Bentley Historical Library Web Archives: Collection Development Policy

Nancy Deromedi and Michael Shallcross
Digital Curation

Version 2.0 (August 2, 2011)

Table of Contents

Introduction

This document describes the Bentley Historical Library Web Archives and the policies that have shaped the development of its collections. Beyond its internal use, this document is intended to serve as a reference for researchers and content owners as well as a model for other archives and libraries engaged in website preservation. In addition to identifying the mission, scope, and key responsibilities of the Bentley Historical Library Web Archives, this policy document articulates the library's underlying criteria and strategies for collection development.

Given the rapid pace of change in the content, features, and functions of websites and the continuous development of website preservation technology, this document will be reviewed on an annual basis and revised accordingly.

Mission of the Bentley Historical Library Web Archives

The Bentley Historical Library was established in 1935 by the University of Michigan Regents to carry out two functions: to serve as the official archives of the university and to document the history of the state of Michigan and the activities of its people, organizations and voluntary associations. The library is currently comprised of four divisions: Michigan Historical Collections (MHC), University Archives and Records Program (UARP), Digital Curation, and Reference and Access Service.

The Bylaws of the University of Michigan Board of Regents (Sec. 12.04) identify the mission of the library's collecting units as follows:

> The Michigan Historical Collections will be maintained for the purpose of collecting, preserving, and making available to students manuscripts and other materials pertaining to the state, its institutions, and its social, economic, and intellectual development. The University Archives and Records Program shall be maintained to collect, preserve, and make available the records generated by the university in the conduct of its business.

The Bentley Historical Library has long recognized the administrative and historical value of online resources and has preserved select websites with open source software applications since 2000. As websites have increased in complexity and significance at the university and across the state, archivists sought a more efficient and cost-effective means to document and preserve significant online records. On July 1, 2010 the Bentley Historical Library began a subscription to the California Digital Library's Web Archiving Service (WAS). This arrangement permits archivists to focus on the identification, appraisal, and description of content while the California Digital Library (CDL) supports the requisite hardware and software infrastructure.

Scope of the Bentley Historical Library Web Archives

The Bentley Historical Library administers two collections of preserved websites (referred to collectively as the Bentley Historical Library Web Archives): the Michigan Historical Collections Web Archives and the University of Michigan (U of M) Web Archives. Each collection reflects the mission and priorities of the institution's two collecting divisions and preserves online resources of unique, essential, and enduring value.

Archived websites parallel and complement the library's manuscript collections and record groups (both paper-based and digital). As such, the MHC Web Archives focus on the activities of individuals and organizations in the state of Michigan, with particular emphasis on religious groups, social justice, ethnic communities, commerce and industry, and politics. The U of M Web Archives, on the other hand, document the university's administration, academics, research, athletic competitions, student life, and cultural activities.

Archivists are diligent in their efforts to identify new, modified, or decommissioned websites so that the Bentley Historical Library Web Archives reflects the scope and fluidity of the University of Michigan web domain and the web presence of prominent individuals and organizations in the state of Michigan. Given the dynamic nature of this enterprise, the Bentley Historical Library Web Archives may be viewed as ongoing projects.

Designated Community of Users

The Bentley Historical Library Web Archives will be of interest and value to researchers with varied interests related to the state of Michigan and the University of Michigan, members of the community at large, and faculty, students, staff, administrators, and alumni of the university.

The Bentley Historical Library welcomes comments, questions, and suggestions in regards to these collections. Please feel free to contact Nancy Deromedi or Michael Shallcross from the Digital Curation Division at bhlwebarchive@umich.edu.

Roles and Responsibilities in a Subscription-Service Model

The Bentley Historical Library's subscription to WAS requires the library and the CDL to fill specific roles and responsibilities. While no active participation is required of content owners, they may take several steps to ensure that websites are preserved as completely as possible.

