

---

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

---

Version 4.0 (September 11, 2014)

---

## **Table of Contents**

<b>Introduction .....</b>	<b>2</b>
<b>Identification of Crawl Target.....</b>	<b>3</b>
<b>Configuration of Web Crawler Settings.....</b>	<b>4</b>
<b>Contextualization of Content .....</b>	<b>7</b>
<b>Metadata .....</b>	<b>7</b>
<b>Tags.....</b>	<b>9</b>
<b>Initiating the Crawl .....</b>	<b>9</b>
<b>Appendix A: Basic University of Michigan Topical Subjects.....</b>	<b>10</b>
<b>Version History.....</b>	<b>11</b>

## **Introduction**

The Bentley Historical Library's 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 four main steps:

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

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 periodically reviewed and revised accordingly.

## **Identification of Crawl Target**

The appraisal and selection of content to be included in the Bentley Historical Library Web Archives are guided by the Library's Web Archives Collection Development Policy (<http://hdl.handle.net/2027.42/94163>), which is reviewed on a regular, ongoing basis.

The Bentley Historical Library will contact individuals (including U-M faculty and student groups), organizations, and voluntary associations to inform them of any relevant web archiving activity and their right to opt out of the crawls or have content suppressed from public view.

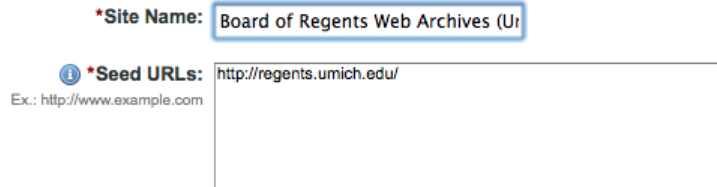
In capturing websites, the Bentley Historical Library employs the CDL's implementation of the Heritrix web crawler (also known as a spider or robot) to copy and preserve websites. A web crawler is an application that starts at a specified URL and then methodically follows hyperlinks to copy html pages and associated files (images, audio files, style sheets, etc.) as well as the website's underlying structure.

The initiation of a web capture requires the archivist to specify one or more "seed" URLs that will be used by the web crawler to 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.

## Configuration of Web Crawler Settings

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.

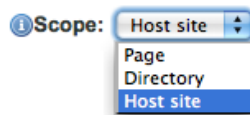


The screenshot shows two input fields. The first is labeled '\*Site Name:' and contains the text 'Board of Regents Web Archives (U'. The second is labeled '\*Seed URLs:' and contains the text 'http://regents.umich.edu/'. Below the second field is an example: 'Ex.: http://www.example.com'.

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 is clear, archivists supply “Web Archives” or “Archived Blog” in the final title. University sites furthermore include “University of Michigan” in their 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 (University of Michigan) Web Archives.”

Once the seed URL(s) and title have been entered, the WAS curatorial interface enables staff members to adjust the following crawl settings:

- **Scope:** defines how much of the site will be captured. The archivist may elect to capture the entire *host* (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>).



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.

Capture Linked Pages:  Yes  No

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 captures linked pages. Archivists will only capture linked pages if required by the structure of a website or to capture contextual information for a high priority web crawl.

- **Honor robots.txt:** The Bentley Historical Library will generally respect all robots.txt exclusions except in cases where donor materials are held in third-party sites (including social media platforms such as Twitter, Facebook, Flickr, Instagram, etc.). In these cases, the Library will not honor the robots.txt exclusions so that it may capture content solely for the purposes of private study and research and in accordance with Fair Use exceptions to copyright law.

Honor robots.txt:  Yes  No \*

The Bentley Historical Library will otherwise respect and adhere to robots.txt exclusions, but may contact content owners to request changes to rules or seek permission to override them.

- **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.

Max. Time:

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.

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

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.

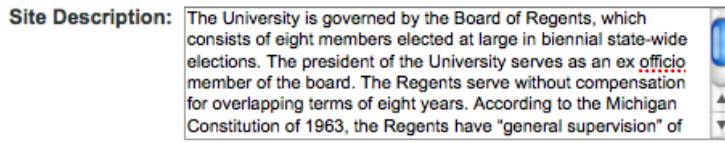
As the foregoing discussion reveals, the accurate and effective configuration of crawl settings must be based on the archivist’s appraisal of content and understanding of the target site’s structure. The failure to consider these factors may lead to a capture that, on the one hand, is narrowly circumscribed 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 descriptive metadata and tags to help contextualize the preserved content and facilitate access. These metadata fields are based on the Dublin Core Metadata Set and the the Bentley Historical Library has established the following conventions to govern their usage.

### ***Metadata***

**Site Description:** permits archivists to contextualize preserved websites with an overview of the creator and/or subject matter.



This step can be simplified by using relevant text from an existing finding aid or the "About"/"More Information" section of a website, if available. Staff should also give some indication of the nature and scope of the content found in the resource (i.e., newsletters, events, curricula, etc.).

