The Web-at-Risk at Three: Overview of an NDIIPP Web Archiving Initiative

Tracy Seneca

Abstract

The Web-at-Risk project is a multi-year National Digital Information Infrastructure and Preservation Program (NDIIPP) funded effort to enable librarians and archivists to capture, curate, and preserve political and government information on the Web, and to make the resulting Web archives available to researchers. The Web-at-Risk project is a collaborative effort between the California Digital Library, New York University Libraries, the Stanford School of Computer Science, and the University of North Texas Libraries. Web-at-Risk is a multifaceted project that involves software development, integration of open-source solutions, and extensive needs assessment and collection planning work with the project's curatorial partners. A major outcome of this project is the Web Archiving Service (WAS), a Web archiving curatorial tool developed at the California Digital Library. This paper will examine both the Web-at-Risk project overall, how Web archiving fits into existing collection development practices, and the Web Archiving Service workflow, features, and technical approach. Issues addressed will include how the reliance on existing technologies both benefited and hindered the project, and how curator feedback shaped WAS design. Finally, the challenges faced and future directions for the project will be examined.

Web-at-Risk Grant Background

The fate of born-digital documents delivered via the Web is precarious. During the last several years, a number of studies have examined the impact of Web references in the scholarly literature of a number of academic fields, and tracked the disappearance of cited references. These

LIBRARY TRENDS, Vol. 57, No. 3, Winter 2009 ("The Library of Congress National Digital Information Infrastructure and Preservation Program," edited by Patricia Cruse and Beth Sandore), pp. 427–441

studies have begun to quantify an experience that Internet users are familiar with: revisiting a cited document to find that it has either changed significantly or is no longer available. Many of these studies measure the fragility of Web citations in terms of their "half life," the point at which half of a sample set is no longer available. In the field of Law, Rumsey (2002) found a half-life of four years for citations in law review journals. Goh and Peng (2007) found a half-life of five years for Web citations in information science journals. These researchers are uncovering the impact of the ephemeral nature of the Web on scholarly publishing, and the threat to future researchers' abilities to retrace an author's cited evidence. However, another realm of information is impacted by the fragility of the Web. As U.S. government information at all levels is increasingly provided via the Web, both the present-day citizenry and future researchers are at risk of losing much of their national heritage. In addition to the volatility inherent in Web publications, the information on agency and political office sites is subject to sweeping change during changes of administration and policy. Government information specialists also face the increasing difficulty of identifying new publications as agencies publish reports directly to the Web, resulting in "fugitive documents" not delivered through traditional government publication streams. In addition to tools for capturing and preserving known documents, librarians and archivists need tools for discovering new publications. The scope of this problem was conveyed in a 2003 Mellon-funded study conducted by the California Digital Library:

> Government information plays a fundamental role in our society—it is a basic foundation of democracy. The data are as diverse as the agencies that create it, ranging from the Department of Health and Human Services and the National Oceanic and Atmospheric Administration, to the California State Coastal Conservancy and the California Trade and Commerce Agency. It is inherently multi-disciplinary in its appeal. And it serves multiple audiences, including research institutions, business enterprises, and private citizens.
>
> Government information is also culturally significant. Memory organizations, government agencies, legal entities, and society as a whole rely on its existence for a variety of essential civic, economic, and political functions. These groups rely on all types of government publications, regardless of format. Digitally published materials are more volatile, uncontrolled, and at much greater risk of being lost than those that are published in printed formats. Unlike the printed publications of U.S. governments, digital ones do not flow through central printing offices, making their existence, number, provenance, and orientation impossible to record. The most volatile and at-risk government information is that which is made available exclusively via the World Wide Web, where 65 percent of all government publications that are distributed by the Government Printing Office, the largest producer of government information, are now placed without printed analog. (California Digital Library, 2003)

In 2005, the National Digital Information Infrastructure and Preservation Program (NDIIPP) awarded a grant to the Web-at-Risk project, a three year collaborative effort to create tools to capture, curate, and preserve government and political information on the Web. This effort is led by the California Digital Library (CDL) with grant partners at the University of North Texas (UNT) Libraries and New York University (NYU) Libraries and in collaboration with the Stanford Computer Science Department and the San Diego Supercomputer Center. The curatorial work is carried out by over twenty-five government information specialists at the University of California (UC) campuses, NYU, UNT, and the Stanford University Libraries. In the course of the grant, these curatorial partners are building a range of Web archives on various topics. Because many of these curators are working at University of California campuses, there is a strong emphasis on California state and local documents, but the overall range of collections includes the Islamic and Middle Eastern Political Web, International Government Organizations and Developing Countries, and the Tamiment Library's archive of leftist political organizations. To date, over 1.4 terabytes of data have been captured by the project's curators and delivered to the Library of Congress as part of the NDIIPP collections.

