"Selecting and Managing Content Captured From the Web: Expanding Curatorial Expertise and Skills in Building Library of Congress Web Archives"

Abbie Grotke, Digital Media Project Coordinator, Office of Strategic Initiatives,
Library of Congress, Washington, D.C. 20540-1310, abgr@loc.gov
and
Janice E. Ruth, Manuscript Specialist in American Women's History, Manuscript
Division, Library of Congress, Washington, D.C. 20540-4680, jrut@loc.gov

DigCCurr2007: An International Symposium in Digital Curation
University of North Carolina, Chapel Hill, NC
April 20, 2007

Abstract: Since 2000, the Library of Congress, as part of its MINERVA program, has been capturing Web sites and developing thematic, event-driven Web archives on such topics as the terrorist attacks of September 11, 2001, national elections, and the Iraq War. A small number of Library staff worked on these projects and gained expertise in crawling technologies, tools, and workflow. In June 2005, Library managers sought to increase the number and expertise of staff involved and to extend the collecting scope beyond event-based Web captures. Staff from the Office of Strategic Initiatives (OSI) partnered with recommending officers, curators, specialists, and processing staff in various Library Services (LS) divisions to conduct a pilot project titled Selecting and Managing Content Captured from the Web (SMCCW). The initiative involved twenty-five staff members working on four distinct projects to develop selection criteria for Web content; examine how technical capabilities enable, affect, or prevent the building of Web site collections; and document the activities required to ensure the continued viability of the content. This paper describes the SMCCW initiative. Project manager Abbie Grotke summarizes the Library's pre-2005 Web archiving activities before presenting an overview of the goals, organization, training needs, and accomplishments of the SMCCW effort. As one of the four team leaders, Janice E. Ruth discusses the project's implementation in the Manuscript Division, describing the staff's appraisal and quality review of captured Web sites, how the work was integrated into existing division priorities and donor relationships, and the skills needed to complete the project.

Introduction

In 2000, the Library of Congress established a pilot project to collect and preserve primary source Web-based materials under the aegis of its MINERVA program. The Library's initial collecting efforts involved identifying sites relevant to particular events. A multidisciplinary team of Library staff representing cataloging, legal, reference, and technology services studied methods to evaluate, select, collect, catalog, provide access to, and preserve these materials for future generations. Since then, the Library has developed thematic Web archives on such topics as the national elections of 2000, 2002, and 2004; Iraq War; and events of September 11, 2001. Until 2004, a small group of Library staff gained practical experience in developing thematic archives while also acquiring a better understanding of crawling technologies, tools, and workflow.

In 2004, the Library's Office of Strategic Initiatives (OSI) created a Web Capture team to support the goal of managing and sustaining at-risk digital content. The team's charge is to build a Library-wide understanding and technical infrastructure for capturing Web content. Working with a variety of Library staff and national and international partners, the team is identifying policy issues, establishing best practices, and building tools to collect and preserve Web content.

Recently the focus has shifted to broadening the knowledge and understanding of Web capture activities among the Library's collection development staff, which includes curators and specialists. As the Library extends its Web capture efforts beyond the relatively narrow scope of event-based collecting, it hopes to tap existing subject expertise within the institution and to increase the number of staff with knowledge and skills to acquire and manage resources archived from the Web.

Selecting and Managing Content Captured from the Web

The Selecting and Managing Content Captured from the Web (SMCCW) project was initiated in June 2005 to establish policies and procedures for selecting Web sites for capture. The project sought to identify categories of Web content, establish selection criteria appropriate to each category, and enhance understanding of issues related to the depth, breadth, and frequency of capture. Another objective was to establish and document Library processes for the selection, storage, and maintenance of captured Web sites. While some processes were already in place, testing these on a wider group of curators would help to refine and standardize how Web collections are developed at the Library. The project also explored other topics, including workflow related to copyright permissions and notifications; roles and responsibilities for life-cycle management of Web resources; cataloging options for collected content; and costs associated with selection, capture, and maintenance.

Twenty-five project team members were selected from Library Services (LS) and OSI to work collaboratively on the project for a period of sixteen months. Four areas of collecting allowed participants to "learn-by-doing":

- Manuscript Division Archive of Organizational Web Sites (Manuscript Division): Web sites of existing donors, including civil rights and political advocacy groups, professional and honorary organizations, memorial groups, and research and educational organizations.

