



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

Version 3.0

September 9, 2013

Table of Contents

<i>Introduction</i>	2
<i>Overview of Procedures</i>	3
<i>Detailed Information on Key Objectives</i>	5
1a. Identify QA Targets.....	5
1b. Check Reports from WAS QA Tools	6
2a. Open Site and Verify Metadata.....	8
2b. Review Crawl Overview.....	10
2c. Crawl Comparison.....	12
3a. Identify and Troubleshoot Issues: Crawl Duration.....	14
3b. Identify and Troubleshoot Issues: HIGH File Count or Volume	15
3c. Identify and Troubleshoot Issues: LOW File Count or Volume.....	17
4. Manual Review of the Archived Website	18
5a. Refining the Seed URL(s)	20
5b. Limiting the Scope and Extent of the Capture.....	21
5c. Expanding the Scope and Extent of the Capture.....	22
5d. Contacting the Content Owner.....	23
5e. Deleting Problematic Crawls and Launching New Ones.....	24
6. Documenting QA Process.....	25
<i>Version History</i>	26
<i>Appendix A: Problem Content and Technical Issues in Website Preservation</i>	27
<i>Appendix B: Working with the WAS Crawl Log</i>	29
<i>Appendix C: Correspondence for Content Owners</i>	32
Request to Remove Robots.txt Exclusions	32
Request to Add Robots.txt Exclusions to Limit Crawls.....	33

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 Given the technical limitations of available tools and the Web’s general lack of uniformity, it is not feasible to perfectly preserve the content, appearance, functionality, and structure of all targeted websites. A list of common technical issues and problematic content types may be found in [Appendix A](#) of the present document.

In performing QA on websites preserved by the University Archives and Records Program (UARP) and Michigan Historical Collections (MHC), the Bentley Historical Library therefore seeks to identify major issues with captures and document remedial actions and communications with content owners. Archivists and graduate students will complete the following actions for each version of a website captured by the Bentley Library:

- Confirm the successful initiation and completion of the capture.
- Verify correctness of capture settings and metadata.
- Determine if any “highly-significant content” is missing from the capture.
 - “Highly-significant content” may be defined as any resource(s) essential for the interpretation of the website or for understanding key functions of the creator.
 - It will not be necessary to note the absence of individual images, audio, video, text, etc. unless that content is critical to the research value of the site.
- Attempt to resolve any outstanding issues by changing capture settings, contacting the content owner, and/or recapturing the site.
- Document outstanding issues as well as issues taken in the WAS Curatorial Notes field.

UARP has designated a number of “high priority” sites for which complete captures are especially important and to which references are made throughout this document. These high priority sites include those of the president, provost, the 19 schools and colleges at the Ann Arbor campus, the Athletic Department, and News Services.

Please note that this document will be periodically reviewed and revised to ensure it reflects the Bentley Historical Library’s current policies and practices.

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

Overview of Procedures

The following is intended to identify key objectives of the Bentley Library's QA process for web archives and is not meant to be a step-by-step workflow.

Click on the links below to be taken to more detailed information on a specific objective. Links at the top of each page can bring you back to this basic overview.

1. Preliminary Steps:

- a. [Identify QA targets](#):
 - i. If returning to a previously initiated round of QA, open previously saved list and continue with objective 1b.
 - ii. If initiating new round of QA, filter sites in the 'Manage Sites' screen of the WAS curatorial interface to produce a list of all content captured since the last QA session began.
 - iii. Save filter results as a complete Web page to NAS location.
- b. [Check reports from WAS QA Tools](#). Filter the report of each tool based upon the time period established in objective 1a:
 - i. Captures with fewer than 10 files
 - ii. Failed captures
 - iii. Captures reached time limit
 - iv. Redirected seed URLs

Complete QA for each site identified in the above tools, paying particular attention to the possible issues identified by WAS.

2. Review Metadata and Crawl Statistics:

- a. [Open the Site Summary in a new tab and verify site metadata](#). Review information in Curatorial Note, if present.
- b. [Review crawl overview: status, statistics and reports](#).
- c. Check file count over time in the *Capture History* tab; if count is static, [conduct Crawl Comparison](#) to determine if site should be deactivated.

3. Identify Potential Issues:

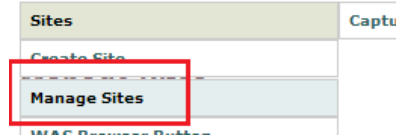
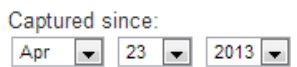

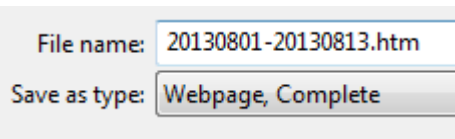
- a. [Crawl duration](#).
- b. [High crawl volume and/or file count](#).
- c. [Low crawl volume and/or file count](#).

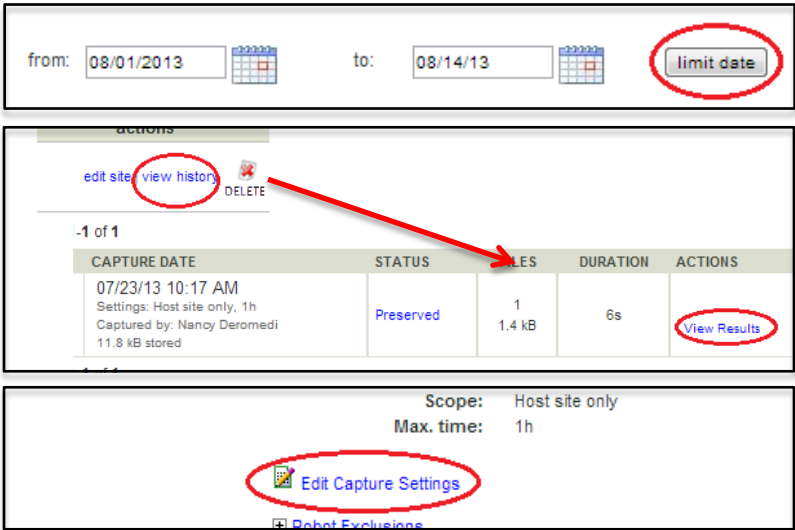
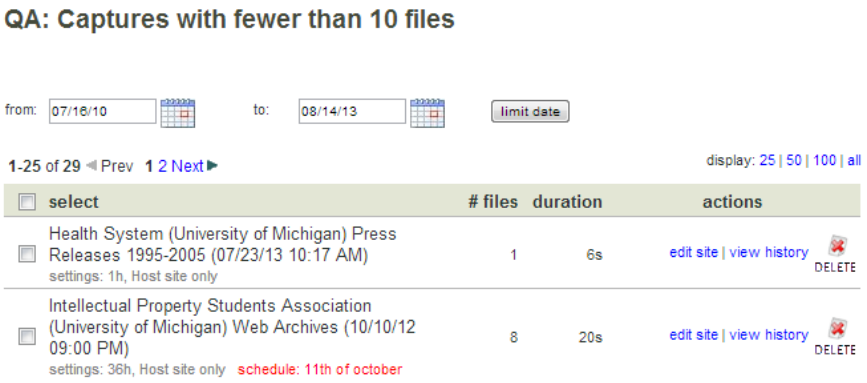
4. [Conduct Manual Review](#): check .CSS file and “highly-significant” content (i.e., resources essential for the interpretation of the website or for understanding key functions of the creator).



















5. Resolve Crawl Issues:
 - a. [Refine the Seed URL\(s\)](#)
 - b. [Limit the Scope and Extent](#) of the Capture
 - c. [Expand the Scope and Extent](#) of the Capture
 - d. [Contact the Content Owner](#) to Revise Robots.txt Exclusions
 - e. [Delete and recapture](#): if the problematic capture is unusable or is devoid of research/historical value, delete the capture and initiate new crawl.

