Collaborations, Best Practices, and Collection Development for Born-Digital and Digitized Materials

Kathleen Murray
Post Doctoral Research Associate
University of North Texas Libraries
P.O. Box 305190
Denton, TX 76203-5190
Email: krmurray@unt.edu

Mark Phillips
Head, Digital Projects Unit
University of North Texas Libraries
P.O. Box 305190
Denton, TX 76203-5190
Email: mphillips@library.unt.edu

Abstract

The University of North Texas (UNT) Libraries is a collaborative partner in the Web-at-Risk project, one of eight preservation projects funded by the Library of Congress under the National Digital Information Infrastructure and Preservation Program (NDIIPP). Early in the project, UNT conducted a needs assessment to identify web archiving issues facing curators and librarians. Key findings in three areas are reported: current challenges, organizational issues, and collection development concerns. These findings informed the development of guidelines and a template for the project's curators to create web collection plans. In addition to the Web-at-Risk project, the Digital Projects Unit at the UNT Libraries has several digital library initiatives with government agencies at the Federal and State levels to preserve and provide access to important collections of born-digital and digitized materials. The library also houses the Portal to Texas History, a digital gateway to the rich collections held in Texas libraries, museums, archives, historical societies, and private collections. Collaborations, best practices, collection development, and key lessons learned from these initiatives are identified. The DPU is also involved in trialing emerging tools and solutions for the libraries' storage architecture and information services. Building on research findings and experience gained in collaborations and projects, future plans for infrastructure and services are discussed.

Web Collection Development

Web-at-Risk Project

The Web-at-Risk project is one of eight collaborative projects funded by the Library of Congress as part of the National Digital Information Infrastructure Preservation Program (NDIIPP). It is a collaborative effort among several institutions, including the California Digital Library, the University of North Texas, and New York University. The California Digital Library is building a prototype Web Archiving Service (WAS) that several librarians, acting as curatorial partners in the project, are trialing. In the initial phase of the project a needs assessment was conducted to identify the issues faced by librarians, researchers, and content providers in relation to collection development for web-published materials and web archives. These assessment activities included an online survey completed by the project's 22 curators, five focus groups with 43 librarians and archivists, and interviews with seven academic researchers and seven content providers.

Needs Assessment Findings

Findings from all activities in the needs assessment were combined in the final analysis (Murray & Hsieh, 2006). The findings were organized in three major areas: (a) challenges in the current environment, (b) organizational issues, and (c) collection development concerns. A discussion of selected findings in each area follows.

Current environment

- Preservation activities for web-published materials, whether born digital or digitized copies, are beyond the capabilities of most libraries' IT staffs and technical infrastructures. Preservation efforts require IT resources at the institutional or organizational level.
- Compared to print materials, web-published materials are expensive to select, capture, and catalog. Few libraries have allocated additional funding to undertake these collection development activities. Many librarians lack the necessary technical skills.
- Content producers, especially smaller publishers, are unable or unwilling to undertake preservation of the materials they publish. This is not an historic responsibility of theirs and they rely on libraries and archives to preserve publications of enduring value.
- The two basic questions librarians ask in regard to identifying web-published materials for preservation are: "Should *we* save this?" and "Is *someone else* already saving it?" Librarians see a need for a nationally coordinated effort that would include a registry of archived web-published materials. Figure 1 depicts a registry to which repositories contribute standard metadata for the web collections in their repositories. Institutions would locate materials in other repositories and subsequently obtain or use these materials in accord with associated intellectual property rights.

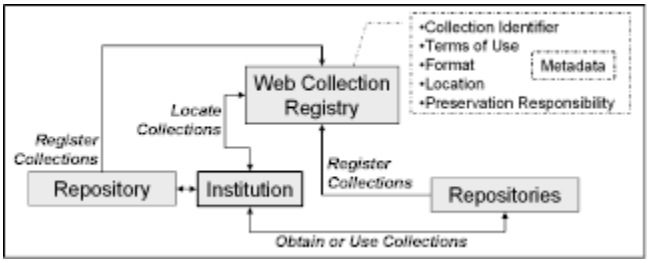

Figure 1. Registry Service for Web Collections

Organizational issues

- The major hurdles libraries and archives face in preserving web published materials are technology, policies, management commitment, and funding. The key funding demands are in the areas of cataloging, preservation, IT support, and staff training.
- Web archiving needs to be melded into the core fabric of an institution or organization in order to receive both the necessary management commitment and the sustained funding for technical infrastructure and staffing.
- Within an institution, cross-departmental, organization-wide collaborations and focus are needed. In many cases, libraries and archives need to deconstruct the barriers between their organizations and their institutional IT departments. Leveraging the expertise each group has to offer is a critical ingredient for creating successful institutional repositories and web archives.
- To both avoid unnecessary redundancy and maximize resource utilization, multi-institutional collaborations are needed to collect and preserve web-published materials. Extending consortia efforts to include web archiving is a possible vehicle for economically addressing the issues currently challenging individual institutions. Figure 2 depicts an institution, such as a large research library, acting as a shared preservation repository service provider. Smaller institutions, agencies, organizations, and publishing houses would provide their materials to the repository and contract with the service provider for a variety of information services, such as ingest, preservation, and metadata services.

