



World Wide Web Overload: Archiving a Messy Web

Scott Reed, Internet Archive
Claude Zachary, University of Southern California
James Jacobs, Stanford University
Henry Lowood, Stanford University

SCA Annual General Meeting 2013

April 12, 2013



What is a Web Archive?

A web archive is a collection of archived URLs grouped by theme, event, subject area, or web address.

A web archive contains as much as possible from the original resources and documents the change over time. It is a priority to recreate the same experience a user would have had if they had visited the live site on the day it was archived.



Specific Web Archiving Use Cases

- **Create a thematic/topical web archive**
- Capture **web content that relates to traditional collecting activity** around the same thematic focus
- **Mandate to preserve institutional memory and history**
- **Support an electronic records system** to meet record retentions requirements
- Capture **state/ local agency publications** no longer being deposited in print form, and collect and aggregate **state/ local government websites**
- Closure crawls



Publicly Available Web Archives

www.archive-it.org

The screenshot shows the homepage of archive-it.org. At the top, there is a navigation bar with links for Home, Help, Search, and Contact Us. Below this, a main heading reads "Explore Collections" with a search box. Three featured collections are displayed: "International Wikipedia Archive", "Public Resource Org", and "European Central Bank". Each collection includes a thumbnail image and a brief description. At the bottom, there is a section for "Explore Collecting Organizations" with logos for various institutions.

www.archive.org

The screenshot shows the homepage of archive.org. It features a dark header with the "Internet Archive" logo and navigation links for Web, Video, Texts, Audio, Projects, About, Account, TVNews, and OpenLibrary. A search bar is prominently displayed. Below the search bar, there are sections for "Announcements" and "Web" (280 billion pages). The "Web" section features the Wayback Machine logo and a search input field. Other sections include "Video" (3,175,679 movies) and "Audio" (1,568,752 recordings).

webarchives.cdlib.org

The logo for the web archiving service features a stylized globe icon composed of overlapping squares. To the right of the icon, the text reads "web archiving service" in a bold, sans-serif font. Below this, a tagline states "Capture today's web ° Build tomorrow's archives."

www.webarchive.org.uk/ukwa/

The UK Web Archive banner features the logo "UK WEB ARCHIVE preserving uk websites" on the left. To the right, a horizontal row of ten small thumbnail images shows various websites archived at different dates: Archived August 2005, Archived November 2005, Archived May 2006, Archived June 2007, Archived March 2009, Archived October 2004, Archived March 2005, Archived November 2006, Archived November 2008, and Archived May 2009.



Web Archiving Service: Archive-It

Archive-It is a subscription service deployed in February 2006

- **Web based application** that allows users to create, manage, access and store collections of digital content
- The service is a **fully hosted solution**, and includes access and storage
- **Provides tools for selection and scoping** including cataloging with metadata
- Ability to **capture content using 10 different crawl frequencies**
- Archived content includes: html, videos, audio, social networking, PDF, images, online newspapers
- **Can browse archived content 24 hours after a capture is complete**; and full text search is available within 7 days
- **Restricted access options** are available



How is Archive-It different than the General Archive (www.archive.org)?

Archive-It:

- Focused collections
- Control over scope and frequency
- Technical support
- Content indexed for full text search
- Content cataloged with metadata
- Archived data can be shipped
- Restricted access options
- Access archived data 24 hours later

General Archive:

- One collection
- Snap Shot every 2 months
- Automated
- Search not available
- Cataloging not available
- Shipping not available
- Public access only
- Access archived data 6 months later



The Web is a Mess

Some common challenges for the archivist:

Selection

Which websites does your institution want to capture? How much of those sites?

Linked Content

How much of a outside links and content do you want to capture, if any?