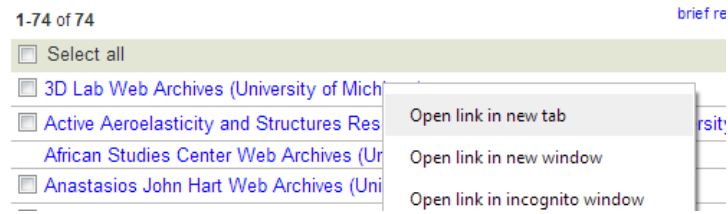
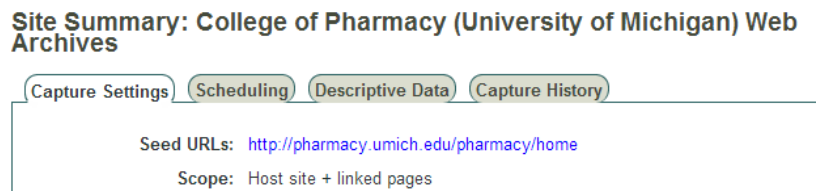
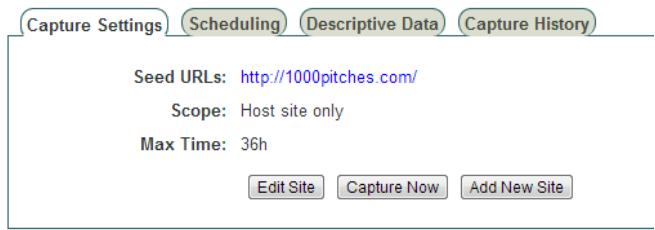
6. [Document QA process](#) in the WAS Curatorial Note field:
 - a. Date QA was conducted (all that is required if no issues were identified)
 - b. Outstanding issues such as blocked .CSS file or highly significant content that is missing. (Include the date to indicate when the issue was current.)
 - c. Actions taken (i.e. changes to crawl settings such as the seed URL, duration, or scope—with date action occurred) and brief summary of any correspondence with content owners (including the date of correspondence and name/email address of contact)

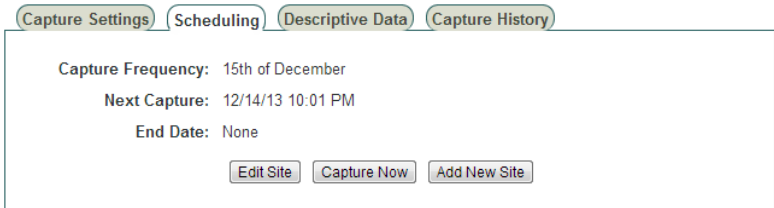
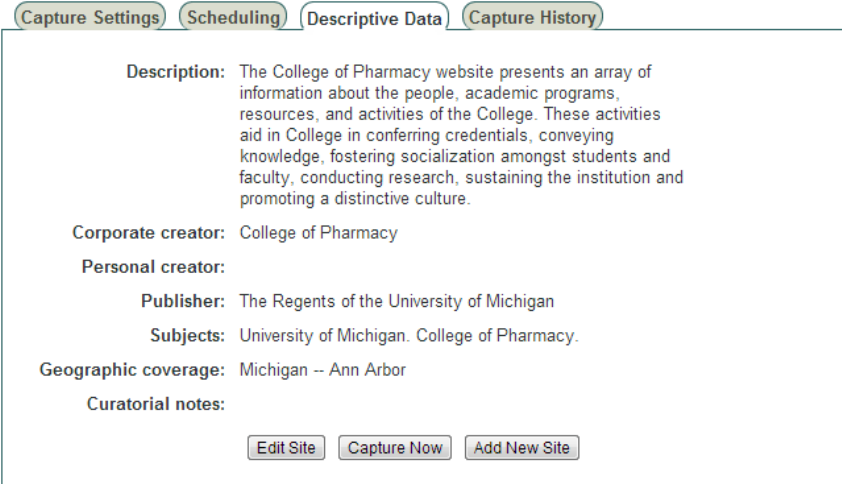
Detailed Information on Key Objectives

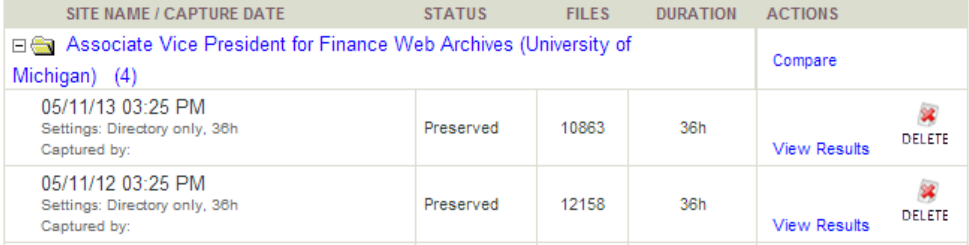

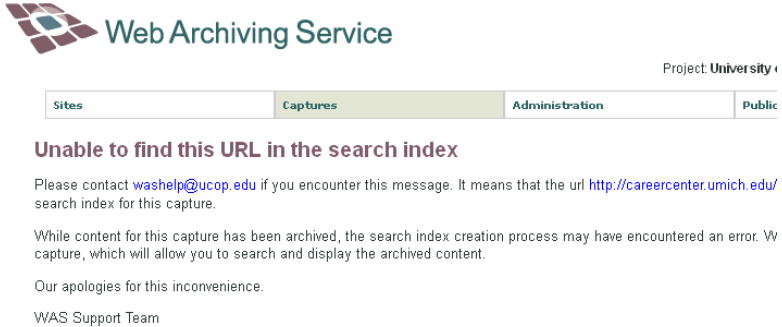
1a. Identify QA Targets	
Procedure	Illustration
<p>1. If you are completing a previously initiated round of QA, open the appropriate list file and proceed to objective 1b, if necessary.</p>	
<p>2. If this is a new round of QA, go to <i>Sites > Manage Sites</i> in the top navigation menu</p>	 <p>A screenshot of a web application's top navigation menu. The 'Manage Sites' button is highlighted with a red rectangular box.</p>
<p>3. Filter the site list to include only those captured since the previous round of QA was initiated (determine date by checking appropriate project folder at \\bhl_nas_1\digital-curation\WebArchive-QA).</p>	 <p>A screenshot of a date filter interface. The text 'Captured since:' is followed by three dropdown menus showing 'Apr', '23', and '2013'.</p>
<p>4. Change the display settings for the sorted list so that “all” records are displayed as “site names.”</p>	 <p>A screenshot of display and sort settings. The 'display' dropdown is set to 'all', and the 'sort by' dropdown is set to 'site name'. Both 'all' and 'site name' are highlighted with red boxes.</p>
<p>5. Use your browser to save the results as a complete webpage to the appropriate project folder at (\\bhl_nas_1\digital-curation\WebArchive-QA).</p> <p>Include the date range of results (starting date to current date) in the filename. Example: <i>20130801-20130813.htm</i></p>	 <p>A screenshot of a browser's 'Save as' dialog box. The 'File name' field contains '20130801-20130813.htm' and the 'Save as type' dropdown is set to 'Webpage, Complete'.</p>

1b. Check Reports from WAS QA Tools	
Procedure	Illustration
<p>General notes:</p> <ol style="list-style-type: none"> a. Filter results by entering appropriate dates and clicking ‘Limit date.’ b. To view content, right-click a ‘view history’ link in the report and open the site in a new tab. From here, click ‘View Results’ for the appropriate capture. c. To review metadata and crawl settings click “Edit Capture Settings” after the capture overview page opens. d. For each site listed in the reports, proceed to objective 2a to continue the QA process, being sure to address the points listed below. 	 <p>The screenshot shows the WAS QA Tools interface. At the top, there are date filters: 'from: 08/01/2013' and 'to: 08/14/13', with a 'limit date' button circled in red. Below this is a table of captures. The first row is highlighted, showing a capture from 07/23/13 at 10:17 AM, with a status of 'Preserved', 1 file (1.4 kB), and a duration of 6s. The 'View Results' link in the actions column is circled in red. A red arrow points from the 'view history' link to the 'View Results' link. Below the table, there is an 'Edit Capture Settings' button circled in red, and a 'Robot Exclusions' link.</p>
<p>1. Captures with fewer than 10 files:</p> <ol style="list-style-type: none"> a. Indicates potential error with seed URL, exclusions in robots.txt file, or technical problem with the CDL WAS crawler. b. Sites with this issue may have a low crawl volume/file count; you may need to employ troubleshooting strategies identified in objective 3. 	<p>QA: Captures with fewer than 10 files</p>  <p>The screenshot shows the 'QA: Captures with fewer than 10 files' section. It features date filters: 'from: 07/16/10' and 'to: 08/14/13', with a 'limit date' button. Below the filters, there is a table of captures. The first row is highlighted, showing a capture for 'Health System (University of Michigan) Press Releases 1995-2005 (07/23/13 10:17 AM)' with 1 file and a duration of 6s. The second row is also highlighted, showing a capture for 'Intellectual Property Students Association (University of Michigan) Web Archives (10/10/12 09:00 PM)' with 8 files and a duration of 20s. The 'View Results' link in the actions column for the second row is circled in red. A red arrow points from the 'view history' link to the 'View Results' link.</p>

