Bentley Historical Library Web Archives
http://bentley.umich.edu bhlwebarchive@umich.edu (734)764-3482 1150 Beal Ave. Ann Arbor, MI 48109

Bentley Historical Library
Web Archives Collection Development Policy and Methodology

Nancy Deromedi and Michael Shallcross
University Archives and Records Program

Version 1.1 (April 11, 2011)

Table of Contents

Abstract

This document reflects the current state of the University of Michigan Web Archives and will serve as the foundation for the University Archives and Records Program's future work on this collection. 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. The document is intended to inform the university community, information professionals, and users of archived University of Michigan websites.

Purpose

The purpose of this document is threefold:

- To state the mandate and scope of the University of Michigan Web Archives.
- To articulate the policy (and underlying criteria and strategies) for collection development.
- To explain the University Archives and Records Program's methodologies for content acquisition, quality assurance, access, and management of intellectual property rights.

Mission of the University Archives and Records Program (UARP)

The Board of Regents founded the Bentley Historical Library in 1935 to serve as the official archives of the University of Michigan. The University Archives and Records Program's mandate to identify and preserve select university records of enduring historical and administrative value, regardless of their form, is made explicit in University of Michigan Standard Practice Guide 601.8. The Bentley Historical Library's *Records Policy and Procedures Manual* further defines the nature and scope of the University Archives and its collecting activities.

Scope of the University of Michigan Web Archives

The University of Michigan Web Archives is dedicated to the documentation and preservation of University of Michigan online resources of unique, essential, and enduring value. These archived websites parallel and complement UARP's manuscript collections and record groups (both paper-based and digital). Many important documents are now available exclusively online and the expansion of interactive features, forms, and collaborative resources have made websites indispensable to the university's daily operation. Although certain content types and features are difficult to capture and render accurately, UARP aims to preserve academic and administrative policies; significant publications and resources; important representations of research, instruction, and creative work; and the overall look and feel of the University of Michigan web domain.

Roles and Responsibilities for the University of Michigan Web Archives

UARP has preserved select university websites with open source software applications since 2000. As the University of Michigan web presence has increased in volume and complexity, archivists needed 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. In addition to UARP and the CDL, University of Michigan units and individual content owners represent a major stakeholder group in this undertaking. The following represents the major roles and responsibilities shared by each of these entities:

The University Archives and Records Program (UARP) will…

- Identify, appraise, and select University of Michigan websites to document its academic, administrative, research, athletic, public service, and social activities and also to reflect the evolving nature of the university's web presence.
- 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:
    - Distinguish 'archived' sites from 'live' content with a prominent banner and statement at the top of each preserved web page.
    - Embargo archived content for six months after its so that the archived copy will not be mistaken for the original or divert viewers from the 'live' site.
    - Suppress content from public view or refrain from website preservation at the request of content owners.
    - Inform individual content owners of their rights.
- Communicate the goals and objectives of the University of Michigan Web Archives to the campus community and beyond.
- Reach out to webmasters when website design or configurations pose issues for the accurate capture of content.
- Work with the Bentley Library's Reference and Access Services and other units on campus to encourage new uses and applications for the University of Michigan Web Archives.
- 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 UARP may provide content and services that are of most value to its clients and patrons.

The California Digital Library (CDL) will…

- 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 in a digital preservation repository 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.
- Ensure that the general public has continuous access to the University of Michigan Web Archives and resolve service outages or technical problems.
- Offer general technical assistance and customer support.

University of Michigan content owners will be able to…

- 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).
- Permit UARP to preserve their website(s).
  - To preserve websites, UARP uses the CDL's version of the Heritrix web crawler (also known as a *spider* or *robot*).
  - 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:

```
User-Agent: cdlwas_bot
Disallow:
```
[1]

[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

- Inform UARP if a website is scheduled to be launched, decommissioned, or undergo significant changes.