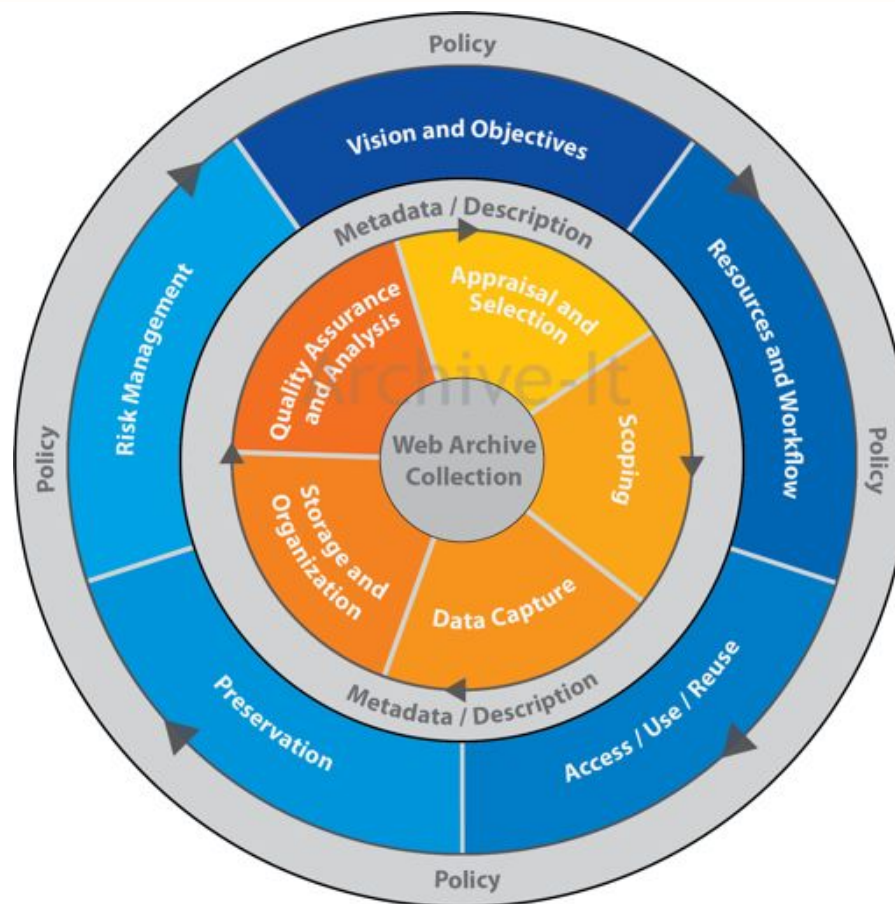
Uncovering a complicated website

content served from multiple hosts, subdomains, poorly constructed etc.





Web Archiving Life Cycle Model



Web Archiving Life Cycle Model white paper available: <http://www.archive-it.org/publications>





Web Archiving Technology

Underlying software tools are ***open source*** and developed by the Internet Archive and IIPC.

- Capture digital content using ***Heritrix*** web crawler.
- View and access captured content using the ***Wayback*** interface.
- Search your archived collections; content is indexed by ***Nutchwax*** (full text search) and ***Solr*** (metadata search) software.
- Store content in ***WARC*** format, an ISO standard for web archiving.



What is a Crawler?

- Crawlers (also known as spiders or robots) are pieces of software that visit websites and index the information included therein (think of Google – it works because of crawlers).
- To archive the Web, Archive-It crawls URLs and captures a copy of the information and files displayed on target websites (PDFs, images, html, etc).
- Can capture an entire site or just one URL, or directory in a site.





How the Crawler Works

- Start with a seed URL
- Check if URL is reachable on live web
- Check embedded content – what does it need to render the page? (CSS, Javascript, Images, etc.)
- Look for links to other pages
- Check if those pages are ‘in scope’, and archive them
- Keep going until either:
 - Can not locate any more links that are in scope
 - Hit the maximum time limit for your crawl



The Web is a Mess

Some common challenges for current web archiving technology:

- **Password protected content**
- **Javascript**
- ***Rich Media*** (videos, Flash content, other interactive media)
- **Social Media** (including Facebook, Twitter, YouTube, etc)
- **“Crawler Traps”**





What makes a site “Archive-Friendly” ?

- We can advocate for more “archive-friendly” websites, those that follow web standards and best practices and are more easily accessed by web archiving technology.
- Tools are being developed and tested in the field to give webmasters and archivists a better sense of what’s a mess, and what’s not.

Checking website <http://www.calarchivists.org/>

Website address:

[Recheck now >](#)

Results

[Summary](#)
[HTML and CSS \(84\)](#)
[HTTP \(2\)](#)
[Media \(9\)](#)
[RSS \(1\)](#)
[Sitemap.xml and Robots.txt \(3\)](#)
[Internet Archive \(1\)](#)
[WARC \(1\)](#)

Overall Rating

64%

Downloads

- One page printable HTML
- Results in EARL XML format

Archivability Facets

Accessibility	91 % (13/14)
Standards Compliance	69 % (11/16)
Performance	100 % (1/1)
Cohesion	8 % (1/12)
Metadata	50 % (1/2)

Website attributes

HTML and CSS	Checking complete! 4 Errors, 15 Warnings, 65 Correct.
HTTP	Checking complete! 2 Correct.
Media	Checking complete! 9 Correct.
RSS	Checking complete! 1 Warnings.
Sitemap.xml and Robots.txt	Checking complete! 2 Errors, 1 Correct.
Internet Archive	Checking complete! 1 Correct.
WARC	Checking complete! 1 Correct.