1b. Check Reports from WAS QA Tools										
Procedure	Illustration									
<p>2. Failed captures:</p> <ol style="list-style-type: none"> Indicates a bad seed URL, technical issues with the target site, or malfunction with CDL WAS crawler. Verify that the seed URL is correct. Delete the new crawl, initiate a new crawl, and document actions in the curatorial note. If this is a persistent problem, the site may need to be deactivated. 	<div style="text-align: center;"> <h3>QA: Failed Captures</h3> <p>from: <input type="text" value="07/16/10"/>  to: <input type="text" value="08/14/13"/>  <input type="button" value="limit date"/></p> <p>There were no failed captures</p> </div>									
<p>3. Captures Reached Time Limit:</p> <ol style="list-style-type: none"> NOTE: this may <i>not</i> be indicative of a problem, as it only means some content remained to be captured at the end of the crawl. During manual review, be sure to check that highly significant content (such as newsletters or degree requirements) was included in the capture. Follow troubleshooting strategies identified in objective 3a (issues with crawl duration). 	<div style="text-align: center;"> <h3>QA: Captures Reached Timelimit</h3> <p>from: <input type="text" value="08/01/2013"/>  to: <input type="text" value="08/14/13"/>  <input type="button" value="limit date"/></p> <p>1-8 of 8 display: 25 50 100 all</p> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr style="background-color: #d9ead3;"> <th style="text-align: left;">select</th> <th style="text-align: left;">status</th> <th style="text-align: left;">actions</th> </tr> </thead> <tbody> <tr> <td><input type="checkbox"/> Animal Diversity (University of Michigan) Web Archives (08/02/13 01:03 PM) <small>settings: 36h, Host site only schedule: 1st of august</small></td> <td>Preserved</td> <td>edit site view history </td> </tr> <tr> <td><input type="checkbox"/> Athletic Department (University of Michigan) Web Archives (08/02/13 11:02 AM)</td> <td>Preserved</td> <td>edit site view history </td> </tr> </tbody> </table> </div>	select	status	actions	<input type="checkbox"/> Animal Diversity (University of Michigan) Web Archives (08/02/13 01:03 PM) <small>settings: 36h, Host site only schedule: 1st of august</small>	Preserved	edit site view history 	<input type="checkbox"/> Athletic Department (University of Michigan) Web Archives (08/02/13 11:02 AM)	Preserved	edit site view history 
select	status	actions								
<input type="checkbox"/> Animal Diversity (University of Michigan) Web Archives (08/02/13 01:03 PM) <small>settings: 36h, Host site only schedule: 1st of august</small>	Preserved	edit site view history 								
<input type="checkbox"/> Athletic Department (University of Michigan) Web Archives (08/02/13 11:02 AM)	Preserved	edit site view history 								
<p>4. Redirected Seed URLs:</p> <ol style="list-style-type: none"> Indicates that the target site redirected the CDL WAS crawler to a new URL and that at least one of the current seeds needs to be updated. Verify the seed URL identified by WAS and edit the crawl settings as needed. In conducting a manual review of the site, be sure to verify that significant target content has been captured. If important content was not captured, it may be necessary to recrawl (with correct seed). 	<div style="text-align: center;"> <h3>QA: Redirected Seed URLs</h3> <p>from: <input type="text" value="08/01/2013"/>  to: <input type="text" value="08/14/13"/>  <input type="button" value="limit date"/></p> <p>1-2 of 2 display: 25 50 100 all</p> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr style="background-color: #d9ead3;"> <th style="text-align: left;">select</th> <th style="text-align: left;">status</th> <th style="text-align: left;">actions</th> </tr> </thead> <tbody> <tr> <td><input type="checkbox"/> Office of the Ombuds (University of Michigan) Web Archives (08/01/13 05:48 PM) <small>settings: 36h, Directory only schedule: 1st of august</small></td> <td>Preserved</td> <td>edit site view history </td> </tr> <tr> <td><input type="checkbox"/> Program on Intergroup Relations (University of Michigan)</td> <td></td> <td></td> </tr> </tbody> </table> </div>	select	status	actions	<input type="checkbox"/> Office of the Ombuds (University of Michigan) Web Archives (08/01/13 05:48 PM) <small>settings: 36h, Directory only schedule: 1st of august</small>	Preserved	edit site view history 	<input type="checkbox"/> Program on Intergroup Relations (University of Michigan)		
select	status	actions								
<input type="checkbox"/> Office of the Ombuds (University of Michigan) Web Archives (08/01/13 05:48 PM) <small>settings: 36h, Directory only schedule: 1st of august</small>	Preserved	edit site view history 								
<input type="checkbox"/> Program on Intergroup Relations (University of Michigan)										

2a. Open Site and Verify Metadata	
Procedure	Illustration
<p>1. Right click on a site name in the saved HTML file and open the Capture History in a new tab.</p> <p>If coming in via a WAS QA Tool report, click the “Edit Capture Settings” link to view all site metadata and crawl settings; otherwise, you may review metadata and crawl settings from the Site Summary screen.</p>	
<p>2. Site name should reflect established BHL conventions and must include “Web Archives”; U-M units must also include “(University of Michigan).”</p> <p>NOTE: If an organization/office (or its site) is renamed, update the following WAS metadata:</p> <ul style="list-style-type: none"> • Site name • Creator and publisher (if necessary) • Site description (note date of the change and include any additional information about the successor organization) • Subject (add the old name as a subject term) <p>If the organization or unit has morphed into a new functional entity, then a new archived Website should be created for it.</p>	
<p>3. Capture settings: <i>Scope</i> should include “linked pages” for high priority U-M sites only. The Maximum Time should be 36 for all sites with the exception of some (like the monthly U-M Gateway capture) intended as only snapshots. Check with a project admin if you are unsure.</p> <p>If the <i>Scope</i> is set to “Directory,” make sure:</p> <ul style="list-style-type: none"> • That the seed URL does not contain a document, i.e. ../index.php • That the seed URL terminates with a trailing slash, i.e. http://bentley.umich.edu/resources/ 	

2a. Open Site and Verify Metadata	
Procedure	Illustration
<p>4. Scheduling: All sites will be captured once a year with the exception of “high priority” UARP targets, U-M campus event sites, and the U-M Gateway (with the possibility of additional exceptions).</p> <p>Make sure that scheduled capture dates are appropriate for sites (i.e. course catalogs should be captured in the fall or winter, not summer). Notify the project administrator if your QA suggests that a site should be captured more or less frequently.</p>	
<p>5. Descriptive Data: Check the Description, Creator, Publisher, Subjects, and Geographic coverage elements to ensure that they follow BHL conventions. Make sure that the appropriate creator type is selected (corporate or personal).</p> <p>You may add additional ‘tags’ that are relevant to the site (via the right-hand column), but do not create new tags without first consulting with the Project Administrator.</p> <p>This is also a great opportunity to review the Curatorial Note to learn about previous issues or actions taken with the site.</p>	

2b. Review Crawl Overview																					
Procedure	Illustration																				
<p>1. To access a specific capture, go to the <i>Capture History</i> tab and click “View Results” for the appropriate date.</p> <p>If you are in the Edit Site screen, click “Save (all tabs)” to return to the Site Summary screen.</p> <p>NOTE: If the number of files has remained static for at least three years (see illustration), the website is likely no longer active and could be a candidate for reappraisal and de-accession. See information on comparing crawls and recommending a site for reappraisal in Stage 9.</p>	 <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="background-color: #d9d9d9;">SITE NAME / CAPTURE DATE</th> <th style="background-color: #d9d9d9;">STATUS</th> <th style="background-color: #d9d9d9;">FILES</th> <th style="background-color: #d9d9d9;">DURATION</th> <th style="background-color: #d9d9d9;">ACTIONS</th> </tr> </thead> <tbody> <tr> <td colspan="5" style="background-color: #d9d9d9;"> 📁 Associate Vice President for Finance Web Archives (University of Michigan) (4) Compare </td> </tr> <tr> <td>05/11/13 03:25 PM Settings: Directory only, 36h Captured by:</td> <td>Preserved</td> <td>10863</td> <td>36h</td> <td>View Results DELETE</td> </tr> <tr> <td>05/11/12 03:25 PM Settings: Directory only, 36h Captured by:</td> <td>Preserved</td> <td>12158</td> <td>36h</td> <td>View Results DELETE</td> </tr> </tbody> </table>	SITE NAME / CAPTURE DATE	STATUS	FILES	DURATION	ACTIONS	📁 Associate Vice President for Finance Web Archives (University of Michigan) (4) Compare					05/11/13 03:25 PM Settings: Directory only, 36h Captured by:	Preserved	10863	36h	View Results DELETE	05/11/12 03:25 PM Settings: Directory only, 36h Captured by:	Preserved	12158	36h	View Results DELETE
SITE NAME / CAPTURE DATE	STATUS	FILES	DURATION	ACTIONS																	
📁 Associate Vice President for Finance Web Archives (University of Michigan) (4) Compare																					
05/11/13 03:25 PM Settings: Directory only, 36h Captured by:	Preserved	10863	36h	View Results DELETE																	
05/11/12 03:25 PM Settings: Directory only, 36h Captured by:	Preserved	12158	36h	View Results DELETE																	
<p>2. If the <i>Status</i> is “Preserved,” click the <i>View Results</i> link and begin the QA process.</p> <p>If the <i>Status</i> indicates an issue you do not recognize, contact washelp@ucop.edu for clarification (and be sure to CC bhlwebarchive@umich.edu).</p>	 <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="background-color: #d9d9d9;">SITE NAME / CAPTURE DATE</th> <th style="background-color: #d9d9d9;">STATUS</th> <th style="background-color: #d9d9d9;">FILES</th> <th style="background-color: #d9d9d9;">DURATION</th> <th style="background-color: #d9d9d9;">ACTIONS</th> </tr> </thead> <tbody> <tr> <td colspan="5" style="background-color: #d9d9d9;"> 📁 Allen Samuels Web Archives (University of Michigan) (3) Compare </td> </tr> <tr> <td>09/30/12 10:53 PM Settings: Directory only, 36h Captured by:</td> <td style="border: 2px solid red;">Preserved</td> <td>52</td> <td>2m 42s</td> <td style="border: 2px solid red;">View Results DELETE</td> </tr> </tbody> </table>	SITE NAME / CAPTURE DATE	STATUS	FILES	DURATION	ACTIONS	📁 Allen Samuels Web Archives (University of Michigan) (3) Compare					09/30/12 10:53 PM Settings: Directory only, 36h Captured by:	Preserved	52	2m 42s	View Results DELETE					
SITE NAME / CAPTURE DATE	STATUS	FILES	DURATION	ACTIONS																	
📁 Allen Samuels Web Archives (University of Michigan) (3) Compare																					
09/30/12 10:53 PM Settings: Directory only, 36h Captured by:	Preserved	52	2m 42s	View Results DELETE																	
<p>3. You may receive an error message that the Web Archiving Service was “Unable to find this URL in the search index.” Unless the <i>Capture History</i> indicated that no files were preserved, this is likely an error with WAS.</p> <p>Take a screenshot and send it with an explanation of what happened to washelp@ucop.edu (and be sure to CC bhlwebarchive@umich.edu).</p>	 <p style="text-align: right;">Project University</p> <p>Sites Captures Administration Public</p> <p>Unable to find this URL in the search index</p> <p>Please contact washelp@ucop.edu if you encounter this message. It means that the url http://careercenter.umich.edu/ search index for this capture.</p> <p>While content for this capture has been archived, the search index creation process may have encountered an error. W capture, which will allow you to search and display the archived content.</p> <p>Our apologies for this inconvenience.</p> <p>WAS Support Team</p>																				

