise necessary to create and preserve an extensive collection of Web-published materials. While information technology professionals have expertise working with networks and digital storage, they rarely understand the long-term implications inherent in the curation and preservation of stored content. It is clear these organizational units will need to work together to achieve success in collecting and preserving Web-published materials. Management support for inter-departmental cooperation is critical to success.

Additionally, long-term funding commitments are required for any Web archiving effort to succeed. To secure these commitments, librarians will need to articulate how collection and preservation of Web-published materials supports the organization's mission, benefits the organization, and provides a valuable service to the community it serves.

Value-Added Services

Although there are clearly many challenges in collecting and preserving Web-published materials, many of the participants in the Web-at-Risk's needs assessment study had positive thoughts about the value-added services these collections could provide. The obvious inherent benefit of Web archives is the provision of persistent access to a wide range of digital or Web-born scholarly materials. However, participants also saw other potential services Web archives could provide. Like Pitschmann, librarians postulated that Web collections could bring together related materials from disparate sources. They also foresaw the ability to create different indexes and impose different classification schemes on a single collection and, thus, enhance the collection's usefulness. Some thought a Web archive could be used as a tool for version tracking, especially for materials with relevance to particular points in time. Still others foresaw that Web archives could provide a valuable role by authenticating the materials in their collections.

CONCLUSION

The quantity and nature of Web-published materials necessitates the creation and preservation of Web collections by many libraries. Although the activities involved in building Web collections are conceptually the same as the activities for building traditional collections, they differ immensely in practice. The unit of selection and acquisition is difficult to define, quality Web-published materials can be difficult to discover, and verifying authenticity is a challenge. Application of metadata is time-consuming and often requires specialized expertise. Preservation requires technical skills and infrastructure beyond the capabilities of most libraries.

Successful Web-collection development will require significant knowledge of how Websites are organized in order to accurately capture the desired materials. Intellectual property considerations must be addressed and the relationship between copyright law and born-digital information closely monitored. Finally, successful preservation of Web collections will require organizational commitment and inter-departmental management endorsement. There must be a commitment both to long-term financial support of the Web archive and to ongoing collaboration among librarians and information technology professionals. In return for this commitment, organizations will enable long-term access to important Web-published materials for the communities they serve.

NOTES

1. The National Library of Australia's *Preserving Access to Digital Information* Website has an introduction to some of the major archiving initiatives around the world. This introduction can be viewed at: http://www.nla.gov.au/padi/topics/92.html (accessed November 30, 2006).

2. The Web-at-Risk project (http://www.digitalpreservation.gov/partners/project_cdl.pdf) is one of eight collection development partnership projects funded in 2004 by the Library of Congress under the National Digital Information Infrastructure and Preservation Program (http://www.digitalpreservation.gov). The project is prototyping a Web archiving service that will enable curators to build, store, and manage collections of Web-published materials in a Web archive.

In 2005, the project's 22 curators, who will build collections of Web-published materials using the service; 43 librarians and archivists, who primarily work in academic libraries; seven university researchers; and seven content providers participated in needs-assessment activities that included an online survey, focus groups, and interviews. The purpose of the needs assessment was to identify the needs and issues librarians, curators, end users, and content providers have specifically regarding collection development for Web-published materials and generally regarding Web archives. Additional information about the needs assessment study and its outcomes can be viewed at: http://web2.unt.edu/webatrisk.

3. Library of Congress, "Partnerships–Digital Preservation (Library of Congress), http://www.digitalpreservation.gov/partners/project.html.

4. Louis Pitschmann, *Building Sustainable Collections of Free Third-Party Web Resources,* (Washington, DC: Council on Library and Information Resources, 2001), http://www.clir.org/pubs/reports/pub98/pub98.pdf (accessed November 30, 2006).

5. Michael Day, *Collecting and Preserving the World Wide Web: A Feasibility Study Undertaken for the JISC and Wellcome Trust,* (Bath, UK: UKOLN, 2003), http://

www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf (accessed November 30, 2006).