- Crisis in Darfur, Sudan (African and Middle Eastern Division): Organizations, news reports, and the responses of government, international organizations and the general public in the U.S. and worldwide to the crisis in Darfur.
- Visual Image (Prints and Photographs Division) Photography, graphic arts, and other visual material Web sites, which complement, expand and enhance the Prints and Photographs Division collections.
- Single Site: Development and coordination of a vetting process to determine which types of Web sites, unrelated to thematic collections, to archive.

Each of the four collection areas had four staff from LS working on collection development, selection, permissions, and quality review. The remaining members of the SMCCW project team supported their activities and provided technical expertise. They included members of the Web Capture team (OSI) and cataloging, metadata, and bibliographic access specialists (LS).

The Library of Congress contracted with the Internet Archive to perform the crawls, beginning in February 2006 and ending in November 2006, using the open source crawler Heritrix (http://crawler.archive.org). Across the four collections, a total of 377 seed URLs were selected, with 294 seed URLs ultimately being collected (the difference is mostly due to lack of response to permission requests. See Figure 1 for project breakdown. The number of objects (html, images, pdfs, etc.) totaled 114.5 million.

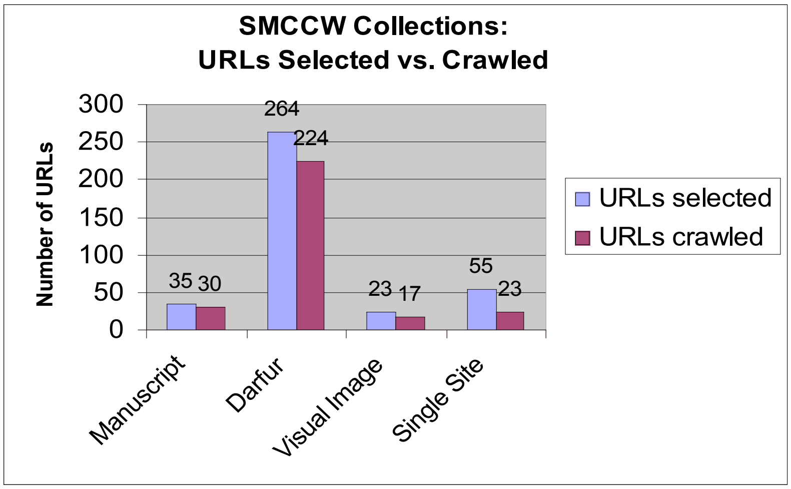

Figure 1: Number of URLs selected vs. collected. Most URLs selected but not collected were due to permissions (lack of response from site owner).

Workshops and Training

Given this symposium’s focus on skills needed by digital curators, a detailed review of the training provided to the newly appointed Web archive curators is in order. Based on prior experience, the Web Capture team knew that curators not previously familiar with Web archiving needed a basic introduction to the field before beginning to develop a Web collection plan. The trainers (project manager and Web Capture team) had briefed enough newcomers already to understand that what was needed for the curators to gain confidence to do the work was training in selection, technology, copyright and permissions, and access.

Selection Workshop

Many of the SMCCW participants were familiar with selecting analog materials for the Library's collections, and some had selected electronic resources to be cataloged, but selecting with the intent of archiving was a new concept to most. Some of the questions they needed to consider were: How do you define the scope of your content area? Which kinds of sites will be selected? (What's in/What's out?) How would you identify candidates for Web capture versus sites that might just be selected for a webliography (linking to but not saving a copy)? How frequently should a URL be collected, and what *is* a Web site? Is it a top-level domain (www.loc.gov), or does it also include different domains "owned" by the same content creator, such as www.copyright.gov and www.digitalpreservation.gov?

Presentations were made describing previous Library of Congress experiences selecting Web content. Each of the four content groups was charged with developing a collection plan in later months, and this workshop set the stage for them to think about what types of Web content to include in their collections.

Technology Workshop

A follow-up workshop on technology and Web content covered basic concepts, terms, and definitions and provided illustrative examples of the tools for capturing URLs and viewing Web archives. Participants reviewed test crawls and discussed how current tools affect crawling results and access. This foundation in Web capture technology proved helpful once the project teams began quality review of the crawls.

Throughout the project, it was sometimes difficult to strike a balance between sharing too much or too little technical information with the curators. For instance: Do they need to understand how a Web site is "scoped," i.e., how the technical staff gives instructions about what parts of a site to archive in a form the crawler can understand? In retrospect, it might have been better to reserve discussion of some of the more difficult concepts until closer to the curatorial analysis of the crawled content (see Quality Review training below). Feedback at the end of the project suggested that most of the technical issues, concerns, and fears curators had during the course of the project were resolved as they began to work with their collections. Ultimately, the curators seemed to have received an appropriate level of understanding of this aspect of the work.