Bentley Historical Library:

- Identify, appraise, and select websites that reflect the mission and collecting interests of MHC and UARP.
- Organize and manage archived websites to complement current holdings at the Bentley Historical Library.
- Provide descriptions and contextual information for materials.
- Mediate access (via metadata, catalog records, and an access interface) to facilitate the search and retrieval of content.
- Respect the intellectual property rights of content owners.
- Communicate the goals and objectives of the Bentley Historical Library Web Archives to researchers, donors, and the professional community.
- Reach out to webmasters when site design or configurations pose issues for the accurate capture of content.
- Work with Reference and Access Services to encourage new uses and applications for preserved content.
- Monitor the development of tools, relevant technical issues, and larger trends in web archiving related to access, description, rights management, etc.
- Respond to questions, comments, and suggestions so that the Bentley Historical Library Web Archives are of maximum value to donors and patrons.

California Digital Library:

- Ensure archivists have reliable and continued access to the web-based interface of the Web Archiving Service.
- Maintain and configure the essential tools and infrastructure for website preservation (these include Heritrix web crawler, NutchWAX search engine, and Wayback Machine archival web browser as well as associated servers, databases, indices, etc.).
- Develop new features and functionalities within WAS to improve the archivist's ability to capture and manage content.
- Provide basic digital preservation activities:
    - Secure storage of captured web content at the San Diego Supercomputer Center
    - Two versions of content on spinning disc with an additional version backed up to tape
    - Fixity checks
    - Disaster recovery

- Host web-ready content from web servers in the University of California Office of the President Data Center in Oakland, CA.
- Offer general technical assistance and customer support.

Content owners:

- Rely upon the Bentley Historical Library to preserve and provide access to multiple versions of select websites over time.
- Follow best practices for the design and maintenance of websites and the presentation of content (cf. UARP's Guidelines for Web-Disseminated Records or Google's Webmaster Guidelines).
- Allow the Bentley Historical Library to preserve their website(s).
    - The library uses the Heritrix web crawler (also known as a *spider* or *robot*) to preserve a version of websites.
    - A web crawler is a software application that starts at a specified URL and then methodically follows hyperlinks to copy html pages and associated files (images, audio files, style sheets, etc.).
    - Our web crawler will only capture publicly available content and cannot access materials that are password protected, require user authentication, or are excluded by robots.txt files. Intranets, private directories, and network-attached storage are strictly off limits.
    - Content owners can ensure that their website(s) will be preserved by including the following exception in the host's robots.txt file[1]:

        User-Agent: cdlwas_bot
        Disallow:
- Inform the library if a website is scheduled to be launched, decommissioned, or undergo significant changes.

[1] A "robots.txt" file is an Internet convention used by webmasters to prevent all or certain sections of websites from being crawled by a robot. The robots.txt must reside in the root of the site's domain and its presence may be verified by typing '/robots.txt' after the root URL (i.e. http://umich.edu/robots.txt). By convention, a web crawler or robot will read the robots.txt file of a target site before doing anything else. This text file will specify what sections of a site the robot is forbidden to crawl. A typical robots.txt exclusion statement is as follows:

User-agent: *
Disallow: /

'User-agent' refers to the crawler; '*' (a wildcard symbol) indicates that the exclusion applies to all robots; and '/' applies the exclusion to all pages on the site. Alternatively, a webmaster might exclude only certain directories (entering each one on a separate line) or open the whole site to a robot (in which case the file would read "Disallow: ". For more information, see http://www.robotstxt.org/.

Intellectual Property Rights

The Bentley Historical Library strives to respect the rights of content owners and to follow professional best practices for intellectual property rights management in website preservation. As part of its subscription to WAS, the library follows the Section 108 Study Group's recommendations for changes to the Copyright Act for website preservation. This group of copyright experts asserts that archives and libraries have the right to capture "publicly available" content (i.e. materials that do not require a password, entry forms, or subscriptions) and that all governmental websites should be freely accessible to web crawlers. To address the rights of content owners, the Bentley Historical Library has taken the following steps (on its own and as a subscriber to WAS):

- The WAS web crawler is configured to respect all exclusions in robots.txt files and will not capture content designated as off-limits by a webmaster.
- WAS will stop a capture if it detects any degradation of service or negative impact on the host's web server.
- Websites will be embargoed for six months so they will not be mistaken for originals or divert viewers from 'live' sites.
- All preserved materials will be prominently labeled as an "archived copy for study and research" to avoid confusion with the live websites.
- Content owners may request that portions of their site be suppressed from public view and can choose to opt out entirely from captures.
- The web archives may contain the websites of individuals or organizations. When such content has been preserved, the Bentley Historical Library will notify content owners that their website has been preserved, inform them of their right to opt out or suppress content, and invite questions or concerns (see Appendix A for a sample letter to U of M faculty).