Collection Development Policy

This collection development policy governs the identification, appraisal, and selection of content for inclusion in the University of Michigan Web Archives. It is based upon specifications set forth in UARP's *Records Policy and Procedures Manual* and further informed by archival principles, professional best practices, and regular analyses of the collections held by the University Archives.

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.

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/.

- 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[2] 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.

[2] The University of Michigan Standard Practice Guide provides organizational charts that detail the institution's administrative hierarchy. "Office of the President: 753000." (October 2010). *Standard Practice Guide*, University of Michigan. http://spg.umich.edu/pdf/753000.pdf. (Accessed February 25, 2011).

- 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.

Social Media

As of March 2011, UARP has not included social media sites such as Facebook, Twitter, or YouTube in the University of Michigan Web Archives. This decision was based upon the active appraisal of content as well as technical limitations inherent to website preservation. Archivists conducted an extensive review of the use Facebook and Twitter by various departments and discovered that these sites largely repeat news and information posted to other university web pages. At the same time, the structure and design of social media sites pose significant challenges for website preservation. Facebook’s robots.txt exclusions and its Automated Data Collection Terms limit the crawler’s ability to access content and Twitter and YouTube pages do not capture and/or display properly.

UARP remains mindful of the significance of social media at the University of Michigan and the potential for such content to be preserved in the archives. As a result, archivists will:

- Monitor the work of the International Internet Preservation Consortium (IIPC) and others to improve access to such sites and capture them more effectively.

- Reassess the use and administrative/historical value of university-affiliated social media sites in 2012.

Acquisition of Content

UARP has developed a workflow for the acquisition of content that reflects features of the Web Archiving Service but adheres to basic archival procedures (such as appraisal and description). Guided by its collecting priorities, regular surveys of the university web domain, and knowledge of significant record creators, archivists created an extensive spreadsheet of potential targets in early 2010 that influenced the initial selection of content for the web archives. This spreadsheet was arranged according to provenance so that archivists could track the various websites associated with different units and organizations. Additional content was unearthed as archivists navigated through web pages and reviewed the contents thereof in terms of UARP's stated selection criteria. To better understand the actual process of website preservation, the workflow may be broken down into three main steps: the identification of the crawl target, configuration of the crawler settings, and contextualization of content with descriptions and metadata.

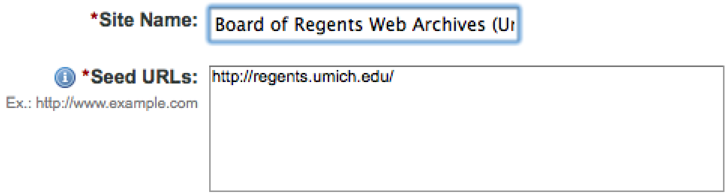
Identification of the Crawl Target and Seed URLs

As mentioned above, UARP uses the CDL's version of the Heritrix web crawler (also known as a spider or robot) to copy and preserve 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.) as well as the websites underlying structure. The initiation of a web capture thus requires the archivist to specify one or more seed URLs from which the web crawling application may begin to archive the target site. Accurate and thorough website preservation requires the archivist to become familiar with a site's content and architecture in order to define the exact nature of the target.

This attention to detail is important because content on many sites is hosted from multiple domains. For example, the Horace H. Rackham School of Graduate Studies hosts the majority of its content at http://www.rackham.umich.edu/ but maintains information on academic programs at https://secure.rackham.umich.edu/academic_information/programs/. To completely capture Rackham's online presence, archivists needed to identify both domains as seed URLs.

At the same time, multiple domains present on a site may merit preservation as separate websites. For example, the Office of the Vice President of Research (http://research.umich.edu/) maintains a large body of information related to research administration (http://www.drda.umich.edu/) and human research compliance (http://www.ohrcr.umich.edu/). Although these latter sites could be included as secondary seeds for the Vice President of Research's site, their scope and informational value led archivists to preserve them separately.

Once the target of the crawl has been identified and defined, the archivist enters the seed URL(s) and site name in the WAS curator interface (see Figure 1).