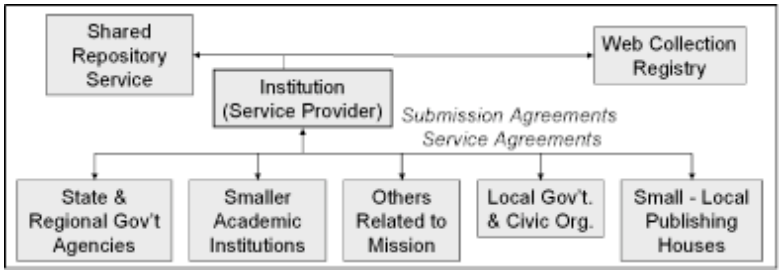

Figure 2. Multi-institutional Collaborations

Collection development concerns

- *User Groups*. Identification of user groups who will be served by a collection of web-published materials is critical to subsequent web collection development decisions. For example, some emerging academic disciplines use source materials differently from how catalogers are accustomed to describing them. In cultural studies, for example, advertisements in trade publications can be more important than articles.
- *Context and Content*. Capture and presentation of web-published materials is dependent on the materials (journal articles or datasets), research discipline (social science or history), and views of an archive (importance of preserving content versus context). For some content providers, their databases and datasets are the meat of their content and to varying extents all other content on their web sites is superfluous. These content providers are not concerned with preserving or replicating their web sites' "look-and-feel". Likewise, in certain research fields (e.g., statistical research in sociology), the original web-context of the source materials is not always critical and researchers would be better served by the ability to interact directly with the statistical datasets. However, for other research disciplines (e.g., anthropology and history) source material context is critically important and preservation and presentation of the original "look-and-feel" of web sites would be important.
- *Quality of Source Materials*. In general, all material types (e.g., a video file, an audio file, and a text file of the same speech) and possibly all formats (e.g., an image in both jpg and tiff formats) should be captured and preserved whenever possible even if the quality is poor. This is important not only for preservation of historical records but because different users will require access to different material types and different research disciplines rely on information contained in specific formats.
- *Frequency of Acquisition*. Assessing change in web-published source materials is often very time-consuming if done through human evaluation by a curator. It is highly desirable for a web capture system to evaluate changes in source materials. Curators can use these change measurements when making judgments about which content changes should trigger re-capture of materials.
- *Authenticity*. While different user groups assess authenticity differently, many need and most would want some authority to provide an assurance of the authenticity of web-published materials in a web archive. Researchers, users, and librarians generally trust that the materials libraries provide are authentic, however many researchers would like an archive to provide an indication of where the original source material is located.
- *Capture Scope*. In several instances, organizational or agency web sites rely on content from other organizations and these archived web sites would be of less value if that external content were not captured. This complicates capture scope decisions and requires more curator investigation into copyright permissions.

- *Crawler Effects*. Curators need to understand how a crawler handles links so that material critical to a collection is captured. Likewise knowledge of file formats that challenge crawler tools is critical so that key file formats are captured.
- *Metadata*. Curator tools are needed to support metadata creation at levels more granular than the web collection and web site levels. Many curators plan to build collections of web-published materials at a discrete object level, such as documents or images.

Guidelines for Web Collection Plans

Web collection plans or policies provide guidance for managing collections of web-published materials created for specific groups of users. The findings from the needs assessment informed a set of guidelines for web collection development plans. The Web-at-Risk project curators used these guidelines and a companion template to create web collection plans. The guidelines, template, and resulting collection plans are available at: http://web3.unt.edu/webatrisk/cpg.php.

Both the findings from the needs assessment and subsequent work with collection development personnel at libraries indicate that many institutions are eager to integrate a collection development service for web-published materials into their institutions. To address the challenges encountered in developing services related to new digital content, the UNT Libraries have implemented a modular infrastructure within their digital library services architecture. New digital initiatives at the Libraries, such as a collection development service for web-published and other types of digital content, are typically led by the Digital Projects Unit (DPU).