2b. Review Crawl Overview	
Procedure	Illustration
<p>4. Examine the basic crawl statistics to see if there could be any potential issues with:</p> <ol style="list-style-type: none"> Crawl duration (extremely long or short) Large crawl volumes and/or file counts Small crawl volumes and/or file counts. 	<div style="text-align: right;"> <p>Job status: finished</p> <p>Start time: 12/14/12 10:01 PM</p> <p>Finish time: 12/15/12 02:13 PM</p> <p>Size: 966.4 MB</p> <p>Files captured: 8750</p> <p>Scope: Host site only</p> <p>Max. time: 36h</p> </div>
<p>5. Check to see if the CDL WAS crawler encountered any robots.txt exclusions. Expand <i>Robots Exclusions</i> under the 'Other Statistics' section. If a robots.txt file was found, click on it to see what exclusions, if any, might have been responsible.</p> <p>NOTE: the presence of a robots.txt file does <i>not</i> mean there will be issues with the capture. Checking will help inform your manual review and let you know how many (if any) directories or resources were excluded.</p>	<div style="border: 1px solid #ccc; padding: 10px;"> <div style="background-color: #d9d9d9; padding: 5px; margin-bottom: 10px;">Other Statistics</div> <div style="margin-bottom: 10px;"> <p>☐ Robot Exclusions</p> <p>The following robots.txt files were discovered on ser are archived in order to document the host server po the archived version of the robots.txt file.</p> <ul style="list-style-type: none"> • https://1000pitches.com/robots.txt </div> <div style="border-left: 1px solid #ccc; padding-left: 10px; margin-left: 20px;"> <p>User-agent: *</p> <p>Crawl-delay: 10</p> <p># Directories</p> <p>Disallow: /includes/</p> <p>Disallow: /misc/</p> <p>Disallow: /modules/</p> <p>Disallow: /profiles/</p> <p>Disallow: /scripts/</p> <p>Disallow: /themes/</p> </div> </div>

2c. Crawl Comparison

Procedure

1. If the number of files has remained static over three years (as displayed in the *Capture History* tab) or if based on your manual review the site no longer appears to be actively used, the website may be a candidate for deactivation.
 - a. To verify that no new content is being added, click the “Compare” link on the *Capture History* page.
 - b. A list of captures will open up; you may need to click the “Process Comparison Data” link and wait for the data to be indexed.
 - c. To compare the captures that have exhibited no change, select the earliest and the most recent by checking the boxes and then clicking “Submit.”
 - d. WAS will produce a summary report that identifies all files that are new, missing, changed, and unchanged from the earlier to the later crawl. The comparison will also provide a sortable list of all URLs in each category (new, missing, changed, and unchanged) and users may filter this list by file type (HTML, PDF, image, video, audio, Office, and compressed) or search for a specific string in the URLs.

Illustration

CAPTURE DATE	STATUS	FILES	DURATION	ACTIONS
05/01/12 02:58 AM Settings: Host site only, 36h Captured by:	Preserved	70	3m 7s	View Results
05/01/11 05:39 PM Settings: Host site + linked pages, 36h Captured by:	Preserved	70	2m 24s	View Results
07/28/10 01:15 PM Settings: Host site + linked pages, 36h Captured by: Michael Shallcross	Preserved	70	2m 49s	View Results

CAPTURE DATE	STATUS
04/20/13 10:26 PM	Process Comparison Data
04/20/12 10:26 PM	Process Comparison Data
04/21/11 12:52 PM	Available
08/17/10 02:20 PM	Available

	CAPTURE DATE	STATUS
<input checked="" type="checkbox"/>	04/20/13 10:26 PM	Available
<input checked="" type="checkbox"/>	04/20/12 10:26 PM	Process Comparison Data
<input type="checkbox"/>	04/21/11 12:52 PM	Available
<input type="checkbox"/>	08/17/10 02:20 PM	Available

Earlier: 04/13/12 08:04 PM
Scope: Host site only; Files: 264; Duration: 7h 15m 7s

Later: 04/13/13 08:04 PM
Scope: Host site only; Files: 866; Duration: 7h 15m 8s

Change Summary

Changed: 157 files (5.0% by size)
New: 709 files (93.0% by size)
Missing: 107 files (1.2% by size)
Unchanged: 0 files (0.0% by size)

display: 25 | 50 | 100
changed | new | missing | unchanged

sort by: URL | file size

http://www.taubmaninstitute.org/	old new	33.5 kB
http://www.taubmaninstitute.org/complete-disclaimer	old new	28.2 kB
http://www.taubmaninstitute.org/_ponent/search/?format=opensearch	old new	792 bytes
http://www.taubmaninstitute.org/_ch/?itemid=108&format=opensearch	old new	807 bytes
http://www.taubmaninstitute.org/_ch/?itemid=110&format=opensearch	old new	807 bytes
http://www.taubmaninstitute.org/_ch/?itemid=112&format=opensearch	old new	807 bytes
http://www.taubmaninstitute.org/_ch/?itemid=116&format=opensearch	old new	807 bytes
http://www.taubmaninstitute.org/_ch/?itemid=117&format=opensearch	old new	807 bytes
http://www.taubmaninstitute.org/_ch/?itemid=118&format=opensearch	old new	807 bytes
http://www.taubmaninstitute.org/_ch/?itemid=120&format=opensearch	old new	807 bytes
http://www.taubmaninstitute.org/_ch/?itemid=121&format=opensearch	old new	807 bytes

Click new and limit to file type PDF, then sort by file size to see potentially significant new publications released on this site.

Click missing to identify content in your archive that might no longer be available on the live web.

Detailed Guides

- Compare Results: Guide

Filetype Limit

All file types

HTML

PDF

Image





MS-Office

Video

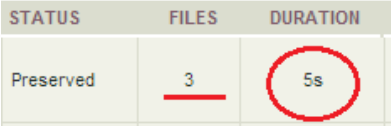

Audio

Compressed

text found anywhere in URL

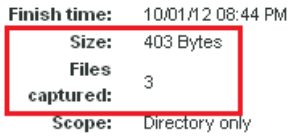

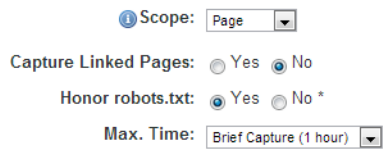
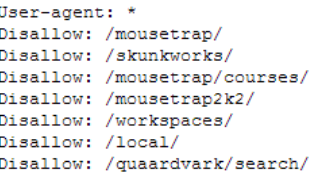

2c. Crawl Comparison	
Procedure	Illustration
<p>2. If your analysis of crawl statistics, manual review, and capture comparison suggest that the site is no longer active, it should be deactivated.</p> <p>From the Manage Sites screen, filter for the site and then click the “Deact” button.</p>	<p><input type="checkbox"/> 2013 Linguistic Institute (University of Michigan) Web Archives (1) Seed URL: http://lsa2013.lsa.umich.edu/ Tags: College of Lit., Science & Arts Status: Preserved Current settings: Host site only, 36h Last captured 2 months ago</p> <p>    EDIT CAPTURE DEACT DELETE</p>