*Site Name: Board of Regents Web Archives (Ur

*Seed URLs: http://regents.umich.edu/

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

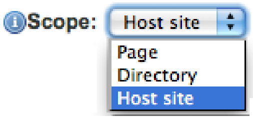
Figure 1

UARP standardizes the names of preserved sites by using the title found at the top of the target web page or, in the absence of a formal/adequate title, the name of the creating unit. UARP follows the best practices for collection titles as established by Describing Archives: a Content Standard (DACS); to ensure that the provenance and nature of the collections are clear in titles, archivists supply "Web Archives" and "(University of Michigan)" to the final title. Complete names for archived websites thus follow the pattern "Board of Regents Web Archives (University of Michigan)."

Configuration of Web Crawler Settings

WAS utilizes the open-source web crawler Heritrix to archive websites. As a command-line tool, this application allows for a wide range of user settings; the curator interface in WAS provides for a more-limited number of options. For each crawl, archivists may adjust the following settings:

- **Scope**: defines how much of the site will be captured. The archivist may elect to capture the entire host site (i.e. http://bentley.umich.edu/), a specific directory (i.e. http://bentley.umich.edu/exhibits/), or a single page (i.e. a letter written by Abbie Hoffman to John Sinclair, featured at http://bentley.umich.edu/exhibits/sinclair/ahletter.php) (see Figure 2).



Scope: Host site

Figure 2

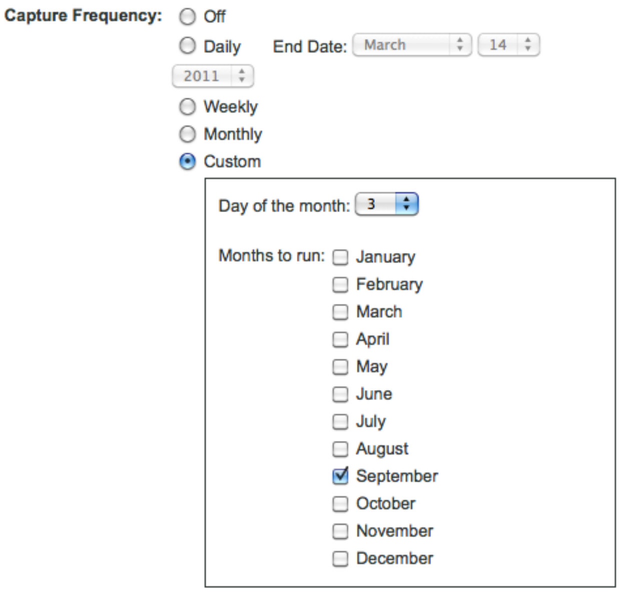
In the interest of having a comprehensive record of the university's web domain, UARP generally uses the "Host site" setting, unless the target is a single directory located on a more extensive host (as, for example, the President's website: http://www.umich.edu/pres/) or a specific page (such as the university online events calendar at http://events.umich.edu/calendar.cfm).

Linked pages: determines whether or not content from other hosts/URLs will be captured; archivists have two options for this setting. If set to "No," the crawler will only archive materials on the seed URL entered by the archivist; if "Yes," the crawler will follow hypertext links one 'hop' to capture linked resources. Capturing linked pages will not result in an indefinite crawl (in which the robot follows link after link after link); instead, the crawler will only capture the page (and embedded content) that is specified by the hypertext link. No additional content on this latter site will be crawled.

UARP captures linked pages as a default to ensure that important contextual information is preserved during web crawls. Archivists may choose to not capture linked pages if the target involves highly specific information limited to a single site (or page) or if the site has a large number of links to external/non-U of M websites. The introduction of the Solr indexing engine to WAS in 2011 will help reduce the amount of duplicate content captured in various crawls.