Digital Projects Unit

The Digital Projects Unit provides leadership and services to digital collections librarians in support of the libraries' digital library initiatives. Digital collections librarians are responsible for identifying new materials for inclusion in the collections they manage and for identifying new non-digitally formatted content for digitization. Three core services in support of librarians are provided: ingest, preservation, and metadata. The DPU staff includes six positions (FTE): a unit head, a metadata librarian, two digital library programmers, a digitization lab manager, and a library associate. Project management and financial oversight are the responsibilities of the Libraries' Assistant Dean for Digital and Information Technologies and the Libraries' Grants and Development Officer. The DPU also currently employs two graduate assistants, three students, and various project specific personnel. The number of employees relates directly to the size and complexity of contracts, research grants, and funded projects, such as the Web-at-Risk project. [http://www.library.unt.edu/digitalprojects/]

Ingest Service

Digital objects are collected or deposited in the archive either by file transfer using ftp, by copying files provided on disk drives or optical media, via web harvesting, or through in-house digitization of analog materials. Audits of content are performed to ensure all files are complete and then the digital objects are added to the ingest service queue. The ingest process creates a Submission Information Package (SIP) for ingest into the libraries' digital archive.

Preservation Service

Analog content of various types are converted to digital formats in the state of the art Digital Projects Lab within the DPU. The lab re-formats content for both preservation and access needs. The DPU also operates the web harvesters used in some digital initiatives.

Metadata Service

The DPU-developed systems for managing the preservation and descriptive metadata associated with digital objects serve as the core for digital initiatives and digital preservation activities at the UNT Libraries. The DPU monitors current trends in descriptive and preservation metadata initiatives and serves as a consultant for partners involved in its digital initiatives. A metadata advisory group within the library provides a community for discussing local implementation issues with various metadata formats. In most cases the DPU digital collections librarian is responsible for coordinating metadata creation. There are several workflows for metadata creation and the DPU consults with content curators to determine the ideal workflow for a given collection.

Existing Collections

CyberCemetery

This collection provides permanent public access to the web sites and publications of defunct U.S. government agencies and commissions. [http://govinfo.library.unt.edu/]

Content provider collaborations

- The UNT Libraries has two partnerships for the CyberCemetery: the Government Printing Office (GPO) as part of the Federal Depository Library Program partnership and the National Archives and Records Administration (NARA) as an Affiliated Archive. Permanent public access and digital preservation activities are carried out on behalf of GPO and NARA.
- Content acquisition is coordinated by library staff in the Government Documents Department.

Standard practices

- Over the years web sites have been acquired both by harvesting web content and by accepting content deposits directly from commissions and agencies. UNT has used several web crawlers including Teleport Pro, in the early years, and subsequently, HTTrack and Heritrix.
- Websites are audited for consistency and completeness and missing files are located and captured when possible.
- Once acquisition and auditing are complete, a site-level metadata record is ingested into the Digital Collections system and the web site is placed in the CyberCemetery.

Collection development

- A web site is a candidate for inclusion in the CyberCemetery if it is a federal government website for an agency or commission that is scheduled to cease its operation. Websites are typically captured as close to the time of closing as possible.
- Monitoring activities in the federal government is challenging and the digital collection librarian responsible for this collection employs various methods for tracking agencies and commissions. In many cases, web sites are identified by the government documents community or by agencies and commissions themselves.
- The CyberCemetery is patronized heavily by users from around the world who discover the collection either through web searches or via links to or searches of the UNT Libraries' Digital Collections system.

Key Lesson Learned

The downstream implications of requirements stated in memorandums of understanding with resource providers can be hard to anticipate and future modifications are sometimes necessary.

Texas Register

A weekly publication, the *Texas Register* serves as the notice bulletin of state agency rule-making. The *Texas Register* contains emergency, proposed, and adopted rules; notices of withdrawn and repealed rules; notices of rule review and other information submitted by state agencies for publication. [http://texinfo.library.unt.edu/texasregister/]

Content provider collaborations

- UNT has formed a partnership with the Texas Secretary of State's office to provide permanent pubic access to the electronic version of the *Texas Register*. A memorandum of understanding governs decisions involved with preservation and access.
- UNT plans to extend this partnership by digitizing back issues of the *Texas Register* in order to offer all historical and current issues of the Register in electronic format.

Standard practices

- The Texas Secretary of State's office makes a weekly deposit to UNT of a single tar file containing the html version and the PDF version of the *Texas Register*.
- The files are uploaded to a local ftp sever located in the library.
- Once received and checked for completeness, the files in the register are placed into the archive and made available to the public.

Collection development

- Collection development policies for this collection are straight forward. Future goals to digitize back issues of the register are consistent with the original intent of the project.

Key Lessons Learned

This collection spans 15 years of born digital information and issues with format migration emerged with the variety of file formats. Additionally, the importance of post-deposit audits was highlighted when several instances of partial data transfers and corrupt files were encountered.