Web-at-Risk in Context

The acquisition and preservation challenges posed by Web content are impacting libraries and cultural memory institutions on an international scale. At around the same time that NDIIPP began exploring solutions to digital preservation challenges in the United States, national libraries around the world faced similar challenges in capturing and preserving Web content produced by and about their countries. In 2003, ten national libraries, the Library of Congress, and the Internet Archive formed the International Internet Preservation Consortium (IIPC) to support the development of standards and open source toolsets for Web archiving. When the Web-at-Risk project got underway in 2005, the work of the IIPC and the Internet Archive had already begun to result in open source solutions and standards to serve as the foundation for this and other Web archiving toolsets.

The open-source crawler Heritrix was developed by the Internet Archive with support from the IIPC and version 1.0 was released in 2004 (Internet Archive, 2008). Heritrix provided two major innovations that made Web archiving efforts more effective and feasible. One was the use of the ARC format for Web archived data (Internet Archive, 1996). Rather than attempting to reproduce the complexity of a server's file system as content is captured, the content is simply appended to a large container file, the ARC file, with separate headers indicating the source of each segment of that data. This innovation insures that each captured file can be found, but also keeps the structure of the Web archive simple and sustainable.

The other innovation Heritrix offered was a modular design that allows for different processors to interpret different file types. The Web is a moving target both in its content and technology. As sites become more interactive their navigation and structure can become enmeshed in JavaScript, Flash, PDF, and other formats. Since Web crawlers work by following links from seed URLs provided by the curator, it is critical to be able to interpret and follow links found in a variety of formats. Heritrix is designed to allow for new processors to be integrated as new Web standards emerge, without having to revisit the fundamental design of the crawler. This extensible design has meant that Heritrix is both very effective and responsive to changes in Web technology. The IIPC also provided support to develop tools for indexing and rendering data from the ARC formatted content that Heritrix returns. While these tools can be run independently, they do not in themselves compose a complete curatorial interface for Web archiving. Rather, they serve as the building blocks for those curatorial interfaces. What librarians need from those curatorial interfaces can be dramatically different depending on the context in which they work. A national library charged with capturing a single national domain faces quite a different problem than a distributed group of institutions building smaller, topical archives or collaborating on shared archives. The task of the Web-at-Risk project has been to address the needs of that more distributed, collaborative environment.

The Web-at-Risk Assessment Findings

User-centered design has been a guiding principle of this project from the outset. In order to ensure that the resulting toolset would meet users' needs and would make this new realm of collection building as intuitive as possible, the Web-at-Risk project began with a strong emphasis on needs assessment. This work was led by Kathleen Murray, post doctoral research associate, University of North Texas Libraries, and included in-depth surveys of the curatorial partners, several focus groups with librarians from a range of institutions, interviews with potential end-users of Web archives, and interviews with content owners. This work is summarized in detail in the "Web-at-Risk Needs Assessment Summary Report" (Murray & Hsieh, 2006). The results of this assessment work underscored the severity of the problem and the frustration that librarians, archivists, and researchers experience as they watch this material disappear, but also brought a few surprises. One result which has shaped the project's requirements was the strong document-centric focus of the curatorial partners. The strength and purpose of Web crawlers, and particularly the Heritrix crawler, is their ability to capture websites as completely as possible. Yet only 44 percent of the project's curators indicated that they planned to capture websites in their entirety (Murray, 2007). From the curators' perspective, the risk to the website may be of less concern than the risk to the documents on it. In many cases, the documents they hope to capture are documents they have long collected, such as state budget reports, grand jury hearings, and environmental impact reports, and can be considered a continuation of a series.

When delivery of these documents moved to the Web, the website itself only served to get in the way of the publications. However, this same group of curators, as well as many focus group participants, referenced both the daunting scale of trying to collect Web content at the document level, and the fact that they often don't know when new publications are available. In a recent follow-up survey with curatorial partners, identifying newly available documents was the single most highly ranked feature request (Murray, 2007). It became clear that the project needed to employ large scale Web capture tools to enable librarians to discover new documents on a more finely grained scale.

