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

Quality Assurance for Bentley Historical Library Web Archives: Guidelines and Procedures

Version 1.0 September 21, 2011

Michael Shallcross
Nancy Deromedi

Bentley Historical Library
Digital Curation Division

Table of Contents

Introduction

Quality assurance (QA) refers to the systematic evaluation of an activity or product "to maximize the probability that minimum standards of quality are being attained."[1] In performing QA on websites preserved by the University Archives and Records Program (UARP) and Michigan Historical Collections (MHC), the Bentley Historical Library (BHL) seeks to ensure the accuracy and integrity of its web archives collections.

BHL staff involved in the preservation and QA of archived websites should have a some understanding of the design and architecture of websites (including links, embedded content, web forms, navigational menus, etc.) as well as basic knowledge of HTML, Cascading Style Sheets (CSS), JavaScript (JS), and other significant web page features. A familiarity with the curatorial interface and basic functions of the California Digital Library (CDL)'s Web Archiving Service (WAS) is also important.

During this process, a BHL QA specialist will:

- Identify incomplete, inaccurate, or unsuccessful web captures
- Determine the underlying causes or issues that led to the substandard captures. This step may require the QA specialist to:
    - Verify crawl settings
    - Review crawl reports and logs
    - Inspect the content, layout, features, and source code of the target site
- Document:
    - Any technical limitations, robots.txt exclusions, or other issues that may have prevented a faithful and accurate capture of a website.
    - Contact information for webmasters (if necessary)
    - Recommendations to delete captures or initiate new crawls

Given the inherent challenges of various content types and the technical limitations of the WAS infrastructure, it is not feasible to perfectly preserve the content, appearance, functionality, and structure of all targeted websites. Although QA may not resolve all issues with a given archived website, careful documentation will help to establish the provenance of content and record actions taken by the archives. Information gathered during QA will also enable the library to revisit problematic captures as web archiving technology continues to mature.

The CDL's release of additional quality assurance tools and reporting features for WAS in late May/early June 2011 will require the revision of these guidelines and procedures. This document will also be reviewed on an annual basis to ensure that the information and procedures contained herein are current and applicable.

[1] "Quality assurance." *Wikipedia* (May 5, 2011). Retrieved on May 6, 2011 from http://en.wikipedia.org/wiki/Quality_assurance.

Assessing Quality: Definitions and Metrics

At a basic level, the 'quality' of a web capture may be ascertained by comparing it to the appearance and functionality of the live target site. As Julien Masanès, Director of the European Archive and a board member of the Living Web Archives, writes:

> The quality of a Web archive can be defined by (a) the completeness of material (linked files) archived within a designated perimeter and (b) being able to render the original form of the site, particularly regarding navigation and interaction with the user.[2]

Ideally, the preserved website should be identical to the 'live' website in its content and behavior. However, Masanès notes that "perfect and complete archiving is unreachable" given the often-complex architecture of websites and the technical limitations of the tools with which archivists must preserve the web. Rather than aim at perfection, the BHL endeavors to capture as completely as possible those websites (or portions thereof) that have been selected and appraised according to the web archives' collection development policy[3] and methodology.[4]

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.

[2] Masanès, Julien. "Web Archiving Methods and Approaches: A Comparative Study." *Library Trends* 54.1 (2005) 72-90. Retrieved on May 9, 2011 from http://muse.jhu.edu/journals/library_trends/v054/54.1masanas.html.

[3] http://bentley.umich.edu/uarphome/webarchives/BHL_WebArchives_Policy.pdf

[4] http://bentley.umich.edu/uarphome/webarchives/BHL_WebArchives_Methodology.pdf

- 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 governing access to web crawlers and the intellectual property rights of uploaded content.

- 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 of Strategies and Methodologies

The metrics by which archivists assess the 'quality' of a web crawl and the resulting archived website range from the highly scientific to the purely observational. A number of studies in recent years[5] have involved the use of complex equations to quantify the 'completeness' (i.e. the breadth and depth of a crawl) and the 'coherence' or 'blur' (i.e. the changes to a site that occur in real time during a crawl) for archived websites. Such approaches require extensive data gathering and in-depth statistical analysis statistical analysis that are beyond the capacity of working archivists who must balance multiple tasks. As an example of this complexity, the following formula articulates the expected number of changes (*B*, or 'blur') to all pages within an archives where *P* = pages archived at time *T*, averaged through an observation interval [0, *n*Δ]:

$$B(P,T,n,\Delta) = \frac{1}{n\Delta}\sum_{i=0}^{n}\lambda_i\omega(t_i,n,\Delta).$$

Such a technique is clearly impractical for the large volume of content preserved by the Bentley Historical Library!