The Bentley Historical Library furthermore strives for transparency in the policies, administration, and activities of its web archives. The Digital Curation Division provides monthly reports on the web archives and seeks to publicize policy and procedural documents for the benefit of donors, researchers, and other information professionals.

Access to the Collections

Content stored in the Bentley Historical Library Web Archives may be browsed, searched, and accessed via the homepages of the University of Michigan Web Archives and the Michigan Historical Collections Web Archives. In addition to these resources, the Bentley Historical Library has developed additional access points to encourage access to its collections of preserved websites. These access points include

- **Bentley Historical Library Web Pages:** In anticipation of the fall 2011 publication of the MHC Web Archives, archivists are developing a unified portal on the BHL homepage that will allow patrons to access both the University of Michigan Web Archives and the Michigan Historical Collections Web Archives. The current U of M Web Archives access page has basic functionalities that will be featured in the new resource: a description of the web archives' content, access options (including full-text search and browsing), an example of preferred citation style, and information on the administration and development of the collections.

- **Finding Aids:** The Bentley Historical Library include archived websites in the finding aids of select record groups and manuscript collections. UARP has already included archived websites as a series in the finding aids for high priority record groups (the Board of Regents, President, Provost, and the 19 schools and colleges). Scope and content notes indicate the overall purpose and function of the archived site and explain that captures will continue on a regular basis and online finding aids contain direct links to archived content. Archived websites will be added to relevant finding aids in the future when archival processing mandates additions to the finding aids.

- **MARC Catalog Records:** The Bentley Historical Library plans to create MARC records for high-priority archived websites (i.e. those of major administrative units and the 19 schools and colleges) in Mirlyn, the University of Michigan Library's online catalog. These basic records will include the names of sites and creators as well as direct links to the archived content. As of July 2011, archivists are experimenting with the batch creation of records (using MarcEdit, a free MARC editing tool) which may then be submitted to the University of Michigan Library or OCLC.

Collection Development Policies

The Bentley Historical Library has articulated collection development policies for both UARP and MHC that govern the identification, appraisal, and selection of content for the respective web archives of each division. These policies are informed by the library's main collecting priorities, archival principles, professional best practices, and analyses of manuscript collections and record groups.

The Bentley Historical Library is mindful of the widespread use and significance of social media and Web 2.0 technologies at the University of Michigan and across the state. Archivists have been unable to preserve social media websites (as of August 2011) due to various technical difficulties. In addition to the challenges posed by the structure and design of social media sites, robots.txt exclusions have severely restricted the library's ability to archive such resources. Moving forward, archivists will work with content owners and the California Digital Library to develop interim solutions and also monitor the work of the International Internet Preservation Consortium (IIPC) to preserve social media sites more effectively.

University Archives and Records Program

The collection development policy for the University of Michigan Web Archives is based upon UARP's *Records Policy and Procedures Manual*, the University of Michigan Standard Practice Guide 601.08, and the mandate set forth in Section 12.04 of the Board of Regents Bylaws.

Selection Criteria

For inclusion in the University of Michigan Web Archives, a website must meet the following criteria:

- The website falls within UARP's collecting scope as it is established by the *Records Policy and Procedures Manual*. It should be created, owned, or used by university units, faculty, or students in carrying out university-related business or functions. This guideline excludes web pages about—but not *by*—the university (such as online articles in *The Chronicle of Higher Education*).

- The website complements or has related material among manuscript collections and record groups. UARP seeks to expand upon existing holdings or develop areas that have been previously under-documented.