The initial needs assessment work was used to guide the early requirements and design of the Web Archiving Service (WAS), which was developed in stages between 2006 and 2008. In order to approach this ambitious development project, the requirements were broken down into separate stages, beginning with simple website capture tools, and then expanding to cover additional features such as analysis, collection building, and administration. Pilot tests with curators followed each stage of development. This allowed developers to break the project down into feasible segments and to integrate user feedback throughout the project. Curators have since taken part in a series of WAS pilot tests as new aspects of the service have been developed, and further assessment work was conducted to measure the success of each phase. With each pilot test, the infrastructure of the service has been enhanced as well, allowing for greater capacity with each testing phase. This user-centered design approach has helped to make the Web Archiving Service interface intuitive and easy to learn. This has been a critical aspect of the project, as most current and anticipated users of the service are balancing this new realm of collection development with an already demanding workload. The pilot testing process also revealed a great deal about the variety of ways curators planned to work together. Many of our curatorial partners are located at different University of California campuses, and while they need to be able to construct independent Web archives, they also have a strong need to survey institutional activity at all campuses, and to be able to draw from materials collected by other campuses. While it was clear from the outset of the project that we were building a tool meant to serve the diverse needs of many different institutions, the need for collaboration and content-sharing became much more pronounced as collection building got underway.

Web Archive Collection Plans

Another important facet of the project was exploring how to balance the unique considerations of Web archiving with traditional collection development practices. One challenge of Web archiving is that it is difficult to know in advance exactly what a crawler will get, or how large a website is until after it has been captured. Even when the curator can control the crawler settings to influence how the capture should run, the design of a site can impact the results in unforeseen ways. By contrast, librarians are accustomed to having a good degree of control when building print collections. In the print realm, in addition to selecting titles on an individual basis, librarians have a number of filters to control what titles are recommended or sent, such as subject classification, publisher, call number, and price. These can all be examined and controlled in advance of actually receiving the content, and there is generally not a great discrepancy between what a selector orders and what is actually shipped.

To explore how it might be possible to apply collection development practices to Web archiving, each curator was asked to write a collection plan in advance of using the tools to capture content. Curators were asked to describe their planned collections as completely as possible, including how the anticipated researchers would interact with the finished archive, and what metadata fields would be needed to describe and discover the content. The collection plans and the guidelines for writing them are available on the project's wiki page (California Digital Library, 2007). One purpose of these plans was to convey the scope of planned collecting activity to project developers. While they couldn't convey the specific amount of storage needed, they could at least convey that several hundred sites would be captured during a given pilot test. Another purpose was to prompt curators to analyze the sites they intended to capture in advance and perhaps better understand how to approach capturing them. For example, one city agency site might be very hierarchical in nature, allowing the curator to capture particular offices separately and perhaps more frequently. Another agency site might be more vaguely distributed, with relevant information being provided from a number of different sources.

The resulting collection plans became a valuable source of assessment in themselves. For example, curators were asked to provide a sample metadata entry for an archived site, to illustrate how they expected researchers to be able to find content. The resulting range of metadata requirements was telling. While a couple of curators stated that simple Dublin Core elements would be sufficient, others ranged from four to over fifty metadata elements. It was clear that metadata needs varied not just from one organization to the next, but from collection to collection, and that a flexible approach to metadata was going to be necessary. Overall, the collection plans were a valuable part of the project; they helped to ensure that the content being collected was in support of each organization's research programs, they helped curators become more deeply familiar with the content, and they provided the CDL support team with a good sense of the workload the Web Archiving Service would have to handle.

The Web Archiving Service Workflow

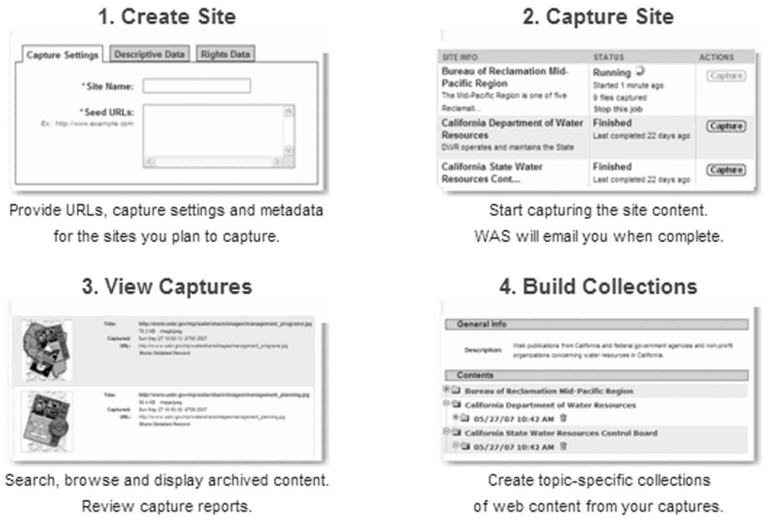
The Web Archiving Service workflow was designed to address the intrinsic steps of Web archiving while incorporating the lessons learned from the project's assessment work as much as possible.

When an institution establishes a WAS account, users can create different projects that serve as workspaces for building Web archives. Each project can have its own list of authorized users, allowing different groups of people to work collaboratively on different projects. Ultimately each project will have a configurable metadata template for describing sites, allowing for the metadata flexibility that was surfaced in the curators' collection plans.