[5] For example, see: Ben Saad, Myriam. "Optimizing the Quality of Web Archives." (August 2010). Retrieved on May 9, 2011 from http://www-poleia.lip6.fr/~bensaadm/Rapport_Aout2010.pdf. Denev, Dmitri, et al. "SHARC: Framework for Quality-Conscious Web Archiving." (August 2009). Retrieved on May 10, 2011 from http://www.vldb.org/pvldb/2/vldb09-350.pdf. Illien, Gildas, et al. "Sketching and checking quality for web archives: a first stage report from BnF." (February 2006). Retrieved on May 8, 2011 from http://bibnum.bnf.fr/conservation/bnf-qualityforwebarchives-feb06.pdf.

The Internet Archive presents a more streamlined approach with its Archive-It service that relies upon the review of crawler reports and logs as well as the direct observation of preserved content.[6] Despite differences between Archive-It and WAS, this general strategy is amenable to the web archives created and maintained by UARP and MHC. Upcoming enhancements to WAS (scheduled for implementation by June 2011) will provide BHL QA specialists with a broader range of tools to conduct QA in a more efficient manner.

[6] Information on Archive-It's approach to QA may be found at https://webarchive.jira.com/wiki/display/ARIH/QA+Checklist.

The Bentley Historical Library's Approach to QA

Because QA is a highly labor intensive process, the Bentley Library will strive to conduct QA in an efficient and productive manner. Please note that differences in the scope and mission of the University of Michigan Web Archives and the Michigan Historical Collections Web Archives will require strategies that are unique to each collection.

University of Michigan Web Archives:
Archivists in UARP have identified high priority targets that will receive a greater degree of scrutiny than low-priority sites.

- These websites include those of the:
    - Board of Regents
    - President
    - Provost
    - 19 schools and colleges
        - Main school/college home page
        - Separate captures (may not be present for all schools/colleges):
            - Degree requirements (graduate and undergraduate)
            - Course catalogs/listings
            - Faculty guide
            - Student handbook
    - Athletic department
    - News Service
- Captures of these sites will receive a thorough and intensive evaluation so that they are preserved as completely and accurately as possible.

Lower priority websites include all others not mentioned explicitly above.

- These sites will undergo a more cursory QA.
- The BHL will make a best effort at capturing such sites, but QA specialists will not 'click-through' them or verify the presence of all content.

Michigan Historical Collections Web Archives:
All MHC websites have been selected to reflect the division's collecting priorities. To maintain an objective approach to collecting and describing this content, Project Administrators will not prioritize content.

Evaluation Goals for *All* preserved websites (U of M and MHC)

- Verify the successful initiation and completion of the web crawl.
    - The seed URL(s) were correctly entered and accurate.
    - There were no significant robots.txt exclusions.
    - Technical issues with the CDL's infrastructure did not result in a cancelled or stalled crawl.
    - The crawl completed before the specified time limit was reached.

- Determine the accuracy of the archived site's layout and appearance.
    - Examine the 'look and feel' of websites
    - Identify notable absences of CSS and/or image files.

Evaluation Goals for *High priority* U of M and *all* MHC websites

- Verify the completeness of the resource's most significant informational content.
    - "Informational content" refers to HTML text as well as audio, video, PDF, Office documents, etc., as opposed to the appearance or structure of a site.
    - "Significant" content refers to information that is essential to the conduct or documentation of the records creators' essential functions. As an example from the U of M Web Archives, web captures for academic departments should include:
        - Course catalogs
        - Curricula
        - Degree requirements
        - Newsletters and publications
        - Information on deans or department heads

      Please consult the appropriate Project Administrators if you are unsure as to the most 'significant' information/aspects of a preserved website.

- Determine the depth and breadth of the capture.
    - Click through the site to see if the crawler captured the resources specified in the crawl settings (i.e. an entire host, a directory, or single page as well as linked resources).
    - Do not try and track down every single page; focus instead examining content accessible from the home page's main navigation menu.

- Evaluate the functionality/behavior of the preserved site.
    - "Functionality" refers to the behaviors and design features that contribute to the appearance, informational content, and/or user interaction with a website. Examples would include:
        - Adobe Flash files
        - JavaScript files
        - Embedded audio and video players, image viewers, page-turner applications, etc.
        - Interactive features necessary to access content or utilize components of webpage (radio dials, drop down menus, web forms, etc.).
    - This objective may require the QA specialist to visit the 'live' website.
    - Although it may be impossible to capture some site features, try to note if any significant resources are missing or disabled in the archived version.