3a. Identify and Troubleshoot Issues: Crawl Duration

Procedure	Illustration						
<p>NOTE: a long or short crawl duration does not necessarily mean that there are problems with a capture. Take the following points into consideration as you determine if highly significant has in fact been preserved or if the crawl contains a large amount of superfluous data. Always use the manual review of a site to fully determine the success or failure of a crawl.</p>							
<p>1. Short crawl duration may indicate an inaccurate seed URL, robots.txt exclusions, or technical issues with the target site or the CDL WAS crawler.</p> <p>You may need to adjust the seed URL(s), contact a webmaster about robots.txt exclusions, or employ another strategy identified in objective 3c (troubleshooting low capture volume/file count).</p>	 <table border="1"> <thead> <tr> <th>STATUS</th> <th>FILES</th> <th>DURATION</th> </tr> </thead> <tbody> <tr> <td>Preserved</td> <td>3</td> <td>5s</td> </tr> </tbody> </table>	STATUS	FILES	DURATION	Preserved	3	5s
STATUS	FILES	DURATION					
Preserved	3	5s					
<p>2. An extremely long capture (full 36 hours) could mean:</p> <ul style="list-style-type: none"> • The crawler was unable to traverse the entire site in the allotted time. • The scope of the crawl was too broad; the crawler has captured superfluous content (may be accompanied by a very large file count). • The crawler fell into a 'trap' such as an online calendar. • There were technical issues with the crawl, related to the host Web server or CDL tools. <p>Depending on the results of your manual review, you may need to adjust the seed URL, limit the crawl scope, or employ another strategy identified in objective 3b (troubleshooting high capture volume/file count).</p>	 <table border="1"> <thead> <tr> <th>FILES</th> <th>DURATION</th> <th>AC</th> </tr> </thead> <tbody> <tr> <td>307978</td> <td>36h 1m 15s</td> <td>Vie</td> </tr> </tbody> </table>	FILES	DURATION	AC	307978	36h 1m 15s	Vie
FILES	DURATION	AC					
307978	36h 1m 15s	Vie					

3b. Identify and Troubleshoot Issues: HIGH File Count or Volume	
Procedure	Illustration
<p>1. If the <i>Size</i> and/or <i>Files captured</i> are unexpectedly large, there <i>might</i> be an issue with the seed URL or crawl settings.</p> <p>To determine if extraneous content has been captured, you may do one (or more) of the following:</p>	<pre>Finish time: 04/30/13 07:54 AM Size: 16.7 GB Files captured: 117126</pre>
<p>2. Verify the accuracy of the seed URL(s). A bad seed URL may result in a wildly inaccurate crawl.</p>	<pre>Seed URL(s): http://www.lsa.umich.edu/daas</pre>
<p>3. Verify that the crawl settings (scope, duration, capture linked pages) are not too broad/inclusive for the target.</p>	<pre>Scope: Host site + linked pages Max. time: 36h</pre>
<p>4. Review the “Hosts Report” to see which hosts were included and how many URLs were captured for each. Look for:</p> <ul style="list-style-type: none"> The capture of large numbers of URLs from hosts other than the seed(s). A large number of URLs in queue (see the last column in the report), as these could indicate an incomplete crawl. 	<p>Hosts Report: Animal Diversity Web Archives (Univ <i>The text below is automatically generated by the Heritrix web crawler</i></p> <pre>[#urls] [#bytes] [host] [#robots] [#remaining] 14165 310343543 animaldiversity.ummz.umich.edu 0 39118 95 2432259 www.ucop.edu 0 0 25 1656 dns: 0 0</pre>
<p>5. Check the “Crawl Log” to see if Heritrix included out-of-scope content, especially for directory or page only crawls. The crawl log might also help you discover if the crawler got stuck in a trap. A large number of incrementally-numbered pages could indicate an issue (such as successive pages in a calendar).</p> <p>See Appendix B for more information on working with the Crawl Log.</p>	<p>Crawl Log: Animal Diversity Web Archives (University of Michigan) (01/09/13 02:50 PM) <i>The text below is automatically generated by the Heritrix web crawler</i></p> <pre>2013-01-09T19:53:29.612Z 1 73 dns:animaldiversity.ummz.umich.edu P http://animaldiversity.um 2013-01-09T19:53:31.814Z 200 472 http://animaldiversity.ummz.umich.edu/robots.txt P http://anim 2013-01-09T19:53:41.187Z 200 24437 http://animaldiversity.ummz.umich.edu/ - - text/html #002 2013 2013-01-09T19:53:41.188Z 1 63 dns:fonts.googleapis.com EP http://fonts.googleapis.com/css?fa 2013-01-09T19:53:41.193Z 1 63 dns:ajax.googleapis.com EP http://ajax.googleapis.com/ajax/lib 2013-01-09T19:53:41.857Z 302 232 https://animaldiversity.ummz.umich.edu/robots.txt LP https://a 2013-01-09T19:53:43.245Z 200 28 http://fonts.googleapis.com/robots.txt EP http://fonts.googlea 2013-01-09T19:53:43.245Z 200 4851 http://ajax.googleapis.com/robots.txt EP http://ajax.googleapi</pre>

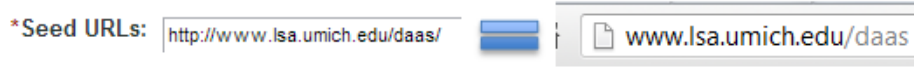


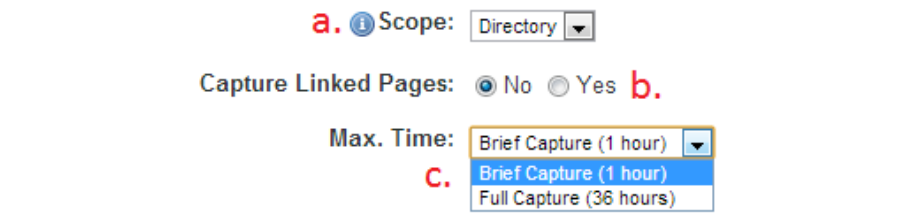

3b. Identify and Troubleshoot Issues: HIGH File Count or Volume	
Procedure	Illustration
<p>6. Check the “Mimetype Report” to see if there is an unexpectedly large number or volume of a particular media type (such as audio or video).</p>	<p>Mimetype Report: Animal Diversity Web Archives</p> <p><i>The text below is automatically generated by the Heritrix web crawler</i></p> <pre>[#urls] [#bytes] [mime-types] 8226 118268752 text/html 5807 130020377 image/jpeg 90 1521334 image/png 46 99207 image/gif 30 489121 text/css</pre>
<p>7. Use information gathered in these steps to guide your manual review of the site. Issues may lead you to reference one or more of the following:</p> <ul style="list-style-type: none"> • Objective 5a (refine the seed URL) • Objective 5b (limit the scope of the capture) • Objective 5d (contact the content owner to request removal or modification of robots.txt exclusions) • Objective 5e (delete problematic capture—only if devoid of research value—and launch new one) 	





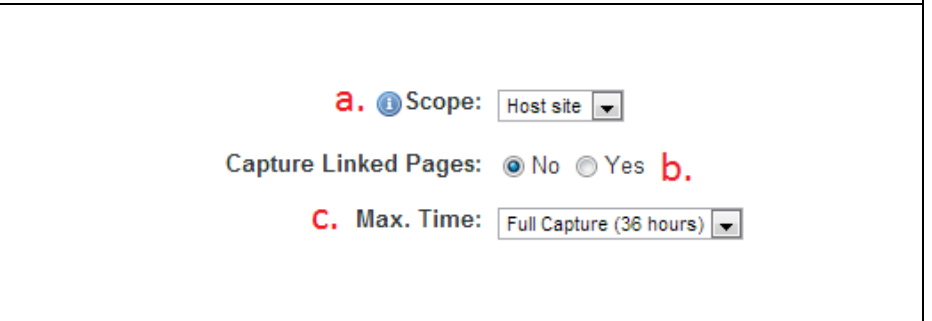



3c. Identify and Troubleshoot Issues: LOW File Count or Volume	
Procedure	Illustration
1. If the <i>Size</i> and/or <i>Files captured</i> are unexpectedly small, there <i>might</i> be an issue with the seed URL, crawl settings, or robots.txt exclusions.	
2. Compare the URL from the target site with the seed URL to verify accuracy. An incorrect seed URL may result in a failed or wildly inaccurate crawl.	
3. Check the crawl settings to determine if they are too narrow or restricted to permit the full capture of the site.	
4. Check to see if robots.txt exclusions are responsible. <ol style="list-style-type: none"> Expand <i>Robots Exclusions</i> under the 'Other Statistics' section. If a robots.txt file was found, click on it to see what exclusions, if any, might have been responsible. 	
5. Check hosts report to see if there are a large number of URLs in queue for the seed URL(s).	
6. Use information gathered in these steps to guide your manual review of the site. Issues may lead you to reference one or more of the following: <ul style="list-style-type: none"> Objective 5a (refine the seed URL) Objective 5c (expand the scope of the capture) Objective 5d (contact the content owner to request removal or modification of robots.txt exclusions) Objective 5e (delete problematic capture—only if devoid of research value—and launch new one) 	