The overall WAS workflow can be seen in the welcome page for a Web Archiving Service Project. From here the curator can define sites, review completed captures, and build collections of captured content (see Figure 1).

Creating Site Entries

The curator's first step is to create site entries that define the rules the crawler will follow when capturing that site. Each site entry requires a name, one or more seed URLs to use as the starting point for the crawler, and basic capture settings. The scope settings available are "site," "page," or "directory." For each of these scopes the user can decide whether the crawler should stay on the original site or if relevant immediately linked pages from other sites should be included. The user can also choose whether the capture should be brief (one hour) or full (up to thirty-six hours). The ideal selection of settings depends on the site being captured.



1. Create Site

Capture Settings | Descriptive Data | Rights Data

*Site Name:

*Seed URLs:
Ex.: http://www.example.com

Provide URLs, capture settings and metadata
for the sites you plan to capture.

2. Capture Site

SITE INFO	STATUS	ACTIONS
Bureau of Reclamation Mid-Pacific Region The Mid-Pacific Region is one of five Reclamati...	Running Started 1 minute ago 9 files captured Stop this job	Capture
California Department of Water Resources DWR operates and maintains the State	Finished Last completed 22 days ago	Capture
California State Water Resources Cont...	Finished Last completed 22 days ago	Capture

Start capturing the site content.
WAS will email you when complete.

3. View Captures

Search, browse and display archived content.
Review capture reports.

4. Build Collections

General Info

Description: Web publications from California and federal government agencies and non-profit organizations concerning water resources in California.

Contents

Bureau of Reclamation Mid-Pacific Region
California Department of Water Resources
05/27/07 10:42 AM
California State Water Resources Control Board
05/27/07 10:42 AM

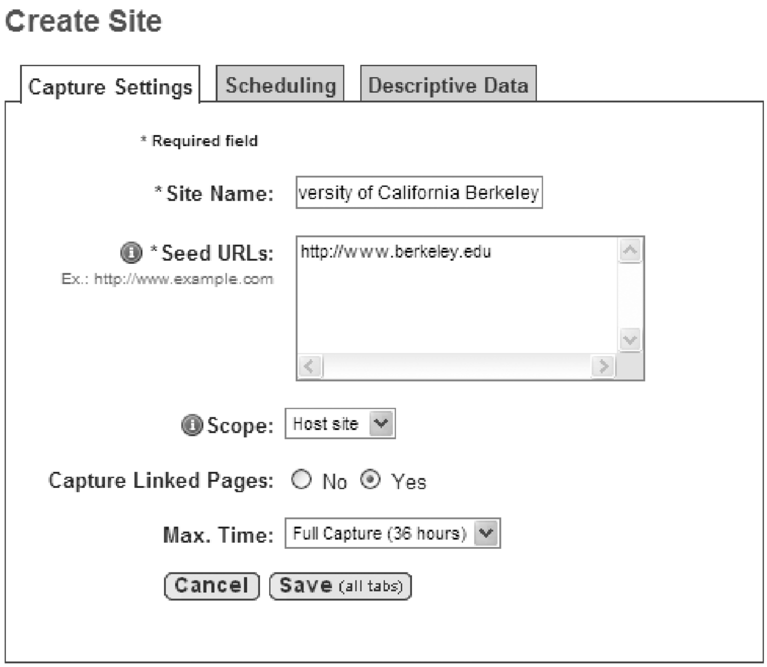
Create topic-specific collections
of web content from your captures.

Figure 1. Web Archiving Service Project Home Page

For example, during the 2005 Southern California Wildfires, a number of blogs provided links to valuable information elsewhere. In this case a "page + linked pages" setting was used to capture the relevant blog entries about the fire as well as the linked content without capturing older, less relevant content from the blog. The user can also set up capture frequency and provide basic metadata about the site (see Figure 2).

Running Captures and Analyzing Results

Once the site entries are created, the curator can capture the site. Captures are specific instances of the site captured at a particular date and time. A typical Web Archiving Service project can contain scores or hundreds of sites, with several captures of each site. WAS provides feedback about capture status and progress, such as the number of documents gathered both in the WAS interface and via an RSS feed. When the crawler has captured all of the documents it can find in the time allotted, the resulting content is then indexed and ingested into CDL's Digital Preservation Repository, and the curator is notified by e-mail that the capture is ready to review. Each capture provides overall reports, such as the frequency of different mimetypes and a list of hosts encountered, and is keyword searchable. These analysis tools may shed new light on the captured con-