What makes a site “Archive-Friendly” ?

- Shares many of the same qualities of SEO optimization- ability to be crawled
- Direct links to content
- Logical site structure, reflected in semantic URLs
- Robots.txt files are written to include archive crawlers, and consider which files and directories are necessary to display the site well
- Sitemaps
- If you need to download software to play media in your browser, it may be difficult to play back in archive form



Archive-It Tools

Test Crawls

Sends out crawler but does not archive content

Tells archivist:

- How many URLs and how much content would have been captured. (including lists of all URLs)
- Whether or not URLs or portions of a site were blocked by Robots.txt
- How long the crawl would take (Minutes? Hours? Days?)
- What kinds of content will be captured (videos, images, pdfs, etc)
- How much data would be archived
- What wasn't in scope automatically that needs to be archived?



Archive-It Tools

8 Post Crawl Reports

Host Report: Shows all content discovered in crawl

Host	URLs	Data	New URLs	New Data	Queued	Robots.txt Blocked	Out of Scope
twitter.com	1,642	15.4 MB	1,633	15.4 MB	0	0	1,894
upload.wikimedia.org	1,548	114.9 MB	60	971.3 KB	2,593	0	47
green.blogs.nytimes.com	1,277	42.5 MB	865	42.4 MB	4,335	0	0
dotearth.blogs.nytimes.com	1,258	41.4 MB	883	41.2 MB	4,060	0	0
www.willstegerfoundation.org	1,191	76.4 MB	1,191	76.4 MB	3,058	0	0
www.climatecrisis.net	1,004	15.5 MB	44	272.9 KB	0	0	0
graphics8.nytimes.com	972	98.1 MB	70	1.8 MB	0	1	986
topics.nytimes.com	905	48.4 MB	880	48.0 MB	3,617	437	98
en.wikipedia.org	820	71.3 MB	791	71.2 MB	66,476	662	21,416
www.nature.org	683	47.0 MB	3	63.7 KB	0	1	342
www.globalwarming.org	643	31.3 MB	464	29.2 MB	1,218	0	0
www.climatichelp.com	518	30.5 MB	0	0.0 bytes	0	0	0



Archive-It Tools

Quality Assurance Tool

select the Seed Url you would like to QA. You will see a screenshot of the archived web page, a list of embedded files that are included in this page, and the reasons why there is an issue.

Ignore robots.txt

[Run Patch Crawl](#)

By clicking on the "Run Patch Crawl" you can capture any embedded URLs that were not captured for the seed URLs in your crawl.

Seed URL
http://www.discoverytheater.org/
http://residentassociates.org/
http://artcollectorsprogram.org/
http://civilwarstudies.org/
http://discoverytheater.org/
http://startstudioarts.si.edu/
http://smithsonianassociates.org/
https://artcollectorsprogram.org/
http://smithsonianassociates.org/start.htm
http://1100jefferson.smithsonianassociates.org/
http://smithsonianassociates.org/ticketing/index.z

<http://www.discoverytheater.org/>



- [Full size screenshot #](#)
- [Full size proxy mode screenshot #](#)

http://www.discoverytheater.org/includes/contentslider.js	issue
http://www.discoverytheater.org/includes/scripts/js-image-sli	Capture issue
http://www.discoverytheater.org/static/js/disclaim-element.js	Capture issue
http://www.discoverytheater.org/images/background-homepa	Capture issue
http://www.google-analytics.com/ga.js	Capture issue
http://www.discoverytheater.org/includes/ribbon.png	Capture issue
http://www.discoverytheater.org/includes/loading.gif	Capture issue

[QA This Crawl Again](#)



Archive-It Tools

The Archive-It staff and community of partners





Looking Forward

4.8 release in just a couple of weeks!

- Ability to crawl content behind a username and password
- Additional QA functionality
- IP authentication for Wayback access to archived content at the collection level.
- Remove specific archived content from full text search
- Additional Metadata functionality:
 - Ability to import seed level metadata.
 - Ability to bulk add/edit document metadata
 - Option to include seed level metadata in OAI-PMH feed

+more!



Learn more about Archive-It!

Scott Reed, Partner Specialist
scott@archive.org

Sign up for an informational webinar: archive-it.org/contact-us

Follow us on Twitter: twitter.com/archiveitorg/

Like us on Facebook: www.facebook.com/ArchiveIt

Check out our blog: blog.archive-it.org/



Panelist Contact Information:

Claude Zachary
czachary@usc.edu

James Jacobs
jrjacobs@stanford.edu

Henry Lowood
lowood@stanford.edu