- The informational/evidential value of the website is made clear in its representation of administration, instruction, research, creative work, competitions, or social events at the University of Michigan. The website should contain meaningful content and adequately illustrate or promote understanding of its subject matter.

- The website and the content therein are unique.

- The website is not merely transactional or related to the delivery of routine products or services.

- The website reflects basic functions or activities associated with colleges and universities: administration, teaching, research, service, student life, and athletic competitions.

To ensure that its policy remains flexible, UARP has identified several exceptions to the above criteria. On a case-by-case basis, archivists may consider websites related to alumni or organizations, individuals, and events affiliated with (but not part of) the university. Archivists may also select a wider range of content in case of important events, breaking news, or upon special request by university units.

Collecting Priorities

The *Records Policy and Procedures Manual* outlines UARP's basic collecting priorities. In developing the University of Michigan Web Archives, UARP has followed these priorities in an initial two-phase process of systematic website preservation. In addition to the ongoing maintenance of existing collections and selection of newly released content, archivists may launch additional phases in response to new projects or initiatives within UARP or developments in the university's online presence.

Phase 1: July 2010 – February 2011

In this phase, UARP initially focused on its highest collecting priority: administrative and academic units, a category that includes all major administrative offices as well as the 19 schools and colleges of the main campus. Sites related to these units were analyzed for the inclusion of content related to research, instruction, and creative work within the schools and colleges. Particular emphasis was placed on collecting web pages related to faculty members from the School of Art + Design and the School of Music, Theatre & Dance, since these individuals and units have been under-documented in existing record groups and collections. This phase also involved preserving websites related to the university's centers and institutes, museums and libraries, and athletic department.

Phase 2: February 2011 -

The second phase of UARP's collection development for the University of Michigan Web Archives involves the broader selection of websites related to prominent faculty members, research projects, and student organizations. Special mention needs to be made in regards to the appraisal and preservation of faculty and student organization websites. In addition to the above-mentioned criteria, the selection of faculty member websites will depend upon:

- The faculty member's prior selection for inclusion in the University Archives.
- The faculty member's professional stature, awards, and recognition (including named chairs).

- Use patterns and frequency of updates for the site in question.

Archivists conducted a survey of student organization websites in 2010 and will use this information as a basis for preservation decisions. The selection of student organization sites will involve this information as well as a consideration of the following guidelines:

- The organization's prior selection for inclusion in the University Archives.
- The stature, history, and organizational viability of the group.
- Use patterns and frequency of updates for the site in question.

The preliminary survey suggested that student groups are using Facebook and Twitter more frequently than traditional websites; UARP may therefore explore the preservation of such content in the future.

Ongoing Activities (as of 2011):
Collection development for the University of Michigan Web Archives will involve the active maintenance and upkeep of archived content and the identification, appraisal, and selection of newly released content in accordance with the above-mentioned priorities. Archivists will evaluate captures and remove content that has significant technical issues and may revisit earlier appraisal decisions if the archived version of a website is missing significant content. Archivists will also review websites of the highest priority groups to ensure that they have not undergone significant changes that could impact preservation (such as changed host names/URLs). This ongoing work will require archivists to stay abreast of news reports and maintain relationships with unit webmasters to be aware of significant changes to or new releases of high-profile sites.

Michigan Historical Collections

The collections development policy for the Michigan Historical Collections Web Archives is based upon the mandate set forth in Section 12.04 of the Board of Regents Bylaws and MHC's existing collecting priorities.

Selection Criteria

Since 1986 the Michigan Historical Collections has used a process of collecting priorities to guide its acquisition of archives and manuscript collections. In selecting websites for permanent preservation, we work within our highest topical priorities, and follow these selection criteria:

- Websites of organizations and persons whose archives we are committed to preserve.

- Websites of other organizations and persons to fill gaps in our collections.

- Websites that are well developed with rich content documenting the work and thought of the person or organization.

- Websites that periodically incorporate new content.

- Websites with content that is not likely to be duplicated in an individual or organization's paper records.