Maximum time: specifies the maximum duration of a crawl. The archivist may select "Brief Capture (1 hour)" or "Full Capture (36 hours)" and the crawl will continue until all content has been preserved (in which case it may end early) or the allotted time period has elapsed. If a session times out before the crawler has finished, the resulting capture may be incomplete. To avoid missing content due to time restrictions, UARP uses the "Full Capture" option by default. Archivists use the "Brief Capture" if the target involves a limited amount of content and the additional crawl time would result in unnecessary content (for instance, the archivist only wants to capture a blog's most recent posts and is not interested in the entire site).

- **Capture frequency**: designates how often a crawl will be repeated. The archivist may elect to crawl a site once or configure the robot to perform daily, weekly, monthly, or custom captures (see Figure 3).



Capture Frequency: ○ Off
○ Daily End Date: March 14 2011
○ Weekly
○ Monthly
● Custom

Day of the month: 3

Months to run:
□ January
□ February
□ March
□ April
□ May
□ June
□ July
□ August
☑ September
□ October
□ November
□ December

Figure 3

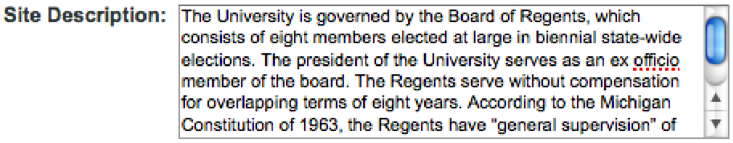
UARP generally chooses the “Custom” option and selects a capture date near the beginning or end of the academic year so that websites will be captured on an annual basis. This strategy is particularly effective with university websites because the vast majority tend to accumulate new content at the top/front of web pages while older content is pushed further down the page or placed in an ‘archive’ section instead of being deleted. Exceptions to this scheduling rule include the President and Provost (captured quarterly), the athletic department (monthly), class schedules (biannually, at the start of each semester), and inactive pages (captured once). Additional exceptions may be made for sites that contain highly significant content that is modified and/or removed on a frequent basis.

As the foregoing discussion reveals, the accurate and effective configuration of crawl settings must be based on the archivist’s appraisal of content and understanding of the target site’s structure. The failure to consider these factors may lead to a capture that, on the one hand, is narrowly circumscribed and incomplete or, on the other, is unnecessarily broad and filled with superfluous information.

Contextualization of Content: Description, Metadata, and Tags

After customizing the settings for the WAS Heritrix web crawler, archivists supply each website with a description, metadata, and tags to help contextualize the preserved website and facilitate access.

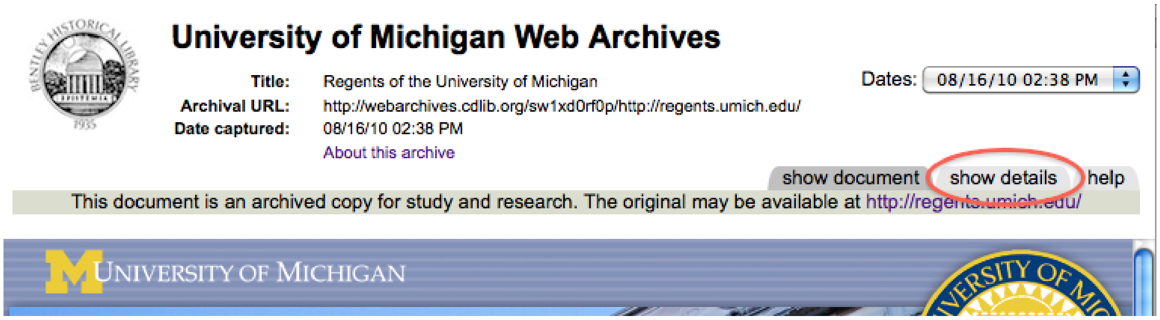
Description: UARP contextualizes each preserved website by providing an overview of the associated record creator or subject matter (see Figure 4).



Site Description: The University is governed by the Board of Regents, which consists of eight members elected at large in biennial state-wide elections. The president of the University serves as an ex officio member of the board. The Regents serve without compensation for overlapping terms of eight years. According to the Michigan Constitution of 1963, the Regents have "general supervision" of

Figure 4

To ensure accurate descriptions, archivists often use text supplied by the websites in an "About Us" or "More Information" section, if it is available. Patrons have ready access to this information from each page in the web archives under the "Show Details" tab (see Figure 5).