Copyright Permissions and Access Workshops

The Library of Congress currently seeks permission from site owners to crawl and display archived Web sites to researchers offsite. A legal adviser to OSI gave a presentation on copyright and Web capture during the workshops so that curators understood their responsibilities in permission planning. For those who had prior experience negotiating literary and reproduction rights with donors or explaining copyright to researchers, these permission concepts were not totally new.

An overview of access tools was also provided, showing approaches taken at the Library of Congress and other institutions to deliver Web archives to researchers. New full-text search tools developed by the Internet Archive were demonstrated. Current limitations of interfaces and the access tools were stressed, as not everything collected can be displayed using these tools. Recognizing the rapid development in this area, the Web Capture team continuously updated the project team on the status of access tools being developed by the Web capture community.

Web Content Digital Collections Management Workshop

In July 2006, another workshop was held to discuss the management of digital collections, specifically Web archived content. Nicknamed "So We Got It---Now What?," this workshop provided an opportunity to gather the participants near the end of the project, evaluate lessons learned, and begin to discuss next steps, including issues related to curation of the archived materials long-term. While discussions were started at this workshop, much remains to be done in this area in the coming year.

Other Training

Another focus of training for the curatorial staff was to familiarize them with a tool that was developed in-house to manage the nomination of Web sites for archiving and to handle the permissions process. This tool, called "The Leaderboard," was custom-built in 2003 by the Library based on existing workflow processes.

Two different curatorial roles were defined for users of the tool, based on prior Web archive collection workflow: first, selection coordinators, and second, recommending officers. The latter were trained to be familiar with the collection plan and were asked to submit URLs using a Web form that would send the recommended URLs to the selection coordinator for review. Since selection coordinators were responsible for the overall coordination of the selection and nomination of URLs, they required more extensive training. They learned how to use the tool to review nominated sites, to assign metadata and send permissions, and to record instructions for the technical team. They became the main contact point when problems or questions arose with URLs in the collection.

A second aspect of training was to explain how to perform quality review (QR). In previous Web archives, only the technical staff performed QR after the URLs were crawled. However, it was believed that review by the content experts was important to make sure that the crawler captured what was wanted by the person or division that had selected the site for archiving. This proved to be quite valuable to the project and improved the overall quality of the archives. It will likely become part of future projects.

Case Study: Manuscript Division Implementation of the SMCCW Project

Although the SMCCW project consisted of four distinct pilot projects, they were similar enough that a close examination of only one–the Manuscript Division Archive of Organizational Web Sites– is sufficient for understanding the timeline, principal tasks, and skills involved in successfully achieving the overall project goals.

Assembling the Project Team

Having been assigned the task of identifying Web sites of organizations suitable for archiving, the division's first step was to assemble a core project team of four staff members, each providing a different perspective based on where he or she worked in the division. The team leader represented the division's "Front Office" consisting of the chief, assistant chief, administrative staff, and manuscript specialists, whose duties involve acquisitions, reference, research, and outreach. Other team members included a reference librarian from the Reference and Reader Services Section, the assistant head of the Preparation Section who provided a processing and technical services perspective, and the division's automation operations archivist. Although these four took the lead, the chief and five other manuscript specialists also became significantly involved with the selection and quality review of the crawled sites.

Timeline and Overview of Project's Implementation

The newly assembled four-person team attended the project's kick-off meeting in early August 2005. Training workshops followed along with internal team meetings to develop a selection strategy, identify eight test sites for capture, and begin work on a permissions plan. In late November, the team held two briefings for the chief and manuscript specialists to orient them to the project, show them some of the test crawls, and solicit their recommendations for additional sites to capture. Rather than have each specialist enter recommendations via the Leaderboard tool, the automated operations archivist prepared a WordPerfect form, which each specialist completed, and from which the team leader extracted information for entry into the Leaderboard. While selections were being made, the team drafted a permission letter for review by OSI's legal advisor.

The first permission letters were mailed in March 2006, and the responses (and questions) started arriving by fax and email shortly thereafter. The specialists answered some questions and referred to the team leader other inquiries requiring more detailed information. In some instances, the team leader sought guidance from the OSI Web Capture team. As permissions were received, the specialists forwarded the signed forms to the team leader, who recorded the information into the Leaderboard and retained the signature copies for the division's case files. Crawls began in April and occurred every month thereafter. Although the core team conducted periodic quality reviews throughout the project, the specialists undertook two detailed reviews – one in early June and another in late August. The last crawl occurred in November 2006.