6. Martin Waller and Robert Sharpe, *Mind the Gap: Assessing Digital Preservation Needs in the UK*, (York, UK: Digital Preservation Coalition, 2006), http://www.dpconline.org/docs/reports/uknamindthegap.pdf (accessed November 30, 2006).

7. Stephen G. Nichols and Abby Smith, *The Evidence in Hand: Report of the Task Force on the Artifact in Library Collections*, (Washington, DC: Council on Library and Information Resources, 2001), http://www.clir.org/pubs/reports/pub103/pub103.pdf (accessed November 30, 2006).

8. Pitschmann.

9. Peter Lyman, "Archiving the World Wide Web," in *Preserving Our Digital Heritage: Plan for the National Digital Information Infrastructure and Preservation Program*, (Washington, DC: Library of Congress, 2002), http://www.digitalpreservation.gov/about/ndiipp_appendix.pdf (accessed November 30, 2006).

10. The following quotes reflect authors' concerns about Web-published materials disappearing:

"The dynamic nature of the Web means that pages and whole sites are continually evolving, meaning that pages are frequently changed or deleted" (Day).

"As ubiquitous as the Web seems to be, it is also ephemeral, and much of today's Web will have disappeared by tomorrow. The implication is clear: if we do not act to preserve today's Web, it will disappear" (Lyman).

"Anyone who has tried to trace a citation to a digital source, only to find that the site no longer exists, understands that dedicated maintenance and resources are required to keep digital sources alive, let alone up-to-date" (Nichols and Smith).

"Web content is not durable; there are no archival guarantees. Information available at any given moment can move or cease to exist without warning" (Pitschmann).

"Failure to take steps to address the issue of digital preservation could lead to irrevocable loss of much of this material" (Digital Preservation Coalition).

11. Digital Preservation Coalition.

12. NISO Framework Advisory Group, *A Framework of Guidance for Building Good Digital Collections,* 2nd ed. (Bethesda, MD: National Information Standards Organization, 2004), http://www.niso.org/framework/framework2.pdf (accessed November 30, 2006).

13. Ibid.

14. Ibid.

15. Waller and Sharpe.

16. The views and opinions articulated in this report reflect the views and opinions gathered from librarians, curators, researchers and content providers that participated in the Web-at-Risk project's needs-assessment activities.

17. Day.

18. Lyman.

19. Ibid.

20. Alexa T. McCray and Marie E. Gallagher, "Principles for Digital Library Development," *Communications of the ACM* 44, no. 5 (2001): 48-54, http://www.proquest.com/ (accessed October 26, 2006).

21. Ibid.

22. Kenneth Thibodeau, "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years," in *The State of Digital Preservation: An International Perspective: Conference Proceedings*, (Washington, DC: Council on

Library and Information Resources, July 2002), http://www.clir.org/pubs/reports/pub107/pub107.pdf (accessed November 30, 2006).

23. RLG, "An Audit Checklist for the Certification of Trusted Digital Repositories: Draft for Public Comment," http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf (accessed November 30, 2006).

24. This indicator of authenticity is different from the integrity indicator identified in the preservation section of this document. Integrity is a measure of the bits included in captured materials stored in an archive. A baseline indication of integrity is established when materials are captured, often by a checksum method.

25. Nichols and Smith.

26. Neil Beagrie and Maggie Jones, "Digital Preservation," in *Preservation Management of Digital Materials: A Handbook,* (York, UK: Digital Preservation Coalition, 2002), http://www.dpconline.org/graphics/digpres/stratoverview.html (accessed November 30, 2006).

27. Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System (OAIS),* CCSDS Publication No. 650.0-B-1 (Washington, DC: CCSDS Secretariat, 2002), http://public.ccsds.org/publications/archive/650x0b1.pdf (accessed November 30, 2006).