Create Site

* Required field

* Site Name: versity of California Berkeley

* Seed URLs: http://www.berkeley.edu

Ex.: http://www.example.com

Scope: Host site

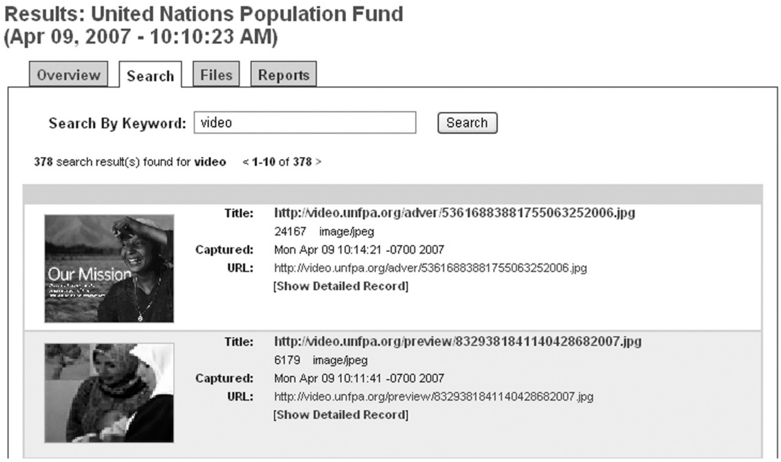
Capture Linked Pages: ○ No ● Yes

Max. Time: Full Capture (36 hours)

Figure 2. Web Archiving Service Site Definition Screen

tent. Even a curator who is very familiar with a site's content will have habitual ways of navigating it on the Web. Once the site is captured and is individually searchable, the full scope of its content comes into view in new ways. Documents that were linked from obscure points on the live site can be surfaced via keyword search, the first step toward addressing the document discovery issues that were a strong source of concern in the assessment findings (see Figure 3).

Content can be displayed and navigated, and each captured page has a detailed display that provides both document and site metadata. Curators can create comments on individual documents that will also show in the detailed display. When multiple captures have been run of the same site on different dates, curators can use a comparison tool to show which files are new since the previous capture, which files have changed, and which files are in the archive that are no longer on the live Web. If each capture was run with different settings, the comparison report will help the curator interpret the impact of those changes. If both captures were run with the same settings, then the comparison report will tell the curator how the site itself has changed between the two capture dates. The compare tool can be filtered by file type, so the user can find only the new PDF files published to the site since the last capture. This feature was developed to directly address the need for document discovery tools, and has been particularly well-received in follow-up assessment with curators. Curators can now routinely capture websites and analyze the results for new or previously undiscovered documents (see Figure 4).



Results: United Nations Population Fund
(Apr 09, 2007 - 10:10:23 AM)

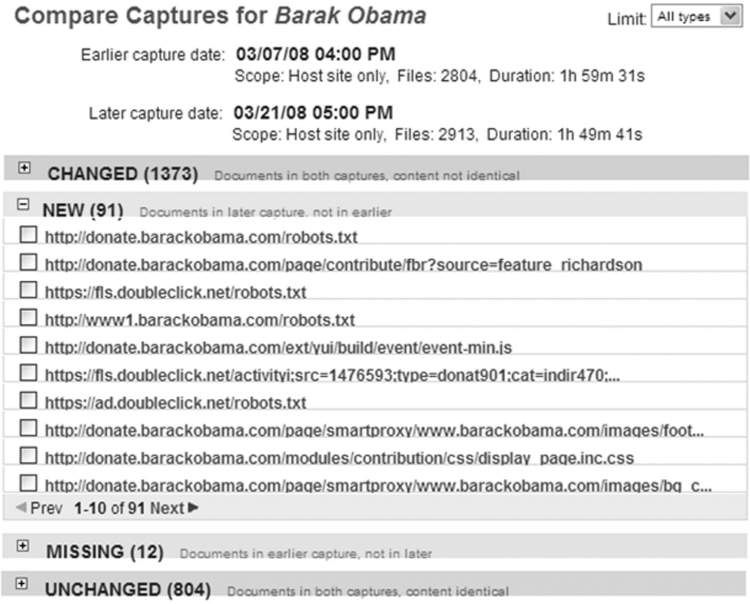
Search By Keyword: video

378 search result(s) found for video < 1-10 of 378 >

Title: http://video.unfpa.org/adver/5361688388175506325 2006.jpg
24167 image/jpeg
Captured: Mon Apr 09 10:14:21 -0700 2007
URL: http://video.unfpa.org/adver/53616883881755063252006.jpg
[Show Detailed Record]

Title: http://video.unfpa.org/preview/83293818411404286 82007.jpg
6179 image/jpeg
Captured: Mon Apr 09 10:11:41 -0700 2007
URL: http://video.unfpa.org/preview/83293818411404286 82007.jpg
[Show Detailed Record]

Figure 3. Web Archiving Service Search Results Screen



Compare Captures for *Barak Obama* Limit: All types

Earlier capture date: **03/07/08 04:00 PM**
Scope: Host site only, Files: 2804, Duration: 1h 59m 31s