QA Procedures for Bentley Historical Library Web Archives

1. For each site, use the QA Spreadsheet to record:
    a. Your initials
    b. The date on which QA was conducted
    c. The number of captures currently held for the site
    d. The date range of the captures (may be a single date).

2. From the "Manage Sites" screen of the WAS curatorial interface, click on the site name to access the "Site Summary." (You may choose to right-click and open in a new tab.)
    a. Capture Settings
        i. Verify that the site name (i.e. "Department of Chemistry Web Archives (University of Michigan)") adheres to BHL conventions.
            1. BHL conventions for site titles may be found in the document: "Bentley Historical Library Web Archives: Methodology for the Acquisition of Content" (pp. 3-4).
            2. Modify site names as needed in step 8 (being sure to respect the original site's name, if possible).
        ii. Check if "linked pages" are being captured:
            1. For U of M content:
                a. Only "high priority" sites should include the capture of linked pages.
                b. For all other sites, linked pages should not be captured to avoid an excessive amount of content in the web archives.
            2. For MHC content, the QA specialist may need to verify if linked content should be captured. (See later steps.)
    b. Scheduling
        i. For U of M:

1. Only “high priority” sites will be scheduled for more than one capture a year (see list on p. 7).

2. Campus event websites (including the Arts Portal, Online Event calendar, etc.) and the Gateway may also be captured more frequently.

3. All other sites should only be captured on an annual basis.

ii. For MHC:

1. If there are multiple captures scheduled, conduct crawl comparisons to see if these are necessary.

2. Check with Project Administrators before adjusting schedule.

c. Descriptive Data

i. Check Description, Creator, Publisher, Subjects, and Geographic coverage elements to ensure that they follow BHL conventions.

1. BHL conventions for metadata entry may be found in the document: “Bentley Historical Library Web Archives: Methodology for the Acquisition of Content” (pp. 7-8).

2. Edit metadata as needed in step 8.

ii. Check “Site Tags” (on right hand side)to see if the archived website could be grouped with other relevant subjects. (This determination may require the QA Specialist to view the archived page.)

1. A full listing of tags for a specific project is available under the “Administration > Mange Tags” menu item.

2. BHL conventions for tagging may be found in the document: “Bentley Historical Library Web Archives: Methodology for the Acquisition of Content” (p. 9).

3. Only Project Administrators may add new tags to the current list. Please inform the appropriate administrator

if you believe that an additional tag (or tags) may be necessary.

d. Capture History
   i. Check general the following for potential issues:
      1. “Status”: may reveal ongoing technical issues
      2. “Files”: could be problematic if extremely low or high
      3. “Duration”: could be problematic if extremely short or timed out

3. Click “View Results” link to access the Crawl Overview
   a. Check seed URL(s) for redirects
   b. In case of an extremely small number of files or short duration, check “Robot Exclusions” statistics to see if the crawler was blocked
   c. In case of an extremely large number of files or in the event that the crawler exceeded the 36 hour duration, check the “Hosts Report” to see how many URLs are remaining for the main seed URL(s)
   d. Pending the review of the archived content, it may be necessary to examine other crawl reports.

4. View archived website
   a. Verify that content is an archived resource (instead of a redirected ‘live’ web page).
   b. Verify that CSS files are present (i.e. pages are *not* text only)
   c. Click on main navigational links (depending upon crawl settings, additional content may or may not have been intended for capture).
   d. For high priority targets, click through the entire site to ensure that significant content and features have been captured.
   e. Troubleshooting:
      i. If a particular resource does not appear in the archive, conduct a search for the URL (search feature available from the main Results screen)

        ii. Viewing the source code of the original page will help to identify web design features or resources that may not have been captured.
        iii. Check live version of archived site (if available) to compare appearance of archived version.
        iv. Check reports/crawl logs to understand issues with the crawl.
            1. Look up specific URLs to see if they were captured.
            2. Trace progress of crawl, identify where issues arose.
    f. If (for MHC or high priority U of M sites) linked pages have been captured, determine if these contain significant information. This may require consulting the “Hosts” report (or others).

5. For sites with multiple captures:
    a. If there are more than 3 captures, only review a sample (i.e. the first, one in the middle, and the most recent).
    b. Check to see if content/features change significantly between captures. Are these frequent captures necessary? Does older content (such as course schedules or news stories) tend to stay on the site as it is updated? Will a less-frequent capture schedule allow us to preserve the same information?