4. Manual Review of the Archived Website	
Procedure	Illustration
<p>As you manually review archived websites, remember that it is impossible to perfectly capture the content, features, and appearance of complex digital objects hosted online. As a result of this fact:</p> <ul style="list-style-type: none"> • The Bentley Library will accept the loss or absence of non-essential images, text, hyperlinks and other resources that do not seriously impact the research/historical value of archived websites. Do not bother noting such minor errors in the WAS Custodial Notes field. • You only need to verify the capture of (a) the site's .CSS file and (b) highly-significant content (i.e. those resources that permit the site to fulfill its functions/purpose and are essential to its administrative/historical value). 	
<p>1. To begin the manual review, click on the hyperlinked seed URL on the <i>Overview</i> tab of the Results screen.</p>	
<p>2. If the site appears as 'text-only' with no layout or design features, it is likely that capture did not include the .CSS file(s). View the source code of the original website (CTRL-U) to determine the path to the .CSS file.</p> <p>If the .CSS file is stored in a separate domain or subdomain, it may be necessary to add it as a separate seed URL. (See objective 5a, refine seed URL.)</p> <p>Alternatively, check the site's robots.txt file to determine if the .CSS file is in a blocked directory. (See objective 5d, contact the content owner to request removal or modification of robots.txt exclusions.)</p>	<ul style="list-style-type: none"> • University of Michigan • School of Dentistry • Accessibility • Contact • Login 

4. Manual Review of the Archived Website	
Procedure	Illustration
<p>3. Click through the site to determine if highly-significant content has been captured. It may be helpful—or even necessary—to view the original live site.</p> <ol style="list-style-type: none"> a. This ‘significance’ will be relative to each site—examples could include include newsletters, events, policies, administrative information, etc. At a basic level, ask yourself if the information present will suffice to give researchers a general idea of the nature and scope of the individual, organization, or event. b. Remember that dynamic content (such as Javascript and Flash), streaming media, and forms will not capture. Don’t worry about or note issues with such content unless they are essential to the site’s use/function. c. For MHC and high-priority UARP sites, click through each major directory/navigation menu; navigation through non-priority UARP sites can be more cursory. 	<div style="border: 1px solid black; padding: 5px; margin-bottom: 10px;"> <pre>User-agent: * Allow: / Disallow: /news_events/events/ Disallow: /news_events/events Disallow: /news_events/week/</pre> </div> <div style="border: 1px solid black; padding: 5px;"> <p>Crawl Log: Animal Diversity Web Archives (University of Michigan) (01/09/13 02:50 PM)</p> <p><i>The text below is automatically generated by the Heritrix web crawler</i></p> <pre>2013-01-09T19:53:29.612Z 1 73 dns:animaldiversity.ummz.umich.edu P http://animaldiversity.um 2013-01-09T19:53:31.814Z 200 472 http://animaldiversity.ummz.umich.edu/robots.txt P http://anim 2013-01-09T19:53:41.187Z 200 24437 http://animaldiversity.ummz.umich.edu/ - - text/html #002 2013 2013-01-09T19:53:41.188Z 1 63 dns:fonts.googleapis.com EP http://fonts.googleapis.com/css?fa 2013-01-09T19:53:41.193Z 1 63 dns:ajax.googleapis.com EP http://ajax.googleapis.com/ajax/lib 2013-01-09T19:53:41.857Z 302 232 https://animaldiversity.ummz.umich.edu/robots.txt LP https://a 2013-01-09T19:53:43.245Z 200 28 http://fonts.googleapis.com/robots.txt EP http://fonts.googlea 2013-01-09T19:53:43.245Z 200 4851 http://ajax.googleapis.com/robots.txt EP http://ajax.googleapi</pre> </div>
<p>4. If significant content is found to be missing, try to determine the cause.</p> <ol style="list-style-type: none"> a. Check the robots.txt file to determine if exclusions are responsible. b. To determine if the crawler even identified the content or encountered an issue, consult the <i>Crawl Log</i> (see Appendix B). 	<div style="border: 1px solid black; padding: 5px;"> <p>Crawl Log: Animal Diversity Web Archives (University of Michigan) (01/09/13 02:50 PM)</p> <p><i>The text below is automatically generated by the Heritrix web crawler</i></p> <pre>2013-01-09T19:53:29.612Z 1 73 dns:animaldiversity.ummz.umich.edu P http://animaldiversity.um 2013-01-09T19:53:31.814Z 200 472 http://animaldiversity.ummz.umich.edu/robots.txt P http://anim 2013-01-09T19:53:41.187Z 200 24437 http://animaldiversity.ummz.umich.edu/ - - text/html #002 2013 2013-01-09T19:53:41.188Z 1 63 dns:fonts.googleapis.com EP http://fonts.googleapis.com/css?fa 2013-01-09T19:53:41.193Z 1 63 dns:ajax.googleapis.com EP http://ajax.googleapis.com/ajax/lib 2013-01-09T19:53:41.857Z 302 232 https://animaldiversity.ummz.umich.edu/robots.txt LP https://a 2013-01-09T19:53:43.245Z 200 28 http://fonts.googleapis.com/robots.txt EP http://fonts.googlea 2013-01-09T19:53:43.245Z 200 4851 http://ajax.googleapis.com/robots.txt EP http://ajax.googleapi</pre> </div>
<p>5. Issues uncovered in your manual review may lead you to reference one or more of the following:</p> <ul style="list-style-type: none"> • Objective 5a (refine the seed URL) • Objective 5b (limit the scope and extent of the capture) • Objective 5c (expand the scope and extent of the capture) • Objective 5d (contact the content owner to revise robots.txt exclusions) • Objective 5e (delete problematic capture—only if devoid of research value—and launch new one) 	

5a. Refining the Seed URL(s)	
Procedure	Illustration
<p>1. Verify that your seed is accurate and correctly entered.</p> <ol style="list-style-type: none"> a. Visit the live site to confirm that the seed URL represents the targeted content. b. If running a “Host” crawl, be sure that the seed URL matches the actual URL on the site’s home page. Web servers will not always redirect the crawler to the appropriate page. c. If running a “Directory”-only crawl, be sure that the seed does not reference a specific page or resource (i.e. “index.html”) and that there is a trailing slash (“/”) at the end of the URL. 	<p>*Seed URLs: http://law.umich.edu www.law.umich.edu/Pages/default.aspx</p> <p>*Seed URLs: http://www.lsa.umich.edu/daas/ www.lsa.umich.edu/daas</p>
<p>2. If necessary, expand the seed URL:</p> <ol style="list-style-type: none"> a. Replace the seed with a broader one (a domain instead of a directory) b. Add an additional seed (i.e. a subdomain or external site). c. Check if the content owner employs two domains (http://site.com as well as http://www.site.com); you may need to include both as seeds to fully capture the site. 	<p> *Seed URLs: http://www.lsa.umich.edu/daas http://isa.umich.edu/daas</p>

5b. Limiting the Scope and Extent of the Capture	
Procedure	Illustration
<p>1. Verify that your seed is accurate and correctly entered.</p> <ol style="list-style-type: none"> a. Visit the live site to confirm that the seed URL represents the targeted content. b. If running a “Directory”-only crawl, be sure that: <ul style="list-style-type: none"> • The seed does not reference a specific page or resource (i.e. “index.html”) • There is a trailing slash (“/”) at the end of the URL. 	 <p>*Seed URLs: <input type="text" value="http://www.lsa.umich.edu/daas/"/>   <input type="text" value="www.lsa.umich.edu/daas"/></p>
<p>2. Based upon your analysis of reports and manual review, you may need to “Edit Capture Settings” and do one (or more) of the following:</p> <ol style="list-style-type: none"> a. Limit the scope to “Directory” or “Page only” (depending on target). b. Disable the option to capture “Capture linked pages” c. Reduce crawl duration to one hour. 	 <p>a.  Scope: <input type="text" value="Directory"/></p> <p>Capture Linked Pages: <input checked="" type="radio"/> No <input type="radio"/> Yes b.</p> <p>Max. Time: <input type="text" value="Brief Capture (1 hour)"/> c. <input type="text" value="Brief Capture (1 hour)"/> <input type="text" value="Full Capture (36 hours)"/></p>

5c. Expanding the Scope and Extent of the Capture	
Procedure	Illustration
1. Verify that your seed URL matches that of the target and is correctly entered.	 <p>*Seed URLs: http://law.umich.edu  www.law.umich.edu/Pages/default.aspx</p>
2. It may be necessary, to add an additional seed if the important content is stored in a subdomain, external site, or variant of the target (i.e. http://site.com as well as http://www.site.com).	 <p> *Seed URLs: http://www.lsa.umich.edu/daas <small>http://www.example.com</small> http://lsa.umich.edu/daas</p>
3. Based upon your analysis of reports and manual review, you may need to “Edit Capture Settings” and do one (or more) of the following: <ol style="list-style-type: none"> a. Expand the scope to “Directory” or “Host” b. Enable the option to “Capture Linked Pages.” Use with great caution, as this will include all manners of superfluous content in the capture. c. Increase the maximum duration of the crawl to 36 hours. 	 <p style="text-align: center;"> a.  Scope: Host site  Capture Linked Pages: <input checked="" type="radio"/> No <input type="radio"/> Yes b. c. Max. Time: Full Capture (36 hours)  </p>