Later capture date: **03/21/08 05:00 PM**
Scope: Host site only, Files: 2913, Duration: 1h 49m 41s

⊞ **CHANGED (1373)** Documents in both captures, content not identical

⊟ **NEW (91)** Documents in later capture. not in earlier

☐ http://donate.barackobama.com/robots.txt
☐ http://donate.barackobama.com/page/contribute/fbr?source=feature_richardson
☐ https://fls.doubleclick.net/robots.txt
☐ http://www1.barackobama.com/robots.txt
☐ http://donate.barackobama.com/ext/yui/build/event/event-min.js
☐ https://fls.doubleclick.net/activityi;src=1476593;type=donat901;cat=indir470;...
☐ https://ad.doubleclick.net/robots.txt
☐ http://donate.barackobama.com/page/smartproxy/www.barackobama.com/images/foot...
☐ http://donate.barackobama.com/modules/contribution/css/display_page.inc.css
☐ http://donate.barackobama.com/page/smartproxy/www.barackobama.com/images/bg_c...

◄ Prev **1-10 of 91** Next ►

⊞ **MISSING (12)** Documents in earlier capture, not in later

⊞ **UNCHANGED (804)** Documents in both captures, content identical

Figure 4. Web Archiving Service Capture Comparison Screen

Building Collections

The final step in WAS workflow is to create a collection of captured content. Here, a collection corresponds to what the curators defined in their collection plans, and can include as much or little of each capture as they choose. For example, when using the capture comparison tool, curators can select only the new PDF files to be added to a collection. Every capture or document added to the collection is indexed, so the collection is full-text searchable. When displaying a document from a collection, the user can see if other versions of that document were captured on other dates, and so navigate the archive in time. So while a project represents all of the content that has been captured, a collection represents only the content that has been selected by a curator. This workflow model was established to address the selective archiving needs of many of our curators. Together, the comparison features described above and the selective collection building model provide a balance between the large scale captures needed for document discovery and the control the curator needs to build the appropriate collection.

The Web Archiving Service and Repository Interoperability: Lessons Learned

The Web Archiving Service was built as an extension of the Digital Preservation Repository (DPR), a Java-based repository system built by CDL.

Early in the project this provided the enormous advantage of building on an already existing repository ingest, storage, tracking, and retrieval services. With some modification, these services were adapted for WAS and did not have to be developed separately. For example the Digital Preservation Repository requires a METS files for submitted content (Library of Congress, 2008); these files are constructed by the organization submitting the content. Since all the material coming into the Web Archiving Service is in roughly the same format (one or many ARC files and a series of crawl reports produced by Heritrix), the DPR's "feeder" service was modified to produce a simple METS file describing the digital object (the capture). The METS file does not attempt to describe the captured website and potentially the tens of thousands of files it contains, but rather describes the ARC files and reports that compose the capture. The other benefit of the DPR base is that the Web Archiving Service inherits features such as ongoing fixity checking and replication services, allowing the WAS developers to focus specifically on Web archiving issues.

There have been, however, some noteworthy drawbacks to this approach. The first challenge became apparent during the project's first pilot test. The DPR has a relatively simple user interface built in Struts, a Java framework for developing Web applications (Apache Software Foundation, 2008). The user interface for the Web Archiving Service is much more demanding, the project's goals and schedule are ambitious, and the design process has been iterative in response to user feedback and lessons learned. The project's developers needed an agile, dynamic user interface development environment and Struts did not meet that need. The second challenge is that the Web-at-Risk project partners at NYU and the University of North Texas do not use the DPR, and the DPR code is not publicly available. During the pilot phase of the grant, curators from all institutions are using the Web Archiving Service hosted on CDL infrastructure, but once the service is fully developed, both NYU and UNT intend to run it locally, and would prefer it to work with their existing repository systems, D-Space and Fedora. Ideally, the Web Archiving Service should be interoperable with a range of repositories, so that it can by used by many different institutions. Finally the workflow itself has raised some issues. Early in the project, designers believed it would be a simpler model to use existing ingest procedures to pull content into the DPR immediately after it was captured. This appeared to prevent the complexity of building a staging area for content, then prompting the user to preserve portions of it. However, this approach has cemented the dependency on DPR, and has also created a high initial processing cost for content that the curator has not yet seen. The display process is also slowed by the extraction of compressed files from the DPR (and the SRB storage system).