University of Michigan Web Archives

Title: Regents of the University of Michigan

Archival URL: http://webarchives.cdlib.org/sw1xd0rf0p/http://regents.umich.edu/

Date captured: 08/16/10 02:38 PM

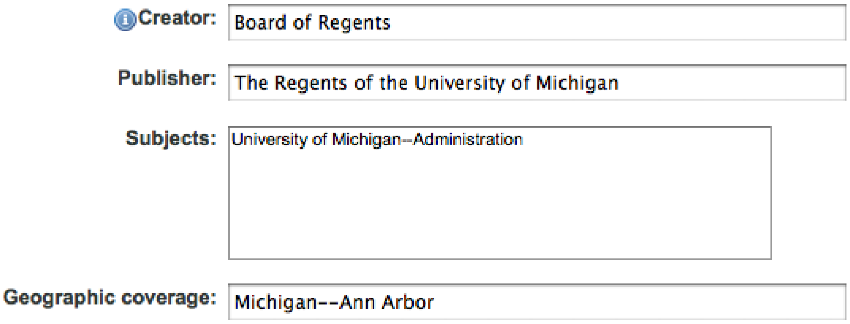
About this archive

Dates: 08/16/10 02:38 PM

This document is an archived copy for study and research. The original may be available at http://regents.umich.edu/

Figure 5

Metadata: The WAS curator interface permits archivists to enter information related to the "Creator," "Publisher," "Subjects," and "Geographic coverage" of each site (see Figure 6).



Creator: Board of Regents

Publisher: The Regents of the University of Michigan

Subjects: University of Michigan--Administration

Geographic coverage: Michigan--Ann Arbor

Figure 6

Although WAS intended these metadata fields to mirror elements in the Dublin Core Metadata Set, UARP needed to establish local definitions and conventions. After extensive discussions among archivists, the following practices were adopted:

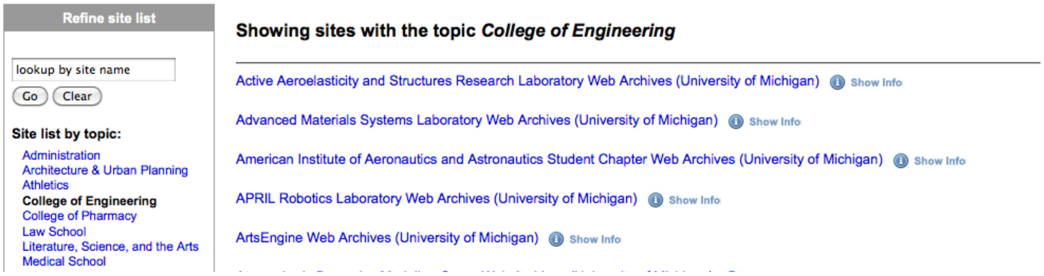
- *Creator* denotes the specific unit that generated or supplied the website's content and not the individual or office that designed the page.

- *Publisher* refers to the Regents of the University of Michigan (as the entity ultimately responsible for the production and presentation of content) for all content hosted from the umich.edu domain (as well as various .org sites administered by university units). For websites administered by individuals (such as a professor's portfolio site) or consortia (such as the Detroit Urban Research Center), the "Publisher" may be the same as the creator.

- *Subjects* express Library of Congress subject authorities that correspond to MARC21 6XX fields. Due to the lack of formatting in this field (and the indeterminate status of their use within WAS), UARP does not include indicators and subfield codes but instead simply enters the primary and secondary descriptors and separates them with double hyphens.

- *Geographic coverage* identifies where the activities described in the site took place. This field usually refers to Ann Arbor, although some research and consortial sites have a broader range. Archivists again utilized MARC21 conventions so that the main geographic entry is followed by the subdivision but did not (for reasons stated above) include the field codes themselves.

