Preserving State Government Digital Information
Minnesota Historical Society

Web Archive Evaluations

Archive-It Testing, 2009

Introduction
Testing of the Internet Archive’s Archive-It\(^1\) service was undertaken by Minnesota Historical Society (MHS) staff between January 17\(^{th}\) and February 23\(^{rd}\) 2009 in connection with the National Digital Information Infrastructure and Preservation Program (NDIIPP) grant to MHS for the preservation of state government digital information. The testing focused on capturing Minnesota Statutes, Laws, and Rules pages and additional contextual information in order to preserve and provide effective access to this collection of information.

A previous test, conducted by MHS staff in May of 2007, attempted to capture web pages of the Minnesota Legislature including the Office of Revisor of Statutes pages containing the Statutes, Laws, and Rules. Problems with content capture encountered during the 2007 test did not occur in 2009 because of changes to the structure of the Revisors’ web pages and improvements to the Archive-It services.

Features
Archive-It captures web pages using an open source internet crawler called Heritrix\(^2\) that copies and indexes web pages in much the same way a search engine robot does. An archiving project begins by setting up a list of web addresses, “seed” URLs, that the crawler will use to begin its search. The Heritrix crawler will then capture all links that contain the seed as the root.\(^3\) Standard crawler configurations may be modified to change the depth of links followed, the hosts, domains, or URL paths crawled.\(^4\) For example, Heritrix will not usually capture subdomains of a seed (it would capture all pages beginning www.mnhs.org, but not www.archive.mnhs.org), but it may be configured to do so by adding scope expansion rules (called SURTs)\(^5\). Embedded content, such as images will normally be captured.

Seeds can be limited to specific website directories or single pages, and users can specify the frequency and duration of crawls. Host constraints can also help focus on the precise portion of a desired site. Page limits can help users zero in a desired content as well.

\(^1\) [https://archive-it.org/](https://archive-it.org/)
\(^3\) [http://webteam.archive.org/confluence/display/ARIH/Archive-It+Getting+Started](http://webteam.archive.org/confluence/display/ARIH/Archive-It+Getting+Started)
\(^4\) [http://crawler.archive.org/articles/user_manual/config.html#scopes](http://crawler.archive.org/articles/user_manual/config.html#scopes)
\(^5\) [http://webteam.archive.org/confluence/display/ARIH/Scope+Rules](http://webteam.archive.org/confluence/display/ARIH/Scope+Rules)
A new scoping tool allows users to better assess the number and breadth of pages that will be captured by a given set of seeds and settings before starting a crawl and can also be used as an analysis tool after a crawl had been conducted. Another new tool, de-duplication, removes identical pages if they have already been captured during previous crawls.

The Internet Archive has recently improved access to archived video files, a feature that state governments such as North Carolina are using. Highlights of the improvements made to the Internet Archive services since 2006 can be found in the “Features List” below.

Results
The 2009 test crawl had a limit of ten seed URLs. The Revisors Office changed the presentation of their material since the 2007 crawl so that it was no longer hidden behind a database search field. The database search functionality is still available, but they have added links to the legislative content through a keyword index and standard page links. In addition, the Internet Archive has made pages with https addresses available for viewing and searching since the 2007 test. Previously, these pages could be captured by the crawler, but were not released for viewing and search.

Because the Revisors Office content page names were in subdomains of the seed pages, scope rules were needed to capture them. Internet Archive staff assistance and documentation explained how to write these and the desired pages were retrieved once they were in place.

A number of reports are available to help evaluate the effectiveness of the captures such as lists of file types, number of URLs crawled, status codes and the number of associated pages, and hosts encountered. A summary report lists all of the seeds and indicates whether the crawl finished or simply stopped when the time limit was reached.

The Minnesota Revisors Office web pages containing the Statutes were successfully captured during this test. A large number of Laws and Rules pages remained in the list of queued documents that had been identified, but not yet archived when the crawl reached its time limit of 36 hours. A list of URLs in the queue was provided in one of the Archive-It crawl reports. Checking a sample of the queued Rules pages consistently found that they were alternate addresses for pages that had already been captured. This probably occurred as a result of recursive link loops created by the new page navigation features. The sampled Laws URLs in the queue however represented only pages that were not archived. Arranging a longer crawl time is recommended when this type of problem is encountered.

One of the seed URLs did not render correctly because the address for the cascading style sheet was outside the domain for its seed. The solution was to add the style sheet page to the seed URL list. Once this problem was identified it was easy to remedy when it occurred again on another page.
Robots text files on the Revisors Office pages prevented Heritrix from accessing some of older files that may have been useful to archive. Capture of this content would require contacting the Revisors Office and asking them to give Heritrix permission to crawl these URLs.