Portal to Texas History

The Portal to Texas History is a multi-institutional repository for cultural heritage objects related to the history of Texas. This project provides a technical infrastructure to facilitate sharing of cultural heritage objects held by museums, libraries, historical societies, and archives throughout the state of Texas. To that end, multiple levels of service are provided, including content digitization, metadata creation, and hosting of digital collections. [http://texashistory.unt.edu/]

Content provider collaborations

- The Portal to Texas History currently has partnerships with over 45 content providers across the state of Texas. Each partner retains full intellectual property rights for their objects held in the Portal.
- Partnership agreements grant permission to UNT to make these materials available to the public free of charge through the Portal.

Standard practices

- After a partnership agreement has been established, Portal staff arranges to digitize new collections or to upload content which has been previously digitized by the partner.
- Metadata records are created at the digital object level both to facilitate resource discovery and to ensure object preservation. The records are packaged with the files and ingested into the system.
- Continuous management and modifications of descriptive metadata is necessary to provide enhanced access to the heterogeneous collections within the Portal.

Collection development

- Content is identified in two ways: new content providers contribute their collections and DPU staff works with UNT historians to identify primary and secondary source materials.
- The Portal's user groups are numerous and varied. They include elementary students, college students, professors, genealogists, and life-long learners. Usage grows proportional to the amount of content added.

Key Lessons Learned

Clear communications and expectations among the many collaborators are critical. The Portal is successful because of the commitment both DPU staff and contributing institutions bring to the provision of rich useful content to a wide audience of users.

Closing: Future Directions for Digital Initiatives

The DPU is investigating emerging tools and solutions for its storage architecture and is planning to implement a replicated architecture for data storage. The intent is to form partnerships to create geographically dispersed storage systems. The DPU is also planning to broaden the scope of its current ingest and metadata services and hopes to develop two new information services: presentation and collection development services. (See Figure 3.)

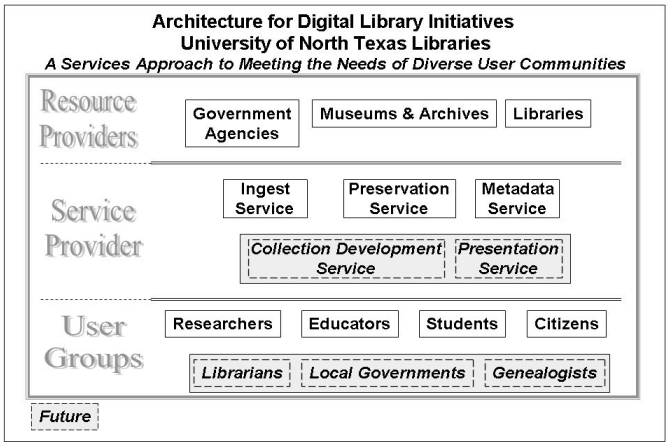

Figure 3. Future Information Services

Ingest Service

- Extend ingest system to recognize characteristics of more file-types
- Streamline current ingest workflow and improve "self-service" deposit into collections

Metadata Service

- Improve user interfaces for metadata creation
- Update metadata analysis tools for post-processing metadata
- Develop visualization tools to assist in metadata quality assurance

Presentation Service

- Informed by a user centered design process
- Targeted for specific user groups: local governments, genealogists, educators
- Developed using a rapid prototype framework

Web Collection Development Service

- Curator tools for web archiving:
  - Specification of crawl configuration parameters: file types, level of capture (URLs, domains, directories), and frequency of reacquisition
  - Application of metadata
  - Determination of new material additions and significant changes to existing source material on web sites in the collection(s)
- Keyword searching and subject browsing
- Migration of captured materials to new formats
- Validation of file integrity

Informed by experience and research, digital library service providers like the DPU are poised to create new services for curators of digital materials. In particular, collection development services for curators and presentation services designed for major user groups should enable improved discovery and dissemination of the rich materials collected in digital libraries.

Reference

Murray, K. & Hsieh, I. (2006, June 18). *Summary report of the needs assessment*. Retrieved March 12, 2007, from http://web3.unt.edu/webatrisk/digccurr2007/NaSumRpt.pdf

Acknowledgements

The Web-at-Risk project, a three-year collaborative project with the California Digital Library and New York University, is funded by the Library of Congress under the National Digital Information Infrastructure and Preservation Program (NDIIPP). Other funding partners for digital library infrastructure and content creation at UNT Libraries include: Texas Telecommunications Infrastructure Fund (LB9), Humanities Texas, Summerlee Foundation, Tocker Foundation, Summerfield G. Roberts Foundation, Forrest C. Lattner Foundation, Institute for Museum and Library Services, TexTreasures, and various private donors.

Copyright