As a result of these issues, work has begun to make the Web Archiving Service independent of any particular repository system, so that it can be used in a wider range of contexts. The first step, taken in the fall of 2006, was to build a separate user interface for WAS, using Ruby on Rails as the development environment (Hetzner, 2007). This decision gave the developers the ability to respond quickly to user input, and helped the project keep pace with a demanding schedule. Developers are currently focused on revising the WAS workflow, so that captured material is held in a staging area for QA review by curators, who can then determine what will be preserved in the repository. The California Digital Library will continue to use the Digital Preservation Repository for materials curators choose to preserve, but ingest procedures for other repositories can be built for other organizations more readily with this modified workflow.

The Demands of a Production Web Archiving System

Another challenge offered by this project was balancing the curators' immediate needs to capture and preserve Web content (and the delivery of those collections to the Library of Congress) with the fact that curators were *pilot testing* a system not yet in production. A production system implies not just that the application is sound, but that it is available on a 24 by 7 basis, that technical support and training are readily available, and that the infrastructure can respond to fluctuations in workload and the inevitable nuisance of network outages and server maintenance. For the project's curatorial partners, this has meant that they have only been able to actively collect materials during testing phases. Particularly when the WAS user interface had reached a point where it was intuitive and fully featured, curators have been eager to use the service on an ongoing basis, while behind the scenes the infrastructure had not yet reached the scale needed to support that service.

For the project developers, the path to a production service has been eye-opening; particularly for an application that takes in data on a potentially very large scale. For example, the Web Archiving Service allows curators to schedule ongoing captures of a site on a daily, weekly, or monthly basis. Curators can also issue individual "one-off" captures of a site as needed. The service currently employs a bank of sixty crawlers, and captures can run for up to thirty-six hours. Additional crawlers can easily be deployed if it is found that the service is hitting capacity, and the project has already reached the point of having hundreds of captures scheduled to run. If all crawlers are busy, captures are queued and run in the order they were requested. It was assumed that most individual site captures would be run during the working week, when curators were there to hit the capture button, so weekly scheduled captures are set to begin at 5:00 p.m. on Friday, when there is less competition for crawlers. However, most data centers tend to run server maintenance in the early hours of Sunday morning. So what most data centers assume is a lull in usage is actually a peak time for the Web Archiving Service. Further, there is no way to predict when a lull time for this service will occur. Finally, if a data center issues an urgent request to shut down services for maintenance, even if no new captures are permitted, the ones already in progress can take up to thirty-six hours to complete. In order to provide dependable 24 by 7 production service, utilities must be created to ensure that the service can either shut down or fail over to other servers on relatively short notice without losing any data.

The infrastructure requirements for conducting Web archiving are also daunting. The need for storage for the archived data is clear, but there is also need for disk space while the capture is running; the captured content needs to reside somewhere until the capture finishes and the content is successfully stored in a repository. There is also a considerable need for processing during both indexing phases: when the initial capture is indexed and when the built collection is indexed, so that both are searchable.

End-user Access

The Web Archiving Service collections are not yet publicly available. The focus of the remaining year of NDIIPP-funded development will be to develop the features needed for public access to both built collections and to individual documents within a collection. This will require the same needs assessment work that was conducted to guide the curatorial tools, so the Web-at-Risk project will be seeking the input of scholars and researchers who will be inclined to use the materials the curators are archiving. This will also involve adding another step to the curatorial workflow: a "publish" action for collections ready for public view. This carries with it a number of required and desired features.

One is the ability to provide more curatorial control over the allowed metadata for each collection. When the Web-at-Risk curators conveyed such a wide-ranging variety of needs for metadata, they also implicitly conveyed a range of visions for what their "completed" archives would look like to end-users. Two of the project's curatorial partners already had significant experience in Web archiving, and had worked on locally built Web archives that are currently available to the public. Both of these, the CyberCemetary (University of North Texas Libraries, 2008) and the UCLA Online Campaign Literature Archive: A Century of Los Angeles Elections (UCLA Libraries, 2008) allow users to browse captured websites by topic. Consequently, as the project approaches public access, the ability to define metadata for each collection and to reflect that metadata back to the end-user by way of searching or browsing options becomes critical.

The need to respond to rights management issues also becomes an important prerequisite to public access. The copyright implications for Web archiving, even for largely public-domain government publications, are still not well understood or established. In March 2008, the Section 108

Study Group released its recommendations for the revision of copyright provisions for libraries, particularly for digital preservation and Web archiving practices. The group recommended that "a new exception should be added to section 108 to permit libraries and archives to capture and reproduce publicly available online content for preservation purposes, and to make those copies accessible to users for purposes of private study, scholarship or research" (Section 108 Study Group, 2008). While promising, this is still only a recommendation, and the Web-at-Risk and other Web archiving projects must proceed with their work in murky waters. What this means for the Web Archiving Service is that the developers will provide features for removing content from public view in the cases where it is critical to do so at the content owner's request, but will not construct a full-fledged rights management and tracking system. While these disputes can be expected, it should be noted that at no point in the pilot testing phase did a content owner contact CDL to request that their sites not be crawled. Further, one United Nations agency has requested a Web Archiving Service account, so as to be directly involved in the preservation of their own content when they know it will change significantly.