Collecting Priorities

Based on the library's mission as established by the University of Michigan Board of Regents to document "the state, its institutions, and its social, economic, and intellectual development" and the historical collecting patterns of the library, the MHC developed a list of 19 topical collecting areas: Agriculture, Commerce and Industry, Communications, Creative Expression, Education, Ethnicity, Family, Gender and Sexuality, Labor, Leisure, Military, Natural Resources, Pioneer Michigan, Politics and public policy, Professionals, Recreation, Religion, Science and Technology, and Transportation.

Within these 21 areas, and working to document the entire state of Michigan, a set of priorities has been developed and is periodically reviewed and adjusted. The process of setting collecting priorities is described by Christine Weideman's "A New Map for Field Work: Impact of Collections Analysis on the Bentley Historical Library"[2] and Judith E. Endelman's "Looking Backward to Plan for the Future: Collection Analysis for Manuscript Repositories."[3]

[2] *American Archivist*, Winter 1991, Vol. 54, Issue 1, pp. 54-60.
[3] *American Archivist*, Summer 1987, Vol. 50, Issue 3, pp. 340-355.

Version History

The Bentley Historical Library will review this collection development policy on an annual basis and make updates to reflect changes to the Web Archiving Service, archived websites, archival best practices, or other relevant issues.

Version No.	Date	Reviewed By	Amendments
2.0	August 2, 2011	Michael Shallcross	Consolidation of policies for both UARP and MHC.
1.1	April 11, 2011	Francis X. Blouin, Director	Clarification of web archiving terminology and selection criteria.
1.0	March 23, 2011	Nancy Bartlett, University Archivist	Clarification of purpose, scope, and roles and responsibilities; general editing.
0.9	March 17, 2011	Brian Williams, Associate Archivist	General editing and clarification of roles and responsibilities.
0.8	March 11, 2011	Nancy Deromedi, Associate Archivist	Clarification of procedures; inclusion of Appendix; general editing.
0.1-0.7	March 8, 2011	Michael Shallcross, Assistant Archivist	Original drafts

Appendix A: Sample Communication to University of Michigan faculty

The University Archives and Records Program (UARP) at the Bentley Historical Library would like to preserve your personal website (at)as a representation of your academic career at the University of Michigan. Since your website is likely to change over time, UARP intends to capture it on an annual basis to document the evolution of your work and accomplishments.

Since its inception in 1935, the Bentley Historical Library has served as the official archives of the University of Michigan. While web pages have long been recognized as valuable university records, UARP launched the University of Michigan Web Archives in July 2010 to capture and preserve a greater number of historically significant sites. As part of this collection, your website would be an important record of your intellectual contributions to the U of M. Additional information on this initiative may be found at http://bentley.umich.edu/uarphome/webarchives/index.php.

This letter is intended to share information about our project and respond to any questions or concerns you might have. UARP will preserve your website with a 'web crawler,' a computer application that methodically copies the site's content and structure. This archived version will then be made available to researchers through the University of Michigan Web Archives at http://webarchives.cdlib.org/a/universityofmichigan. UARP will only capture and preserve publicly available materials and will never copy content that is password protected or requires registration or data entry. In addition, all preserved content will be embargoed for six months before being made public and will then be prominently labeled as an "archived copy for study and research" to avoid confusion with your live website. This process involves no special preparation of the website and is designed to have no negative effects on your web server's performance.

As the content owner, you have the right to opt out of allowing UARP to preserve your website and you may also request that specific sections of your site be suppressed from public view in the Web Archives. If you would like your site to become part of the University of Michigan's official archives, no further actions are required on your part.

Please contact Michael Shallcross or Nancy Deromedi by phone (734.764.3482) or by email (bhlwebarchive@umich.edu) should you have any questions or concerns about the Bentley Historical Library's University of Michigan Web Archives or if you are interested in establishing an archive of additional content from your academic career. For more information about faculty collections in the University Archives, please see http://www.bentley.umich.edu/research/um/facpapers.php.

Thank you for your time and consideration.