**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).

Corporate creator:	<input type="text"/>	Corporate creator:	<input type="text" value="University of Michigan, A. Alfred Taubman College of Architect"/>
Personal creator:	<input type="text" value="Sheng, Bright"/>	Personal creator:	<input type="text"/>

If the creator is an individual his or her name should be recorded in the 'Personal Creator' field. If the creator is an organization, voluntary association, or University of Michigan unit, their name should be recorded in the 'Corporate Creator' field. A site cannot have both a personal and a corporate creator; in the event that both an individual and an organization are associated with a site, choose the most relevant entity to be recorded in the creator field. The other(s) should be mentioned in the 'Site Description' field and may also be included in the 'Subjects' field if they have made major contributions to the site or are featured prominently.

Personal or corporate names should always appear as established by the Bentley or by the Library of Congress. To find correct form of name for the creator:

- Search name in MIRLYN as 'Author.' (Remember to search personal names in inverted form: surname, forename).
- If this search brings no results, search the name as 'Subject.'
- If the name is not found in MIRLYN, repeat the search in the Library of Congress (LC) Authorities (<http://authorities.loc.gov/>).

- If you find similar entries for personal names, be sure to identify the correct one.

If the name already exists, enter it in the exact form, including dates and the complete form of the name, if applicable. If the name is not found in MIRLYN or the LC Authorities, you will need to create a new name:

- For personal creators enter surname, forename, and middle initial (i.e., Smith, John J.).
- For corporate creators, enter the complete name as it is officially used (i.e., “Michigan Environmental Council,” not “MEC.”)
- For the University of Michigan units and organizations, the proper form of name is constructed as following (including periods after name segments):
  - University of Michigan. School/College of [...].
  - University of Michigan. Office of [...].
  - University of Michigan. Department of [...].

**Publisher:** refers to the entity ultimately responsible for the production and presentation of content. For University of Michigan websites, the Regents of the University of Michigan are recognized as the collective publisher for all affiliated sites in the “edu” domain. Outside the university, similar situations may arise in which a group or organization is formally identified as being responsible for presenting information or holding copyright.

Publisher:       Publisher:

Aside from these instances, ‘Publisher not identified’ should be entered in the ‘Publisher’ field.

Publisher:

**Subjects:** provides relevant terms that denote the nature of content in the web archives and facilitate patron searches. These may include personal names, topical subjects, or geographic areas. Use Library of Congress subject authorities (<http://authorities.loc.gov/>) that correspond to MARC21 6XX fields. Due to the lack of formatting in this field, do not include specific MARC codes (i.e., 650 or 651), indicators, or subfield codes. Instead, simply enter primary and secondary descriptors and separate them with double hyphens.

Subjects:       Subjects:

Each subject entry should end with a period and a new subject should be separated from the previous with a carriage return (i.e., the “Enter” key).



If the library holds an archival collection for the creator, an existing catalog record may provide useful subject terms. For University of Michigan websites, it may also be helpful to use subjects from the list of basic terms in Appendix A. High-priority university sites (which include the Board of Regents, President, Provost, and 19 schools and colleges) should receive additional subject terms to improve the visibility of content.

**Geographic coverage:** used to identify the place of publication. Geographic location in this field should be entered in the format corresponding to MARC field 260. Examples: 'Ann Arbor, Mich.' or 'Detroit, Mich.'

Geographic coverage:  Geographic coverage:

For all University of Michigan web sites 'Ann Arbor, Mich.' should be entered. For non-university web sites, if place of publication is not known, 'Place of publication is not known' should be entered.

## 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.

The screenshot shows a web interface for refining a site list. On the left, there is a search box labeled 'lookup by site name' with 'Go' and 'Clear' buttons. Below it, a 'Site list by topic:' section lists various categories, with 'College of Engineering' highlighted. On the right, the main content area is titled 'Showing sites with the topic College of Engineering' and lists several web archives, each with a 'Show Info' link. The listed sites include: 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), and ArtsEngine Web Archives (University of Michigan).

Tags are currently employed in both the University of Michigan and Michigan Historical Collections Web Archives; additional ones will be created as the collections continue to expand and as archivists receive feedback from users.

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.

## Initiating the Crawl

With the inclusion of description, metadata, and tags, the archivist may initiate the web crawl and successfully conclude the workflow for content acquisition.

## **Appendix A: Basic University of Michigan Topical Subjects**

University of Michigan -- Administration.  
University of Michigan -- Planning.  
University of Michigan -- Finance.  
University of Michigan -- Faculty.  
University of Michigan -- Curricula.  
University of Michigan -- Curricula -- Catalogs.  
University of Michigan -- Degrees.  
University of Michigan -- Admission.  
University of Michigan -- Students.  
University of Michigan -- Students -- Social life and customs.  
University of Michigan -- Students -- Societies, etc.  
University of Michigan -- Societies, etc.  
University of Michigan -- Research.  
University of Michigan -- Buildings.  
University of Michigan -- History.  
[Subject] -- Study and teaching  
    Ex.: China -- Study and teaching.  
University of Michigan -- Sports.  
University of Michigan -- Public services.

## **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>
4.0	September 11, 2014
3.0	May 21, 2014
2.0	August 2, 2011
1.0	March 23, 2011