5d. Contacting the Content Owner

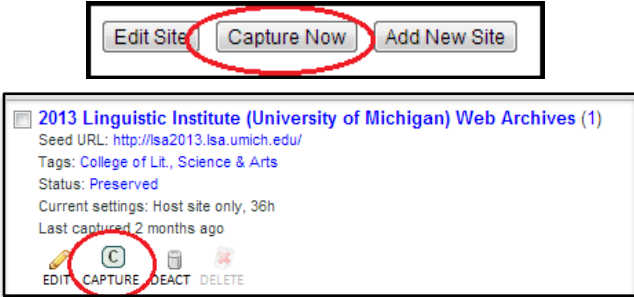
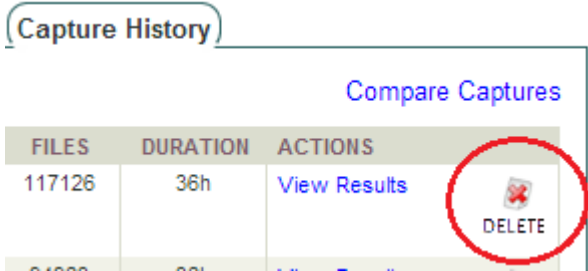
1. Content owners can do a variety of things to improve conditions for Web crawling: adjusting robots.txt exclusions, optimizing content and structure, refining navigation paths, etc. At the same time, it may be difficult (if not impossible) to find contact information for content owners and even if a message is sent, nothing may ever come of it.

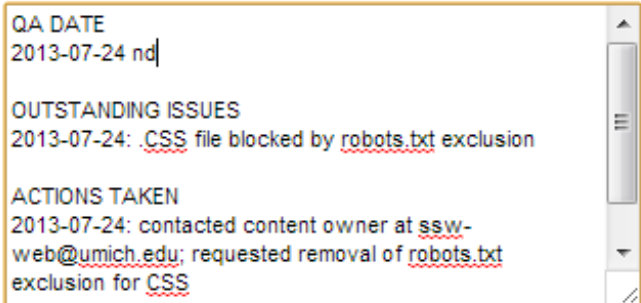
NOTE: See [Appendix C](#) for templates to be used in requesting content owners to take action in regards to robots.txt exclusions.

2. Requests may be for the content owner to:

- a. Modify or remove an exclusion (or time limit or other feature) so the CDL WAS crawler may access additional content.
- b. Impose a restriction so the CDL WAS crawler may avoid a crawler trap or a large body of superfluous content.

3. You may also recommend that a Project Administer contact a content owner if there are particular features of a site that make capture problematic or if additional information is needed about the site structure to satisfactorily conduct the crawl.

5e. Deleting Problematic Crawls and Launching New Ones	
Procedure	Illustration
<p>1. To verify that the revised settings have worked (or to acquire a more accurate capture), you may crawl the site again by clicking the “Capture Now” button. This new crawl will be reviewed in the next round of QA.</p> <p>If the content owner complies with a request, you may need to go back to the site and initiate a new crawl. From the Manage Sites screen, filter for the site and click the “Capture” button. This new crawl will be reviewed in the next round of QA.</p>	 <p>The illustration shows a screenshot of a web interface. At the top, there are three buttons: 'Edit Site', 'Capture Now', and 'Add New Site'. The 'Capture Now' button is circled in red. Below this is a card for a site titled '2013 Linguistic Institute (University of Michigan) Web Archives (1)'. The card contains details like 'Seed URL: http://lsa2013.lsa.umich.edu/' and 'Status: Preserved'. At the bottom of the card, there are four icons: a pencil (EDIT), a camera (CAPTURE), a power button (DEACT), and a trash can (DELETE). The 'CAPTURE' icon is circled in red.</p>
<p>2. If you have launched a new crawl and the previous capture is without research value and/or takes up an inordinate amount of space, it may be necessary to delete the problematic capture. Consult with Digital Curation, if necessary.</p> <p>Go to the Capture History tab on the Site Summary screen and click the “Delete” button for the appropriate crawl.</p> <p>NOTE: Do not delete a crawl if this action will result in a noticeable gap in the preservation of a site. It is more important to document versions of a site across time than to save storage space.</p>	 <p>The illustration shows a screenshot of a 'Capture History' table. The table has columns for 'FILES', 'DURATION', and 'ACTIONS'. The first row shows '117126' files and a '36h' duration. The 'ACTIONS' column for this row contains a 'View Results' link and a 'DELETE' button. The 'DELETE' button, which features a trash can icon, is circled in red.</p>

6. Documenting QA Process	
Procedure	Illustration
<p>1. If necessary, create a QA DATE heading in the Custodial Note field and enter the current date (YYYY-MM-DD) followed by your initials.</p> <p>If there are no issues or required actions, you do not need to enter any more information.</p>	<p>Curatorial Notes:</p>  <p>Curatorial notes will not display to the public.</p>
<p>2. If necessary, create an OUTSTANDING ISSUES heading. Include the date (YYYY-MM-DD) to note when the issue first arose and only record major issues (such as blocked .CSS file or absence of highly significant content).</p> <p>Try to be specific in referring to any problems so that future QA staff can follow up on the problem.</p> <p>If you discover that a previously reported issue has been resolved, delete it from the list.</p>	
<p>3. If necessary, create an ACTIONS TAKEN heading. Include the date (YYYY-MM-DD) and record any actions you took: modification of seed URL (noting previous URL) and changes to crawl settings (duration, scope, links). Also include a brief summary of any correspondence with content owners (with name/email address of contact, if available).</p>	

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:
3.0	September 9, 2013	Michael Shallcross, Assistant Archivist
2.0	March 20, 2013	Elise Reynolds, QA Specialist
1.0	September 21, 2011	Michael Shallcross, Assistant Archivist

Appendix A: Problem Content and Technical Issues in Website Preservation

Common Website features and known issues with crawling technology make it difficult for archivists to capture the exact form, functionality, and content of sites as they are experienced on the 'live' web. To compound matters, a crawler may successfully capture a site, but the Wayback Machine may not be able to properly render it for end-users.

Be aware of the following types of content and technical issues during the QA process:

- **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, an Internet convention used by webmasters to prevent all or certain sections of websites from being captured by a web crawler.
 - The Bentley Historical Library will respect all robots.txt exclusions.
 - If significant historical or administrative information is blocked from capture by robots.txt exclusions, the library may ask content owners to revise these settings.
 - While robots.txt exclusions are usually found in a file at the root level of the host domain, they may be incorporated in HTML headers:

```
<meta name='robots' content='noindex,nofollow' />
```
- **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**: Web crawlers are unable to properly capture such material because:
 - Streaming audio and video content often provide a progressive download of content via non-HTTP protocols (such as RTMP or MMS).
 - Embedded players require the source code of the application to render and function properly.
- **Social media sites**: Sites such as Facebook, Twitter, Flickr, and YouTube pose challenges due to:
 - Their structure and design (i.e. the use of JavaScript and multiple links to embed video content).
 - Policies governing access to web crawlers (documented in terms of service and enforced via robots.txt exclusions) 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:
 - Password or [Captcha](#) authentication.

- Content accessed by drop-down menus, radio dials, or form entry.
- Databases.
- 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 U-M Law School homepage is at <http://www.law.umich.edu/Pages/default.aspx> rather than <http://www.law.umich.edu/>, as might be expected.
- 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.
 - The .CSS file may be located in a directory that is excluded from the capture by the robots.txt
 - Sometimes, a .CSS file may inherit 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 some versions of 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.

Appendix B: Working with the WAS Crawl Log

The Crawl Log records every operation performed by Heritrix as it crawls a website. It therefore allows you to retrace every request sent by the crawler and determine what content was included in the crawl and what was missed.