Audio of House and Senate committee hearings was successfully archived. Capture of video was not successful because the video player, not the content was imbedded in the pages. Finding the correct seed URL to target the video content without getting additional, unwanted content also proved somewhat difficult.

Search and Access
Access to archived pages is available once a crawl is completed. Archive-It provides a basic and advanced keyword search function. Advanced searches can specify “all words,” “none of the words,” “any of the words,” and “exact phrase” strings. The search can also be limited by host or file type and by dates. Searches for exact URLs can be conducted via the Internet Archive’s access tool, the Wayback Machine.

Searching captures is a key to evaluating the effectiveness of the archiving effort. Unfortunately, the keyword search function was not working on the collection during the testing period. Date limited searching also did not work. Search problems may have been a glitch related to improvements to the text indexing being implemented at the time of the test. Subsequent keyword searches on the 2009 collection provided good results that may have benefited from the improvements to the relevance ranking delivered by the NutchWAX search engine upgrade. URL searches were performed without any problems.

Summary
The test captured most of the desired material from the Revisors office although the crawl timed-out with a large number of files remaining in the queue because of recursive link loops created by keyword access added to some of the pages relatively recently. This experience demonstrated the importance of time limits for some crawls. Good scoping prior to starting a crawl and adding constraints could have mitigated this problem. Missing files on the other hand would have benefited from increasing the crawl time for that particular seed.

Rendering for two of the pages was incorrect until seeds were added for the out-of-scope cascading style sheets and scope rules were required for seeds URLs where subpage addresses varied from the norm. Without the assistance of the Internet Archive’s partner specialist, Molly Bragg, this might have taken a considerable amount of time to diagnose and fix. She quickly spotted and fixed the problem. The Internet Archive generally encourages contacting partner

---

6 http://www.archive.org/web/web.php
7 http://archive-access.sourceforge.net/projects/nutch/
specialists whenever such questions arise and during the testing Molly Bragg was readily available to offer any assistance required.

The attempt to capture video files of legislative committee meetings managed to archive the media player file itself, but not the actual video files. Capture of these files apparently failed due to the way in which they were posted to the web, making this content simply unavailable for archiving.

Similarly, database search forms still present an insurmountable obstacle to internet archiving, just as they do to search engine indexing. Fortunately, none of the desired material in this test had to be ruled out for that reason, but anyone wishing to preserve such information should not consider internet archiving a viable choice unless the data provider has posted a sitemap\(^8\) that includes the database records. A webmaster may post a sitemap file containing URLs for all pages on their website so that search engine crawlers can find otherwise hidden content. The Heritrix crawler would also be able to take advantage of a sitemap to find and archive data records.

**California Digital Library, Web Archiving Service Testing, 2009**

**Introduction**

Development of California Digital Library (CDL) Web Archiving Service (WAS) was a direct result of the CDL’s Web-at-Risk project. Launched in 2005 with funding from a Library of Congress, National Digital Information Infrastructure and Preservation Program (NDIIPP) grant, the CDL website describes the project as, “a four and one half-year effort … to develop tools that enable librarians and archivists to capture, curate, preserve, and provide access to web-based government and political information.” The project primarily focused on state and local government information.\(^9\)

As an NDIIPP partner, Minnesota Historical Society (MHS) project staff were invited to participate in testing the new service. The tests were conducted between February 13\(^{th}\) and March 17\(^{th}\) 2009.

---

8 http://www.sitemaps.org/index.php
9 http://www.cdlib.org/inside/projects/preservation/webatrisk/
Features

The CDL’s WAS employs the same open source web archiving tools used by Archive-It; the Heritrix crawler to collect content and the Wayback Machine interface combined with a NutchWAX search engine to provide access to the archived web pages.

The tools available to specify the scope of web pages captured were comparable to those offered by Archive-It. Crawls could be limited to a page, directory, or the host website for the seed URL. The “Capture linked pages” setting allows the user to direct Heritrix to follow external links one hop off from the host domain, if desired. Associated images and style sheets are ordinarily captured and do not require special instructions to do so. The WAS does not implement SURTs (see above), so the crawler cannot capture content found on subdomains in a website unless a seed is entered for each one. For example, to obtain a comprehensive capture of Minnesota legislative branch pages, seeds would be required for http://www.leg.state.mn.us/, http://www.house.leg.state.mn.us/, http://www.senate.leg.state.mn.us/, https://www.revisor.leg.state.mn.us/, and http://www.commissions.leg.state.mn.us/d. Heritrix can collect content from HTTP and HTTPS URLs, but not from FTP sites. Users can specify the frequency and duration of crawls. Frequently changing content can be collected daily, but an end date must be set to avoid unintentionally running up a huge number of files. Crawls may also be set to run for as little as an hour or for up to 36 hours.