<u>Tags</u>: WAS also allows archivists to "tag" archived websites with one or more subject terms to facilitate user access to content. Archivists created tags that identified significant groups of interrelated content; the "College of Engineering" tag thus identifies all archived websites that are created, maintained, or associated with this particular college. When browsing the University of Michigan Web Archives' Site List, a user may select a tag to review only those archived websites associated with a specific subject (see Figure 7).



Site list by topic:

Administration
Architecture & Urban Planning
Athletics
College of Engineering
College of Pharmacy
Law School
Literature, Science, and the Arts
Medical School

Showing sites with the topic *College of Engineering*

Active Aeroelasticity and Structures Research Laboratory Web Archives (University of Michigan)

Advanced Materials Systems Laboratory Web Archives (University of Michigan)

American Institute of Aeronautics and Astronautics Student Chapter Web Archives (University of Michigan)

APRIL Robotics Laboratory Web Archives (University of Michigan)

ArtsEngine Web Archives (University of Michigan)

Figure 7

As of March 2011, UARP has created 24 tags that include the abbreviated names of the university's 19 schools and colleges, "Administration," "Athletics," "Faculty," "News & Events," and "Museums, Libraries, & Culture," and "Student Organizations." Additional tags will be created as the collections continue to expand and as archivists get feedback from users. Archivists may modify or delete tags and all sites that are denoted by the affected tags will inherit these changes. Many sites in the web archives do not have tags because they do not fit into these established categories and tagging is only effective when there are a significant number (i.e. five or more) of related sites. Archivists may, however, add tags to existing archived websites should the need arise.

With the inclusion of description, metadata, and tags, the archivist may initiate the web crawl and successfully conclude the workflow for content acquisition. Archivists regularly meet to discuss the status of the web archives and review difficult appraisal and content management decisions.

Quality Assurance

Quality assurance is a labor-intensive process that is made more challenging by the complex and evolving nature of website design. Due to the size of the University of Michigan Web Archives,[3] archivists are unable to review each preserved website to gauge the relative success of its capture. While a special effort is made to examine captures of the highest-profile sites (such as the President's), UARP has adopted several strategies to deal with quality assurance in a more effective and efficient manner. Archivists identify website features during the appraisal process that may cause technical issues during the crawl and then follow a basic workflow to provide general quality assurance for captured content.

Known Technical Issues in Website Preservation

Technical limitations inherent to website preservation make it difficult to capture the exact form, functionality, and content of websites as they are experienced on the 'live' web. Even if a web crawler has successfully captured a site, material may not display or function properly for end-users. The following types of content are known to be particularly difficult to capture and/or display:

- Dynamic scripts or applications: Common web design elements such as JavaScript and Adobe Flash may be very problematic for website preservation. Developers may employ relative links to JavaScript files that are difficult to capture and the underlying code may need to contact the originating server (which wreaks havoc with the display of archived sites when, for example, a script causes the archived page to be automatically updated to the 'live' version). Likewise, sites designed with Flash are difficult to capture since they may require interaction with the original host to display and/or be navigated.

- Streaming media and embedded players: Streaming audio and video content often relies on links to a live server and are therefore difficult to capture. The content itself may require plug-ins to be properly displayed (i.e. 'streamed') to archival users. Embedded players require the source code of the application to render and function properly.

- Social media sites: As mentioned above, sites such as Face book, Twitter, and YouTube pose challenges based upon their structure and design (i.e. the use of JavaScript and multiple links to embed video content) as well as policies.

[3] As of March 7, 2011, the web archives is comprised of 614 sites that total 462 GB of data.