Crawl Log: 1000 Pitches Web Archives (University of Michigan) (12/14/12 10:01 PM)

The text below is automatically generated by the Heritrix web crawler

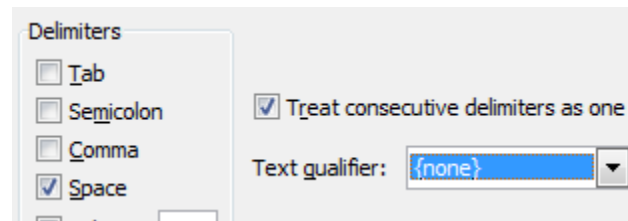
```
2012-12-15T03:02:55.959Z 1 55 dns:1000pitches.com P http://1000pitches.com/ text/dns #001 20121215030254676+16 sha1:PRS3A5LZVFL7FY6BUQERKMJT53HNF3YP - -
2012-12-15T03:02:58.156Z 301 185 http://1000pitches.com/robots.txt P http://1000pitches.com/ text/html #002 20121215030257995+138 sha1:ECTYC6BV4IY4ZFTBUVIJYVOQ6UM7C
2012-12-15T03:02:59.757Z 200 1561 https://1000pitches.com/robots.txt PR http://1000pitches.com/robots.txt text/plain #001 20121215030258159+561 sha1:ONMDDIRORSASBJM:
2012-12-15T03:03:00.321Z 301 185 http://1000pitches.com/ - - text/html #002 20121215030300189+125 sha1:ECTYC6BV4IY4ZFTBUVIJYVOQ6UM7CUHW - 3t
2012-12-15T03:03:11.001Z 200 16214 https://1000pitches.com/ R http://1000pitches.com/ text/html #001 20121215030309771+404 sha1:PG24RU42X4CY3VPR5R35UD7DYDR5SMAX - -
2012-12-15T03:03:11.001Z 1 50 dns:www.w3.org REP http://www.w3.org/1999/xhtml/vocab text/dns #002 20121215030310806+1 sha1:ZB5EZSKCTBBVT53ZG3R7WBR22CVVAQ3M - -
2012-12-15T03:03:13.385Z 200 2727 http://www.w3.org/robots.txt REP http://www.w3.org/1999/xhtml/vocab text/plain #001 20121215030313060+201 sha1:Y5WKVPVSRQ6BVPGBJ7D:
2012-12-15T03:03:15.668Z 301 243 http://www.w3.org/1999/xhtml/vocab RE https://1000pitches.com/ text/html #001 20121215030315424+216 sha1:HDLDKTHVLATUOT3RGQQWNU5O4I
2012-12-15T03:03:18.247Z 200 30287 http://www.w3.org/1999/xhtml/vocab/ RER http://www.w3.org/1999/xhtml/vocab text/html #001 20121215030317683+332 sha1:3UBKZ7Y3EYOFAI
2012-12-15T03:03:21.743Z 200 111671 https://1000pitches.com/sites/default/files/js/js 76Em0JjbxymggyJV52e7aWds-FfoBE2JT7HPacb MY.js RE https://1000pitches.com/ applic
```

The data is arranged into 12 tab-delimited columns. You can do keyword (CTRL-F) searches in the web interface or download the report and import it into an Excel spreadsheet so that you can sort on columns and do more analysis.

Column	Data (see http://crawler.archive.org/articles/user_manual/analysis.html)
1	Timestamp in ISO8601 format, to millisecond resolution
2	HTTP status codes (will be a negative number if URL processing was unexpectedly terminated)
3	Size of the downloaded document in bytes
4	URI of the document downloaded NOTE: will include full path and filename of resource
5	Breadcrumb codes showing the trail of downloads that got us to the current URI
6	URI that immediately referenced this URI ('referrer')
7	Document mime type
8	Id of the worker thread that downloaded this document
9	Timestamp indicating when a network fetch was begun (and millisecond duration of the fetch, separated by a '+' character)
10	SHA1 digest of the content only (headers are not digested)
11	'Source tag' inherited by this URI, if that feature is enabled
12	"Annotations", if any have been set. Possible annotations include the number of times the URI was tried ('-' if never retried)

To import into an Excel spreadsheet:

1. Save as (or copy and paste into) a plain text file. (In Google Chrome: save as a web page, change extension to .txt, and then delete the HTML tags at the beginning and end).
2. Open a new Excel workbook; go to the “Data” menu tab and under the “Get External Data” area, click “From Text.”
3. Browse to the file; a “Text Import Wizard” dialogue window will open.
4. Choose “Delimited” data type and click Next.
5. The Delimiter should be “Space.” Check the box next to “Treat consecutive delimiters as one” and change “Text qualifier” to {none}



6. Click Finish and then OK to put the data in the Existing worksheet (at =\$A\$1).

A	B	C	D	E	F	G	H	I	J	K	L
2012-12-15T03:02:55.959Z	1	55	dns:1000pitches.com	P	http://1000pitches.com/	text/dns	#001	20121215030254676+16	sha1:PRS3A5LVFL7PY6BUQERKMJT5-	-	-
2012-12-15T03:02:58.156Z	301	185	http://1000pitches.com/robots.txt	P	http://1000pitches.com/	text/html	#002	20121215030257995+138	sha1:ECTYC6BV4IY4ZFTBUVTJYVOQ6U-	-	-
2012-12-15T03:02:59.757Z	200	1561	https://1000pitches.com/robots.txt	PR	http://1000pitches.com/	text/plain	#001	20121215030258159+561	sha1:ONMDDIRORSASBJMJHPGVPM5-	-	-
2012-12-15T03:03:00.321Z	301	185	http://1000pitches.com/	-	-	text/html	#002	20121215030300189+125	sha1:ECTYC6BV4IY4ZFTBUVTJYVOQ6U-	-	3t
2012-12-15T03:03:11.001Z	200	16214	https://1000pitches.com/	R	http://1000pitches.com/	text/html	#001	20121215030309771+404	sha1:PG24RU4ZX4CY3VPR5R35UD7DY-	-	-
2012-12-15T03:03:11.001Z	1	50	dns:www.w3.org	REP	http://www.w3.org/1999/	text/dns	#002	20121215030310806+1	sha1:ZB5EZSKCTBBVT53ZG3R7WBR22-	-	-
2012-12-15T03:03:13.385Z	200	2727	http://www.w3.org/robots.txt	REP	http://www.w3.org/1999/	text/plain	#001	20121215030313060+201	sha1:Y5WKVPVSRQ6BVPGBJ7D4XQZG-	-	-
2012-12-15T03:03:15.668Z	301	243	http://www.w3.org/1999/xhtml	RE	https://1000pitches.com/	text/html	#001	20121215030315424+216	sha1:HDLDKTHVLATUOT3RGQQWNU5-	-	3t
2012-12-15T03:03:18.247Z	200	30287	http://www.w3.org/1999/xhtml	RER	http://www.w3.org/1999/	text/html	#001	20121215030317683+332	sha1:3UBKZ7Y3EYOFALQK64LIZ3XIK67-	-	-

You are now ready to analyze your data! You might:

1. Sort on column 4 (URIs captured) to group resources by domain or directory.
2. Sort by column 2 (HTTP response code) to see which documents failed to download due to 404 or other errors.
3. Sort by column 7 (mime type) to group format types and review which files (for instance, videos on a certain page) were captured.
4. Trace the route taken by the crawler from a URL that you know has been captured (for instance, to see if other files embedded on a particular page were captured).

The possibilities, while not endless, abound! Consult with Digital Curation if you have questions about how to work with the report.

Appendix C: Correspondence for Content Owners

Request to Remove Robots.txt Exclusions

My name is [YOUR NAME] and I work in the [DIVISION NAME] of the University of Michigan's Bentley Historical Library, which serves as the official archives of the University of Michigan and also documents the activities of the people, organizations, and voluntary associations of the state of Michigan.

Archivists at the Bentley have determined that your website at [URL: <http://...>] represents important aspects of [the University of Michigan's intellectual life OR Michigan's cultural and/or socioeconomic life] and warrants long-term preservation as a historical record. To this end, we have attempted to preserve an archival copy of your website with the California Digital Library's Web Archiving Service (WAS; for more information about WAS, see <http://webarchives.cdlib.org/p/webmasters>).

In reviewing a recent archived version, we discovered that our web crawler was blocked from capturing content stored in the following location(s):

[list blocked folders]

I therefore write to ask if you could please modify your site's robots.txt exclusions so that we may preserve a more accurate representation of your Website. Our user-agent is: cdlwas_bot

For more information on the Bentley Historical Library and our Web archives, please see: <http://bentley.umich.edu/dchome/webarchives/guidelines.php>. If you have questions or comments, please send them to bhlwebarchive@umich.edu or call Nancy Deromedi or Michael Shallcross at 734-764-3482. Thank you very much for your time and consideration.

Request to Add Robots.txt Exclusions to Limit Crawls

My name is [YOUR NAME] and I work in the [DIVISION NAME] of the University of Michigan's Bentley Historical Library. The Bentley serves as the official archives of the University of Michigan and also documents the activities of the people, organizations, and voluntary associations of the state of Michigan.

Archivists at the Bentley have determined that your Website at [URL: http://...] represents important aspects of [the University of Michigan's intellectual life OR Michigan's cultural and/or socioeconomic life] and warrants long-term preservation as a historical record. To this end, we have preserved an archival version of your Website with the California Digital Library's Web Archiving Service (for more information, see <http://webarchives.cdlib.org/p/webmasters>).

In reviewing a recent archived version, we discovered that our web crawler is capturing a large amount of unnecessary data related to [DESCRIBE SUPERFLUOUS CONTENT].

I therefore write to ask if you could include the following exclusion in your site's robots.txt file so that we may preserve a copy of it in a more efficient manner:

User-agent: cdlwas_bot

Disallow: [CONSULT WITH DIGITAL CURATION]

For more information on the Bentley Historical Library and our Web archives, please see: <http://bentley.umich.edu/dchome/webarchives/guidelines.php>. If you have questions or comments, please send them to bhlwebarchive@umich.edu or call Nancy Deromedi or Michael Shallcross at 734-764-3482. Thank you very much for your time and consideration.