WAS does not provide special scoping tools, but they do encourage users to check crawl results, delete unsatisfactory captures, and run them again after adjusting the scope. A “compare” feature allows users to analyze the data from any two crawls to determine the effect of changes in settings or changes in content by providing lists of files that have been changed, added, deleted, or that remained unchanged.

The WAS offered basic training in form of WebEx demonstrations (archived for reference purposes10) that covered all aspects of web archiving – from content selection, seed identification, and crawl settings, to the basics of quality assurance checking after content has been captured.

The roll-out of public access to the archive just after the testing period11 12 meant that WAS provided additional training on how to set-up public interfaces for collected material. This included a discussion of legal issues related to intellectual property and copyright that reviewed the findings of the Section 108 study group formed by the Library of Congress.13

10 http://was.cdlib.org/user/training
11 http://www.digitalpreservation.gov/news/2009/20090715_article_was.html
12 http://webarchives.cdlib.org/
13 http://www.section108.gov/
Results
The WAS test focused on capturing Minnesota Statutes, Laws, and Rules pages. Contextual material that was not included in the Archive-It test, such as House and Senate Journals and historical biographies of state legislators were added to the scope of this collecting effort.

The WAS provides a set of reports similar to Archive-It’s. A summary report allows users to view the number and size of files returned for each seed. File types, response codes, hosts encountered with a map of their locations, and a list robots.txt files are also available, as well as a full list of captured files. Keyword searches can be performed against text in any of the archived pages.

Sampling of results consistently showed that contextual materials appeared to have been successfully archived. Minnesota Revisors Office seeds, however returned "Not in archive" messages when attempting to render pages in the Wayback Machine. The Revisors pages that were captured and displayed did not have any style sheet formatting and it was not possible to use website navigation elements to click through captured pages. WAS staff determined that all of these problems appear to have been caused by the link rendering/rewriting engine they use to work around a problem the Wayback Machine has displaying material from secure addresses, such as those used for the Revisors web pages.

Capture of the Revisors pages was also incomplete. The crawl appears to have timed-out once again before all of the pages were archived. The large number of pages, website structure and crawler settings were all identified as potential causes of this problem. WAS staff recommended extending the crawl period beyond the 36 hour limit and narrowing the scope to only the host pages. The scope of all test crawls were set to capture the host plus linked pages.

The crawls did not seem to encounter problems with any particular file type, but random checks found a PDF file that was not captured. The reason for this was not determined.

Once again, audio files successfully transferred into the archive, but some of the video files did not because the method used to post these files simply renders them unavailable to crawlers.

Search and Access
Archived pages may be searched in a variety of ways through the administrative interface once a crawl is completed. Keyword and string searches can be performed on a collection exactly as they would be made on an internet search engine, but Boolean searches aren’t supported. The search can also be limited by file type, but not by date. Searches across collections but within the same “project” are not currently supported. If a user wishes to search across a variety content generated by seed URLs from different hosts, these should be entered in the same collection.

Since the WAS is a new archiving service, a few glitches were inevitable. File indexes did not work properly during testing. This made checking the crawl results difficult because keyword
searches were not fully functional and URL searching was not available. Lists of all files captured for a given seed were prohibitively slow to browse.

While it is possible to see that a crawl ran for the maximum allowable time, it is difficult to determine if the entire site was captured. The WAS does not provide an indication of whether a crawl was completed or simply timed-out, nor does it list any queued files at the time the crawl stopped as the Archive-It service does. This combined with the issues mentioned previously made it extremely difficult to assess whether or not all desired files were captured.

MHS opted not to publish archived pages during the test period, when the new public interface became available. WAS can serve as a dark archive as well as a public access point for archived web sites. Collections can be made public at any time, should the archivist decide to do so. WAS administrators discourage removing material once it has been published, though, to avoid breaking links created by others while the pages were available.

**Summary**
Contextual information posted on the Minnesota Legislative Reference Library and the Minnesota House and Senate’s websites was successfully archived. Pages from the Minnesota Revisor of Statutes were also largely archived, but captured pages did not render correctly and navigation was not functional because of problems with the Wayback Machine’s ability to display secure websites.