Although additional work is needed to make the crawled sites accessible to researchers, three principal tasks, each requiring a slightly different skill set, emerged during the course of the project: appraising and selecting sites, securing permissions, and conducting quality review of the captured data. Other tasks and skills would have likely been involved if this project occurred in a different institutional setting that lacked the technical support and contracting services provided by the OSI Web Capture team.

Appraising and Selecting Content

In making selection decisions, the staff drew on its subject expertise in American history; its skill in appraising paper-based collections; its knowledge of an organization's mission and record-keeping practices; and its experiences navigating the Web and evaluating for itself and the Library's patrons the authenticity and accuracy of information found there. Also useful were understanding how Web sites are put together, instruction in how to select a seed URL, and training in what the Web crawling software cannot capture.

Already equipped with some of these skills and newly trained in others, the core project team drafted an initial selection plan, which was linked to the division's existing collection policies and holdings. This plan was further refined in discussions with the manuscript specialists. The division considered not only the current and future research value of the information contained on the Web sites but also the Library's relationship to the donor; the extent to which the Web site complemented, duplicated, or expanded information in the organization's paper-based records in the Library's custody; and the significance of an organization's activities in relation to the division's overall documentation strategies. Not surprisingly, differences of opinion surfaced, and the project sparked useful debates regarding the research value and donor benefit of collecting organizational Web sites. These discussions, in turn, led to a broader dialogue about other digital materials, which are now beginning to arrive with incoming paper-based collections.

The division was hesitant to seek permission from some organizations whose records it had previously declined or deaccessioned. In other cases, it decided differently. Even if it chose to decline additional paper records from specific groups, it still hoped to document their activities in some way and decided that capturing Web sites offered a compromise.

Although some staff expressed initial skepticism about the research potential of many Web sites, closer examination revealed a wide range of official documents, research studies, audio and video recordings, press releases, agendas and conference proceedings, blogs, electronic newsletters, and other sources documenting people, events, and activities likely to be of lasting research interest. Also deemed important was the need to document how organizations, many of which had been established early in the twentieth century, were incorporating technology and using the Web to reach new audiences and carry forth their mission into the twenty-first century.

Although diverse in content, the captured sites all have a connection to existing Manuscript Division collections. They fall into several broad categories. First, there are sites for non-governmental, voluntary organizations, including civil rights and political advocacy groups whose records the division holds. Examples include the National Urban League, Leadership Conference on Civil Rights, National Consumers' League, National Council of Jewish Women, and the League of Women Voters (LWV).

The last is an interesting use of this project to expand the scope of the Library's holdings, in that the Manuscript Division holds the paper records of only the national office of the LWV. Initially the team planned to collect only the national office Web site, but became impressed with the information found on many of the state and local sites and decided that capturing them provided a way to supplement the documentation in the paper archives without committing the division to acquire and preserve the local records. Fortunately the division's existing relationship with the LWV helped to secure permission from the national board to capture all the League's state and local sites without having to submit separate permission requests. With the addition of these state and local sites, the scope of the division's project expanded from thirty sites to approximately five hundred.

Other captured sites, whose records the division holds, include professional and honorary organizations such as the American Historical Association and American Studies Association; memorial groups such as the Vietnam Veterans Memorial Fund; and research and educational organizations, including two with whom the division has an ongoing relationship regarding their oral history programs.

The division also decided to collect the Web sites of organizations whose records it does not hold, but which are directly related to collections of personal papers in the division.

These included organizational Web sites related to Frederick Cook and Sigmund Freud, and the Web site of the National Association of Criminal Defense Lawyers, founded by Sam Dash, whose papers the division recently acquired. Recognizing that Web archiving is a service that may appeal to potential donors, the pilot also included the Web site of the National Endowment for Democracy, whose records the division has solicited.

Many organizations were thrilled to be included in the Library's pilot project. Of the organizations who responded to the division's permission request, none opted out. A few also expressed interest in having the Library explore a possible Intranet capture, which could ensure the preservation of records made available only in electronic form or offer an alternative to collecting some paper records that require more extensive processing or more costly storage than digital versions. Also, electronic files may facilitate improved access to the information. That having been said, the division needs to consider the costs of capturing and storing Web resources and weigh those costs against the research potential of the captured sites and the lost research potential of other traditional collections which may be forfeited because funds are shifted to acquire Web resources.