28. PREMIS Working Group, *Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group,* (Dublin, Ohio: OCLC, May 2005), http://www.oclc.org/research/projects/pmwg/premis-final.pdf (accessed November 30, 2006).

29. National Library of Australia, "Preservation Metadata for Digital Collections," http://www.nla.gov.au/preserve/pmeta.html (accessed November 30, 2006).

30. If the Web archive service is provided by the librarian's own institution or organization, a service agreement may not be required. However, it is still important to identify organizational roles and responsibilities in the preservation effort and to ensure that supporting policies are in place within the organization.

31. Lyman.

32. More information about the Section 108 Study Group is available at http://www.loc.gov/section108/index.html (accessed November 30, 2006).

WORKS CITED

Beagrie, Neil, and Maggie Jones. "Digital Preservation," in Chap. 2 in *Preservation Management of Digital Materials: A Handbook.* York, UK: Digital Preservation Coalition, 2002. http://www.dpconline.org/graphics/digpres/stratoverview.html (accessed November 30, 2006).

Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System (OAIS),* CCSDS Publication No. 650.0-B-1. Washington, DC: CCSDS Secretariat, 2002. http://public.ccsds.org/publications/archive/650x0b1.pdf (accessed November 30, 2006).

Day, Michael. *Collecting and Preserving the World Wide Web: A Feasibility Study Undertaken for the JISC and Wellcome Trust.* Bath, UK: UKOLN, 2003. http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf (accessed November 30, 2006).

Lyman, Peter. "Archiving the World Wide Web," in *Preserving Our Digital Heritage: Plan for the National Digital Information Infrastructure and Preservation Program*. Washington, DC: Library of Congress, 2002. http://www.digitalpreservation.gov/about/ndiipp_appendix.pdf (accessed November 30, 2006).

McCray, Alexa T., and Marie E. Gallagher. "Principles for Digital Library Development." *Communications of the ACM* 44, no. 5 (2001): 48-54. http://www.proquest.com/ (accessed October 26, 2006).

National Library of Australia. "Preservation Metadata for Digital Collections." http://www.nla.gov.au/preserve/pmeta.html (accessed November 30, 2006).

Nichols, Stephen G., and Abby Smith. *The Evidence in Hand: Report of the Task Force on the Artifact in Library Collections*. Washington, DC: Council on Library and Information Resources, 2001. http://www.clir.org/pubs/reports/pub103/pub103.pdf (accessed November 30, 2006).

NISO Framework Advisory Group. *A Framework of Guidance for Building Good Digital Collections,* 2nd ed. Bethesda, MD: National Information Standards Organization, 2004. http://www.niso.org/framework/framework2.pdf (accessed November 30, 2006).

Pitschmann, Louis. *Building Sustainable Collections of Free Third-Party Web Resources*. Washington, DC: Council on Library and Information Resources, 2001. http://www.clir.org/pubs/reports/pub98/pub98.pdf (accessed November 30, 2006).

PREMIS Working Group. *Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group.* Dublin, Ohio: OCLC, May 2005. http://www.oclc.org/research/projects/pmwg/premis-final.pdf (accessed November 30, 2006).

RLG. "An Audit Checklist for the Certification of Trusted Digital Repositories: Draft for Public Comment." http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf (accessed November 30, 2006).

Thibodeau, Kenneth. "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years," in *The State of Digital Preservation: An International Perspective: Conference Proceedings*. Washington, DC: Council on Library and Information Resources, July 2002), http://www.clir.org/pubs/reports/pub107/pub107.pdf (accessed November 30, 2006).

Waller, Martin, and Robert Sharpe. *Mind the Gap: Assessing Digital Preservation Needs in the UK*, (Digital Preservation Coalition, 2006), http://www.dpconline.org/docs/reports/uknamindthegap.pdf (accessed November 30, 2006).

Received: 12/06/06
Revised: 12/16/06
Accepted: 12/20/06

doi:10.1300/J502v01n02_02