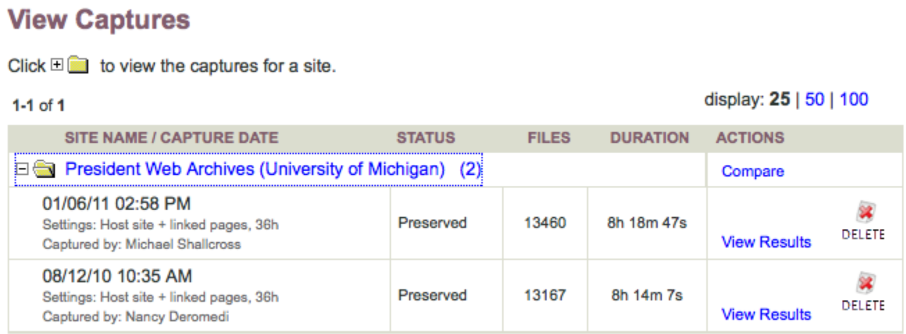
- Form or database-driven content: The web crawler is unable to capture content that requires a user to interact with the website. This category includes content accessed by drop-down menus, radio dials, password entry, or Captcha authorization. This category includes database-driven sites (such as those built with Drupal or SharePoint content management systems) that can sort and display content with user-provided fields.

- Robots.txt exclusions: For a variety of reasons, content owners may elect to exclude all or part of their websites from capture by web crawlers. These exclusions are documented in a web host's robots.txt file. UARP may request that these exclusions be revised, but will respect the content owners' decision.

Review and Assessment of Preserved Websites

Although it is not possible to manually check every page within a website to see if a crawl has been impacted by one of the above factors (or other technical issue), archivists can identify potentially problematic captures via features in the WAS interface. While additional WAS quality assurance features are in development (as of March 2011), UARP has developed a basic workflow to conduct quality assurance on archived websites.

- As an initial step, archivists check the "View Captures" screen of the WAS curatorial interface to see if the crawl was completed and if any problems may have arisen (see Figure 9).



View Captures

Click ⊞ to view the captures for a site.

1-1 of 1

display: 25 | 50 | 100

SITE NAME / CAPTURE DATE	STATUS	FILES	DURATION	ACTIONS	
⊟ President Web Archives (University of Michigan) (2)				Compare	
01/06/11 02:58 PM Settings: Host site + linked pages, 36h Captured by: Michael Shallcross	Preserved	13460	8h 18m 47s	View Results	DELETE
08/12/10 10:35 AM Settings: Host site + linked pages, 36h Captured by: Nancy Deromedi	Preserved	13167	8h 14m 7s	View Results	DELETE

Figure 9

The "Status" column may indicate a variety of outcomes: the crawl may be (a) ongoing, (b) paused or subject to technical difficulties, (c) completed and in the midst of processing, or (d) successfully preserved (in which case additional information may be available). Archivists then check the "Files" and "Duration" columns and check for those captures with relatively low or high figures for crawl duration or the number of files captured.

- The archivist reviews the individual crawl results for those captures that appear problematic. The archivist may consult the crawl log (a text file generated by the Heritrix robot) or search through the crawl results to see if significant files or web pages have been included in the capture.

- An extremely short crawl (say, four seconds) or a scant number of files may indicate an error with the seed URL or the presence of robots.txt exclusions. If the site has blocked the crawler, the archivist then contacts the webmaster to explain the purpose of the University of Michigan Web Archives and request an exception to the exclusion.

- If the crawl exceeded its maximum time, this may result in an incomplete capture or indicate the presence of a 'crawler trap.' To resolve these issues, the archivist may have to:
    - Refine or modify the seed URL
    - Enter certain directories as seeds (and adjust the crawl scope accordingly)
    - Prevent the crawler from following associated links.
    - Limit the time allotted for the capture from "Full (36 hours)" to "Brief (1 hour)"

- If the cause of an incomplete or inaccurate crawl can be identified, the archivist deletes the problematic capture, adjusts the crawler configuration, and launches a new crawl. If the problem cannot be determined, additional technical support from WAS staff may be necessary.

Intellectual Property Rights

UARP 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, UARP adheres to 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, UARP 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.

- 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 will contain the personal home pages of select faculty members and other individuals. UARP will distribute communications to these content owners to explain the purpose of the University of Michigan Web Archives, inform them of their right to opt out or suppress content, and invite questions or concerns (see Appendix A).