MHS participated in the beta testing of the CDL’s WAS just as staff were preparing to roll-out initial public access. Predictably, trouble shooting assistance from a partner specialist was not comparable to the kind of service Archive-It could offer.

**Quality assurance checking is fundamental to web archiving.** Reports and access tools provided by the archiving service will help or hinder a user’s ability to assess the quality of an archiving effort. While basic reports were available, WAS did not offer some information - like crawl completion status - that would help a web archivist determine the success of a crawl. Service administrators welcomed feedback on the usefulness of the tools provided and will no doubt offer enhanced quality checking features in the future.

The WAS training materials, in the form of webcasts and written guidelines, are excellent and cover, not just the mechanics of web archiving, but collection planning and guidance on policy decisions relating to access and use of collections.

**Archive-It and WAS Testing Conclusions**
Both the Internet Archive and CDL WAS tests encountered examples of content that is difficult or impossible to capture, such as the Minnesota Senate committee meeting videos. The Internet Archive and other web archiving services have documented some of the common barriers to archiving efforts. Significantly, they point out that anything which prevents the Heritrix crawler
from capturing information will also prevent other crawlers, including those used by internet search engines, from finding the information and making it available to potential users. Some typical causes of archiving failures are:

- Navigation menus created using JavaScript, typically those that show a list of choices when a mouse rolls-over part of a web page, cause problems. Even when content from web page that use JavaScript can be archived, reproducing the menus and other JavaScript files can be difficult for the Wayback Machine.
- Dynamically generated web pages of other types also cause problems. Some dynamic content can be easily stored in an archive and some of it fails completely. When a dynamic page renders standard html, the archive will usually function correctly. When a dynamic page contains forms, JavaScript, or other elements that require interaction with the originating host, the archive will not contain the original site's functionality.\(^\text{14}\)
- Any functionality on a web page that needs to contact the originating server in order to work will fail when archived.
- Video poses any number of problems, some of which are related to the issues mentioned above. Video may not be embedded in a page, but instead send the user to outside content (such as YouTube). Streaming media may use transfer protocols other than HTTP and proprietary software that may not be recognized by the crawler.\(^\text{15}\text{16}\)
- Password Protected Sites cannot be crawled. Most archive services restrict the scope of collecting to publicly available material.
- Content hidden behind forms and database driven content cannot be captured because the Heritrix crawler cannot use the forms and query databases to retrieve information. However, if there are links into the raw content, crawlers can follow those. Also if there is an XML site map, this can be included in a seed list and links included on the site map can be followed.
- Sometimes a webmaster will use a robot.txt exclusion to prevent some or all of the content on a site from being crawled.\(^\text{17}\)

Scoping sites prior to a crawl to determine if a seed will lead the crawler to the desired content is clearly an art that benefits from a good knowledge of website construction and some research prior to running a crawl. The Internet Archive’s new scoping tool should improve this process, especial for less knowledgeable users. The more focused the content acquisition goals are, the more important it is to find the right combination of seeds or seeds and constraints. Even an organization with relatively broad archiving goals can make better use of their storage budgets by designing a targeted content capture.

\(^\text{14}\) http://www.inter4m.com/archive-it/faq.html
\(^\text{15}\) http://en.wikipedia.org/wiki/Streaming_media#Protocol_issues
\(^\text{16}\) http://liwa-project.eu/images/uploads/d1-1_1_requirements_beg_v1.0.pdf
\(^\text{17}\) https://webarchive.jira.com/wiki/display/ARIH/5+Challenges+of+Web+Archiving
Services such as the Internet Archive and the CDL’s WAS do an excellent job of providing a record of the substance and appearance of information provided on the web at a given point in time and they are particularly good at documenting change to that content over time. Access, while important to use of the captured material, is necessarily of secondary importance. Full-text indexing of archived web pages makes content relatively easy to find, but the NutchWAX tool used to index and search web archive files does not return results on par with a Google search of the open web. Assembling material that may be easily searched together in a collection offers another possible incentive to capture pages in a web archive.18

The most successful web archiving projects will have goals that are consistent with the strengths of the archiving service. Digital records that can be effectively preserved in other forms or do not change often and can easily be made available on the live web may not be ideal candidates for an Internet Archive collection. In the most recent Minnesota test, the contextual information for the legislative processes such as meeting minutes, research, and reports were a good match because new content is posted frequently and not all of it is consistently collected and preserved in digital form. The Statutes, Laws, and Rules pages bring the original documents that serve as reference points to the other materials into the collection, even though the digital version of these records may be more likely to be preserved in some other way. The cost of preserving and accessing records in some other digital form, the relative value and suitability of the records, and the goals of the organization considering a collections program must be weighed against the fees and value provided by the archiving service.

