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

INTERNET ARCHIVE

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

www.archive.org

webarchives.cdlib.org

www.webarchive.org.uk/ukwa/

INTERNET ARCHIVE

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

INTERNET ARCHIVE

The Web is a Mess

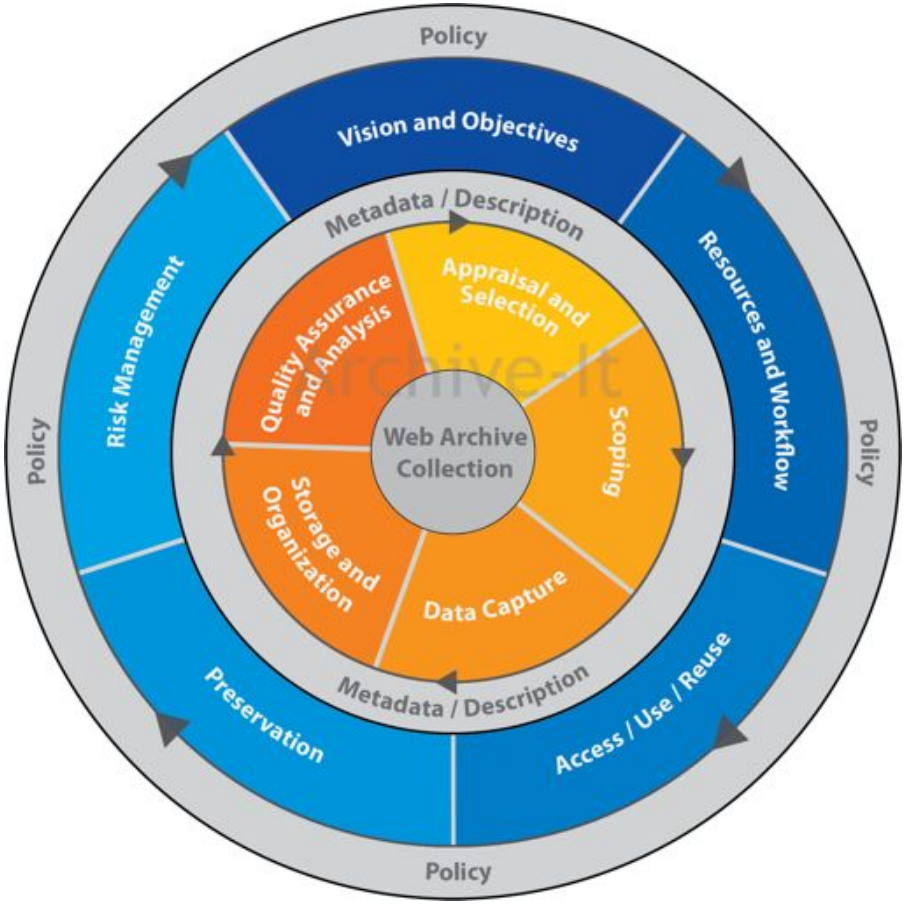
Some common challenges for the archivist:

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

INTERNET ARCHIVE

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.

INTERNET ARCHIVE

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

INTERNET ARCHIVE

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"**

INTERNET ARCHIVE

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.

www.ArchiveReady.com

INTERNET ARCHIVE

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?

INTERNET ARCHIVE

Archive-It Tools

8 Post Crawl Reports

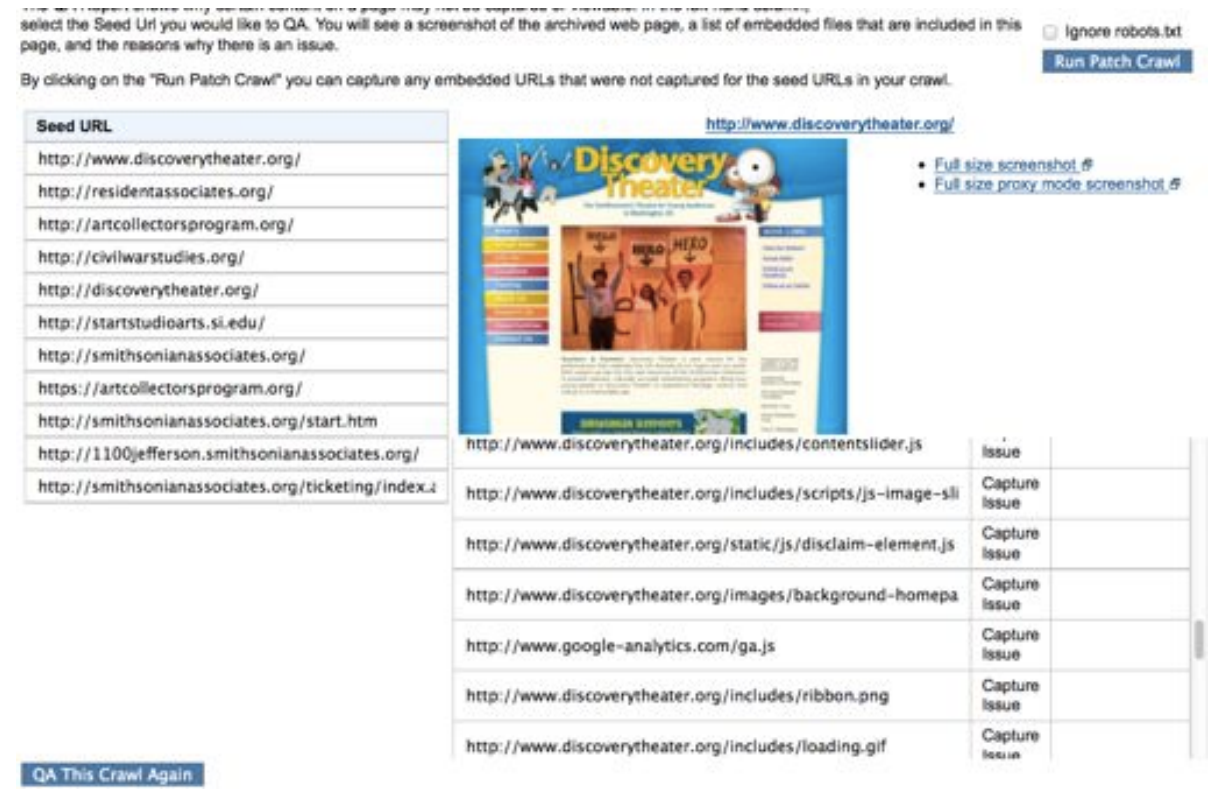
Host Report: Shows all content discovered in crawl

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	1,994
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	683	41.2 MB	4,060	0	0
www.willstegerfoundation.org	1,191	76.4 MB	1,191	76.4 MB	3,059	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	66,478	662	21,418
www.nature.org	693	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
[illegible]	[illegible]	[illegible]	0	0.0 bytes	0	0	0

INTERNET ARCHIVE

Archive-It Tools

Quality Assurance Tool



INTERNET ARCHIVE

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