Access to the Collections

The University of Michigan Web Archives was created not only to preserve important online resources, but also to ensure unfettered access to these materials for patrons of the Bentley Historical Library. On one level, this commitment extends to the actual search and retrieval of content. During the capture process, the Heritrix web crawler preserves all the target site's components in their original formats inside WARC (Web ARChive) file wrappers. Patrons use the WAS implementation of the open source Wayback Machine to access and render the archived materials. Descriptions, metadata, and tags provided by archivists contextualize content and facilitate navigation and browsing. Information on search strategies for the web archives may be found on its Help page.

Archivists have also developed various access points to encourage patrons' use of the University of Michigan Web Archives. Although the collections will be indexed by Internet search engines and thus be made accessible to a global audience, UARP recognizes that preserved websites will benefit from archival mediation and description, as do other resources at the Bentley Historical Library. To this end, archivists have provided the following richly contextualized access points in addition to the main portal of the web archives:

- **UARP Web Pages:** UARP hosts general guidelines as well as an access page on the Bentley Historical Library website. The former resource provides an overview of the web archives as well as information on the roles and responsibilities of various stakeholders. The latter resource offers a description of the web archives' content, access options (including search and browsing), preferred citation style, and information on the administration and development of the collection. The access page was designed to provide a thorough orientation for users and to promote UARP's transparency in its decisions regarding the policies and administration of the web archives.

- **UARP Finding Aids:** UARP will place standardized language into the finding aids of its most prominent record groups, those of Board of Regents, President, Provost, and the 19 schools and colleges. The archived websites will be included as a new series (or, in some cases, as a continuation of an existing "Archived Website" series) and the date range will indicate that the captures commenced in 2010 and are ongoing. The EAD version of the finding aid will furthermore have a direct link to the persistent URL of the archived site. The series' scope and content note will also indicate the overall purpose and function of the archived site and that captures will continue on a regular basis.

- **MARC Catalog Records:** As of March 2011, UARP plans to include records for each site in the University of Michigan Library's online catalog, Mirlyn, and to provide direct links from there to the archived content. As such, the records will not be highly detailed; basic information related to site names and creators will suffice to give end-users a toehold to access content.

  Due to the large number of records that would have to be created, the process will need to be automated. To accomplish this goal, archivists are working closely with WAS technicians and administrators of the MLibrary's Aleph ILS. Although details are still being resolved (as of March 2011), UARP is experimenting with the batch creation of records using MarcEdit, a free Windows-based MARC editing tool. Archivists will also develop protocols to prevent the creation of duplicate records for content.

Designated Community of Users

The University of Michigan Web Archives will be of interest and value to researchers, alumni, and members of the community at large (including students, administrators, faculty, staff, and area residents) as well as to units of origin that may refer to earlier versions of their sites for ready reference or administrative purposes. UARP welcomes comments, questions, and suggestions in regards to these collections. Please feel free to contact Nancy Deromedi or Michael Shallcross at bhlwebarchive@umich.edu.

An Ongoing Project: the University of Michigan Web Archives

Archivists are diligent in their efforts to identify new, modified, or decommissioned websites so that the archives can best reflect the scope and fluidity of the University of Michigan web domain. This vigilance has required archivists to monitor University of Michigan news reports and cultivate relationships with webmasters and other personnel to learn of pending changes to the institution's web presence. At the same time, UARP actively manages archived content within the collection. In addition to quality assurance reviews and the deletion of incomplete or problematic crawls, archivists will remove archived websites (or parts thereof) at the request of content owners. Given the dynamic nature of this enterprise, the University of Michigan Web Archives may be viewed as an ongoing project.

Version History

UARP will review this collection development policy and methodology on an annual basis and make updates to reflect changes to the Web Archiving Service, the University of Michigan web domain, archival best practices, or other relevant issues.

Version No.	Date	Reviewed By	Amendments
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 for description; 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.