Permissions Process

Unlike the other three pilot projects, which sent form emails via the Leaderboard tool, the Manuscript Division opted to use a manual permissions process to foster and capitalize on existing donor relations. The division mailed personalized form letters to 33 of the 35 organizations whose Web sites had been selected. These letters were modified slightly for each recipient and signed by the specialist who recommended the site or who had a relationship with the organization. Of the 33 contacted, 20 responded favorably after the first letter, while others needed follow up inquiries. Eventually the division received permission from 28 of the 33 organizations contacted by letter and 2 contacted by form email, for a total of 30 of 35 sites or 86.6%. This statistic compares quite favorably to the 30 percent success rate recorded previously during other Library of Congress Web captures that relied exclusively on the email-generated permission requests.

Sending individualized permission requests was time-consuming and likely not scalable, although it was effective. It also provided the specialists with the opportunity to connect with donors with whom they had not had recent contact; learn of changes in personnel, mission, or record-keeping that may impact future additions to the Library; and educate organizations on the importance of archiving other born-digital materials they are creating. Sometimes, the initial permission request initiated an ongoing dialogue, through which the division later learned of organizations' plans to expand, modify, or relocate their live sites. In one instance, an organization alerted the specialist that its Web site was moving from one host institution to another, necessitating a change in the crawling instructions. In another case, a specialist learned that an organization was adding a blog to its Web site, again prompting a change in the crawling instructions to include the blog's URL which was different than the organization's other URLs.

Quality Review

In May 2006 after the first two crawls were completed, all division participants met to discuss how to conduct quality review on the captured sites. The core project team had received some training on this task, and it attempted to convey what was learned to the manuscript specialists. The quality reviews revealed that the crawled sites captured both more and less information than anticipated. With respect to the first scenario, when reviewing sites for possible inclusion in the pilot, some specialists concluded that parts of a Web site were suitable for capture while other parts were ephemeral and not as worthy of permanent retention. There was no easy or economical way, however, to implement these appraisal recommendations during the crawl. It was easier to have the crawler capture the entire site than for the technical team to compose site-specific instructions that excluded portions of the site. Also, for contracting purposes during the pilot, all sites were captured on a monthly basis, which was often more frequent than recommended.

Getting more content than expected, however, was not nearly as troubling as losing content. The most common problems noted during quality review were missing pages (including content that was not captured because portions of a site used a different URL than the seed URL initially selected or because the crawler could not capture pages generated from Web-enabled databases which require passwords or other user input); broken links; Javascript-enabled menus and links that redirect you out of the crawled version and into the live Web; and pages which the crawler captured but could not be located easily because the Wayback machine did not support the flyway menus, drop down boxes, or database-enabled searches designed to access those pages originally. Also some links became inoperable in the Wayback machine because of inconsistencies in the usage of upper and lowercase letters in the original coding.

The Manuscript Division pilot confirmed the need for quality review by division staff as well as by the technical support team, since the two groups are often looking for and finding different types of problems. Questions remain, however, about how best to implement that review. Some staff seemed better suited to finding problems with the crawled pages. The unevenness in the quality review may have been due to lack of clarity about what was expected; lack of time; lack of proofreading experience; lack of technical understanding; or greater acceptance of errors. Staff knowledgeable about Web design and who are familiar and willing to use both Internet Explorer and Firefox often uncovered more or different problems than staff who used only one browser or who lacked knowledge about how a Web site is constructed. A checklist to remind reviewers what to look for and to ensure consistent review across sites would be helpful. Even if the quality of the technical review is enhanced through checklists and better training, there will still remain, however, the need for a single point person to monitor the division's crawled assets, solicit and augment feedback from the specialists, verify reported problems, and compile consistent errors messages for the technical team.

Conclusion

Although the pilot project has ended, the Manuscript Division remains interested in continuing its Web archiving activities, despite unresolved questions about costs, workflow, and technology. The other pilot projects seem equally satisfied with their results. Certainly the Web Capture team believes that the SMCCW project has helped widen the expertise and understanding throughout the Library of how to select and manage content captured from the Web. Notable progress was made in resolving curation issues related to two stages of the digital life cycle--selection and acquisition of content. More work remains to be done on the other stages, particularly those related to providing and improving researcher access to Web archives and sustaining collections over time. In the coming year, the Library hopes to explore issues related to "Single Site" archiving and intends to study various access issues (cataloging, full-text searching, access tools). It is eager to implement the lessons learned from the SMCCW and to improve its collection and preservation of Web resources for use by future generations.