Finally, the ability to generate persistent URLs for direct access to archived documents, either from links or from catalog records is also a critical component of public access. For Web documents that are part of a series begun in print, this will tie together a library's holdings for that series whether print or digital, and will provide access to an archived copy not at risk of disappearing. However, as the public grows increasingly accustomed to finding government publications on the Web, over time they will become less inclined to search library catalogs to find them. While catalog records are important for some archived Web content, the Web archiving community will need to find ways to place that archived content where the user is most inclined to look for it.

In Summary

At three and a half years, the Web-at-Risk project has produced a user-friendly large-scale Web archiving application, guided by user input, with a workflow that allows for collaboration and flexible collection building. The Web Archiving Service is on the verge of going into production for the UC campuses and the project's grant partners. The emerging field of Web archiving has been an exciting, constantly evolving environment. Open-source tools became available during the course of the grant that completely altered the scope of work, and new challenges and requirements arose that altered it again. Moving forward, the possibilities offered by these archives are even more exciting. Researchers and writers may have new tools for archiving content as they cite it, so that their references are persistent, and point to a version of the document that looks exactly as it did on the day it was cited. Looking beyond the defensive position of preventing the loss of these documents, new prospects for research are offered once this data is housed in an archive. Tools for visualizing trends in Web content and for conducting text analysis against collections of Web content on a particular event are only part of what might become possible as librarians and archivists enter a new realm of collection development.

References

Apache Software Foundation. (2008). *Struts.* Retrieved July 8, 2008, from http://struts.apache.org/

California Digital Library. (2003). *Web-based government information: Evaluating solutions for capture, curation, and preservation, an Andrew W. Mellon funded initiative of the California Digital Library.* Retrieved June 28, 2008, from http://www.cdlib.org/programs/Web-based_archiving_mellon_Final.pdf

California Digital Library. (2007) *Web-at-Risk collection plans.* Retrieved August 16, 2008, from https://wiki.cdlib.org/WebAtRisk/tiki-index.php?page=WebCollectionPlans

Goh, D. H., & Peng, K. N. (2007). Link decay in leading information science journals. *Journal of the American Society of Information Science and Technology, 58*(1), 15–24.

Hetzner, Erik. (2007). *Using Ruby on Rails.* Presented at the Digital Library Federation Spring 2007 Forum. Retrieved August 16, 2008, from http://www.diglib.org/forums/spring2007/presentations/hetznerA.pdf

Internet Archive. (1996). *ARC file format.* Retrieved July 8, 2008, from http://www.archive.org/web/researcher/ArcFileFormat.php

Internet Archive. (2008). Heritrix home page. Retrieved June 29, 2008, from http://crawler.archive.org/

Murray, K., Hsieh, I. (2006). *Needs assessment survey report.* Retrieved June 29, 2008, from http://web3.unt.edu/webatrisk/na_toolkit/Reports/survey_data_analysis_final_05Jan2006.pdf

Murray, K. (2007). *Prioritization of WAS enhancement ideas.* Retrieved June 31, 2008, from http://wiki.cdlib.org/WebAtRisk/tiki-download_file.php?fileId=269

Library of Congress. (2008). *Metadata encoding and transmission standard.* Retrieved June 31, 2008, from http://www.loc.gov/standards/mets/

Rumsey, M. (2002). Runaway train: problems of permanence, accessibility, and stability in the use of Web sources in law review citations. *Law Library Journal,* 94, 27–39.

Section 108 Study Group. (2008). *Section 108 study group report: Executive summary.* Retrieved July 8, 2008, from http://www.section108.gov/docs/Sec108ExecSum.pdf

UCLA Libraries. (2008). *ULCA online campaign literature archive: A century of Los Angeles elections.* Retrieved July 8, 2008, http://digital.library.ucla.edu/campaign/

University of North Texas Libraries. (2008). *CyberCemetary.* Retrieved July 8, 2008, from http://govinfo.library.unt.edu/

Tracy Seneca is the Web archiving services manager at the California Digital Library. She has an extensive background in designing and developing Web applications for libraries, including a copyright tracking application for electronic reserves, tools for creating Web-based research instruction and tools for managing library subject guides. She came to application development for libraries by way of bibliographic instruction but also has experience in collection development and public service. She has presented frequently on issues raised in Web archiving at the Digital Library Federation Forum, the International Internet Preservation Consortium General Assembly, the American Library Association Annual Conference, and other events. She received a Master of Library and Information Studies from the University of California, Berkeley in 1995 and a Master of Arts in Applied Technology from DePaul University in 2004.