

# **Bentley Historical Library Web Archives: Methodology for the Acquisition of Content**

Nancy Deromedi and Michael Shallcross  
Digital Curation

Version 2.0 (August 2, 2011)

---

## **Table of Contents**

- Introduction ..... 2**
- Identification of Content ..... 3**
- Configuration of Web Crawler Settings..... 5**
- Contextualization of Content ..... 7**
  - Description: ..... 7
  - Metadata ..... 8
  - Tags..... 9
- Version History..... 10**

## **Introduction**

The Bentley Historical Library's Digital Curation Division has developed a methodology and workflow for the acquisition of content. These procedures are based on the available features of the California Digital Library (CDL)'s Web Archiving Service (WAS) as well as standard archival practices (such as appraisal and description). This document provides an overview of the Bentley Historical Library's methodology for website preservation.

The actual process of website preservation may be broken down into three main steps:

1. Identification of the crawl target
2. Configuration of the crawler settings
3. Contextualization of content

Guided by collecting priorities, surveys of relevant websites, and knowledge of significant individuals and organizations, archivists identify potential targets for preservation. By standardizing the configuration of web crawler settings and addition of metadata and descriptions, archivists are able to ensure that websites are preserved in a manner that is consistent, efficient, and cost-effective.

Given the fast pace of change in web archiving technology and ongoing development of features and functionalities in WAS, this methodology document will be reviewed on an annual basis and revised accordingly.

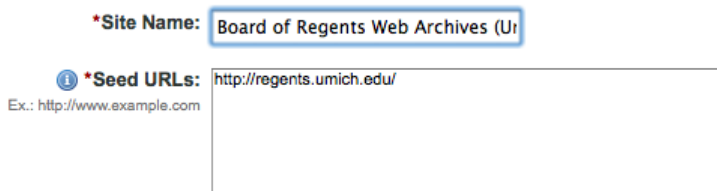
## **Identification of Content**

The Bentley Historical Library employs the Heritrix web crawler (also known as a spider or robot) to copy and preserve websites. As a subscriber to WAS, the Bentley Library relies upon an implementation of Heritrix specially configured and maintained by the CDL. A web crawler is an 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 requires the archivist to specify one or more seed URLs from which the web crawling application will preserve 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 may be hosted from multiple domains. For example, the University of Michigan's 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/](https://secure.rackham.umich.edu/academic_information/programs/). To completely capture the Rackham School'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 University of Michigan's 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 curatorial interface (see Figure 1).



The image shows a screenshot of a web form. The first field is labeled "\*Site Name:" and contains the text "Board of Regents Web Archives (U". The second field is labeled "\*Seed URLs:" and contains the text "http://regents.umich.edu/". Below the second field is an example text "Ex.: http://www.example.com".

Figure 1

The Bentley Historical Library 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 creator (i.e. the individual or organization responsible for the intellectual content of the site). The library follows the best practices for collection titles as established by Describing Archives: a Content

Standard (DACS); to ensure that the nature of the collections are clear, archivists supply “Web Archives” in the final title. The University Archives and Records Program (UARP) furthermore includes “University of Michigan” in titles to highlight the provenance of websites. Complete names for sites in the University of Michigan Web Archives 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 curatorial 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).

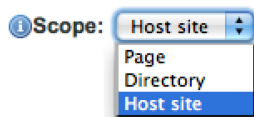


Figure 2

To thoroughly capture target websites, the Bentley Historical Library generally uses the “Host site” setting, unless the target is a single directory located on a more extensive host or a specific page.

**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.

To avoid preserving extraneous content, the Bentley Historical Library by default does not capture linked pages. Archivists will only capture linked pages if it required as a result of website design or if it is necessary to capture contextual information for a high priority web crawl.

**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, the Bentley Historical Library 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:    
  
 Weekly  
 Monthly  
 Custom

Day of the month:

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

Figure 3

Archivists generally choose the “Custom” option and select an annual capture date, being mindful of important events/dates that might result in updates to the target site. (For instance, University of Michigan sites are captured near the beginning or end of the academic year.) This strategy is particularly effective with ‘aggregative’ websites in which new content is placed at the top/front of pages while older information is moved further down the page or placed in an ‘archive’ section. For high priority targets (such as the University of Michigan Office of the President) or sites with a large turnover of important content, captures may be scheduled on a more 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**

After the configuration of crawl settings, archivists supply each website with a description, metadata, and tags to help contextualize the preserved content and facilitate access.

### **Description:**

WAS provides a 'Site Description' field so that archivists may contextualize preserved websites with an overview of the creator and/or subject matter (see Figure 4).

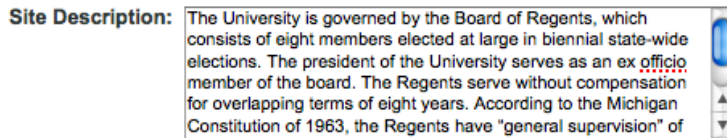


Figure 4

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).



Figure 5

## Metadata

The WAS curatorial interface permits archivists to enter information related to the “Creator,” “Publisher,” “Subjects,” and “Geographic coverage” of each site (see Figure 6).



The image shows a screenshot of a web form with four input fields. The first field is labeled 'Creator:' with an information icon and contains the text 'Board of Regents'. The second field is labeled 'Publisher:' and contains 'The Regents of the University of Michigan'. The third field is labeled 'Subjects:' and contains 'University of Michigan--Administration'. The fourth field is labeled 'Geographic coverage:' and contains 'Michigan--Ann Arbor'.

Figure 6

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

- *Creator* denotes the individual or organization that generated or supplied the website’s intellectual content (and not merely the web designer who created the page).
- *Publisher* refers to the entity ultimately responsible for the production and presentation of content. Although the publisher may often be identical to the creator, the Regents of the University of Michigan are recognized as the collective publisher for all sites affiliated with the university. Similar situations may arise with other archived websites.
- *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), the Bentley Historical Library 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. 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.



## Tags

WAS also allows archivists to “tag” archived websites with one or more subject terms to facilitate user access to content. Archivists have therefore created tags that identified significant groups of interrelated content: for example, the “College of Engineering” tag identifies all archived websites that are created, maintained, or associated with this particular college. When browsing the site list of a public archives, a user may select a tag to review only those archived websites associated with a specific subject (see Figure 7).



Figure 7

Tags are currently employed in both the Bentley Historical Library Web Archives; additional ones will be created as the collections continue to expand and as archivists receive feedback from users. Management features in the curatorial interface allow archivists to modify or delete tags; all sites that are denoted by the affected tags will inherit these changes (see Figure 8).

## Manage Tags for Michigan Historical Collections Web Archives



Figure 8

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.

## **Version History**

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

<b>Version No.</b>	<b>Date</b>	<b>Reviewed By</b>	<b>Amendments</b>
2.0	August 2, 2011	Michael Shallcross	Consolidation of methodologies 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	General editing.
0.9	March 17, 2011	Brian Williams, Associate Archivist	General editing.
0.8	March 11, 2011	Nancy Deromedi, Associate Archivist	Clarification of procedures for description; general editing.
0.1-0.7	March 8, 2011	Michael Shallcross, Assistant Archivist	Original drafts