6. If there is a notable problem with the crawl, identify the underlying cause and document the issue on the QA spreadsheet.
    a. Robots.txt exclusions
    b. Crawl limits (timed out)
    c. Display errors:
    d. Seed redirect
    e. ‘Live links’—rendering error
    f. Missing .css files
    g. Resources not in archive (partial)
    h. Seed issues: did not capture (at all)

    i. Crawl of unusual size
    j. Adjust crawl frequency

7. Make recommendations on the QA Spreadsheet in regards to:
    a. Back up spreadsheet while working on it
    b. The deletion of a previous crawl.
        i. Deletions should be reserved for crawls that were misdirected, erroneous, or never completed (due to robots.txt or technical issues).
        ii. In some cases, excessively large captures (i.e. greater than 4 GB) may need to be deleted to preserve space.
    c. The initiation of a new crawl.
    d. Reducing the crawl frequency of high-priority sites
    e. Communication with the contact owner if it will be necessary to request a modification of the robots.txt file or resolve another issue with the site. Try to identify and record the name/email address of the site's webmaster or main contact.

8. Edit crawl settings:
    a. "Capture Linked Pages"
        i. For U of M content:
            1. Only "high priority" sites should include the capture of linked pages.
            2. For all other sites, the capture linked pages setting should be changed to "**No"** to avoid an excessive amount of content in the web archives**.**
        ii. For MHC content, the QA specialist may need to
    b. If you determine that the web archives need to capture a smaller/wider range of content, make one (or more) of the following changes (and note in the QA Spreadsheet):
        i. Decrease/increase scope (host, directory, or page)

    ii. Decrease/increase maximum crawl time (1 or 36 hours)
    iii. Recommend the deletion/addition of additional seed URLs on the QA Spreadsheet.

  c. While crawl schedules should be accurately set at the time of capture, check with an archivist if the frequency for a site seems too low/high.

Common Issues and Problems with Web Captures

- Crawler traps: These are essentially infinite loops from which a robot is unable to escape. Online calendars are among the most common examples. The crawler will start with the present date and capture page after page of the calendar until the crawl expires without preserving more meaningful site content. The resulting capture may have a very large number of files and will likely reach the maximum time setting before finishing.

- Unexpected seed redirects: The web crawler may be unexpectedly redirected from the target seed URL and begin the crawl on a random page (sometimes completely unassociated with the original seed URL). The redirection may truncate the crawl, cause important content (such as a home page) to be missed, or may lead to a crawler trap.

- Inaccurate seed URLs: Some sites require the crawler to start at a specific web page instead of a basic domain name. For instance, the accurate capture of the U of M Law School required http://www.law.umich.edu/Pages/default.aspx to be included as a seed (instead of just http://www.law.umich.edu/). Other sites will require the crawler to start at "…/home" or "…/index.html." Failure to include accurate seeds may result in a failed crawl, unexpected redirect, or a crawler trap. The BHL QA specialist may need to visit the live website to identify the exact URL from which the crawler should begin.

- Robots.txt files: A "robots.txt" file is an Internet convention used by webmasters to prevent all or certain sections of websites from being captured by a web crawler. 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; * is a wildcard symbol that indicates 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 by leaving the field blank after "Disallow." The BHL QA specialist may need to contact the webmaster or content owner to request a modification of a site's robots.txt file.

Missing CSS files: If a site's CSS file is not included in the capture, then the archived version will be displayed in a text-only version. If this occurs, check the HTML source code to identify the CSS file's URL. It may be located in a directory that is excluded from the capture; therefore, examine the robots.txt file and, if necessary, contact the webmaster to request that the exclusions be modified. Sometimes, a CSS file may inherent features from an additional CSS file; in this case, the original CSS file must be included in the capture and may need to be added as a separate seed URL.

- Content management systems: Database-backed content management systems may be difficult to completely crawl (especially if content is not directly linked via anchor tags or is dynamically generated). In addition, the default robots.txt file for Drupal will prevent the site's CSS file from being captured. Default privacy settings on Wordpress blogs also need to be corrected so that they permit the blog to be visible to everyone, including search engines.

Version History

The Bentley Historical Library will review these QA guidelines on an annual basis and make updates to reflect changes to the Web Archiving Service, archival best practices, and other relevant issues.

Version No.	Date	Reviewed By	Amendments
1.0	September 21, 2011	Michael Shallcross, Assistant Archivist	Clarification of procedures; more detailed references to BHL policies; general editing.
0.9	August 3, 2011	Michael Shallcross, Assistant Archivist	Revised to include information related to MHC Web Archives.