18 http://en.wikipedia.org/wiki/Web_archiving
Archive-It Features List

Improvements to the documentation and to the user interfaces are made on a continuous basis.

Archive-It Release Notes Highlights

1.5 [http://webteam.archive.org/confluence/display/ARIH/Archive-It+Feature+Release+1.5](http://webteam.archive.org/confluence/display/ARIH/Archive-It+Feature+Release+1.5)
   - Addition of a bookmarklet allows seeds to be added easily while browsing a website.
   - Search results can be sorted by date and date range.

2.0 [http://webteam.archive.org/confluence/display/ARIH/Archive-It+2.0+Feature+Release+Notes](http://webteam.archive.org/confluence/display/ARIH/Archive-It+2.0+Feature+Release+Notes)
   - Bulk seed edits wizard added.
   - Search results inside the application now return hits both from the archived pages and seed metadata assigned to seeds.

2.4 [http://webteam.archive.org/confluence/display/ARIH/Archive-It+Feature+Release+2.4](http://webteam.archive.org/confluence/display/ARIH/Archive-It+Feature+Release+2.4)
   - Subscriber’s home page now displays total number of documents that they have archived.
   - Duplicate seed warning function when using the "add seeds" link.

2.6 [http://webteam.archive.org/confluence/display/ARIH/Archive-It+Feature+Release+2.6](http://webteam.archive.org/confluence/display/ARIH/Archive-It+Feature+Release+2.6)
   - Search results can be defined by document, host and date on the public website.
   - “Robots.txt” report will show the number of documents discovered but not crawled due to robots.txt exclusions for each host with a link to view and download the specific documents excluded.
   - “Queued” will show the number of documents discovered but not crawled due to the crawler's time limit with a link to view the documents that were still in queue when the crawl ended.  Regarding documents that show up as queued, sometimes this can happen due to a crawler trap on the site you are trying to crawl…In many instances a crawl trap can be avoided with some simple modifications.  The Archive-It partner specialist can help you set these up.

2.8 [http://webteam.archive.org/confluence/display/ARIH/Archive-It+Feature+Release+2.8](http://webteam.archive.org/confluence/display/ARIH/Archive-It+Feature+Release+2.8)
   - Email reminders before start and after completion of crawls.
   - Collection Management page provides a more direct interface to collection controls including the ability to start crawls manually.
   - One-time and semiannual crawl frequencies added
   - Collections can be browsed either by URL or searched by keyword (NutchWAX)

2.10 [http://webteam.archive.org/confluence/display/ARIH/Archive-It+Feature+Release+2.10](http://webteam.archive.org/confluence/display/ARIH/Archive-It+Feature+Release+2.10)
   - Public and private interfaces allow easy switching between English and Spanish on the public and private Archive-It websites.
   - Institution logos may be uploaded to the public website for branding.
• One or more of eight topics may be assigned to collections to facilitate more effective and focused browsing on Archive-It's public site.
• Seed lists on public collection pages can be edited.
• Users may browse partners, in alphabetical order, by their institution type, or by topic
• More effectively capture and archive videos hosted on www.youtube.com however, some difficulties in playing back videos captured from YouTube continue (if a YouTube page is used as a seed, too many documents may become in scope, so constraints must be added to limit crawling).

2.12 http://webteam.archive.org/confluence/display/ARIH/Archive-It+Feature+Release+2.12
• Partners are able to add a list of hosts to block instead of blocking one host at time
• Extended Crawl Frequencies (Tests, one time, monthly, and quarterly crawls can be extended from 3 to 5 days. Crawls run semi-annually and annually can be extended from 5 to 7 days).
• Partners can now view archived videos much more easily as part of their collections. Re-playing archived videos is an ongoing part of our software development, and this improved access provides partners enhanced access to archived media.

3.0 http://webteam.archive.org/confluence/display/ARIH/Archive-It+Feature+Release+3.0
• Regularly scheduled, twice per month training sessions will be offered.
• Scope-It tool added to help analyze existing and prospective crawl
• Use of the Heritrix de-duplication feature means after an initial crawl is conducted only new or changed content will be saved on subsequent crawls.
• Software upgrade improving overall performance of the NutchWAX search engine as well as providing better ranking and relevancy in the search results.
• Adoption of WARC as the preservation file format. WARC is an extension of the previously used ARC format.

Archive-It Help Wiki http://webteam.archive.org/confluence/display/ARIH/Welcome