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
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.
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
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

<table>
<thead>
<tr>
<th>Host</th>
<th>URLs</th>
<th>Data</th>
<th>New URLs</th>
<th>New Data</th>
<th>Queued</th>
<th>Robots.txt Blocked</th>
<th>Out of Scope</th>
</tr>
</thead>
<tbody>
<tr>
<td>twitter.com</td>
<td>1,642</td>
<td>15.4 MB</td>
<td>1,633</td>
<td>15.4 MB</td>
<td>0</td>
<td>0</td>
<td>1,994 MB</td>
</tr>
<tr>
<td>upload.wikimedia.org</td>
<td>1,548</td>
<td>114.9 MB</td>
<td>60</td>
<td>971.3 KB</td>
<td>2,593</td>
<td>0</td>
<td>47 MB</td>
</tr>
<tr>
<td>green.blogs.nytimes.com</td>
<td>1,277</td>
<td>42.5 MB</td>
<td>965</td>
<td>42.4 MB</td>
<td>4,335</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>dotearth.blogs.nytimes.com</td>
<td>1,258</td>
<td>41.4 MB</td>
<td>683</td>
<td>41.2 MB</td>
<td>4,060</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><a href="http://www.willsteferfoundation.org">www.willsteferfoundation.org</a></td>
<td>1,191</td>
<td>76.4 MB</td>
<td>1,191</td>
<td>76.4 MB</td>
<td>3,059</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><a href="http://www.climatecrisis.net">www.climatecrisis.net</a></td>
<td>1,004</td>
<td>15.5 MB</td>
<td>44</td>
<td>272.9 KB</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>graphics8.nytimes.com</td>
<td>972</td>
<td>98.1 MB</td>
<td>70</td>
<td>1.8 MB</td>
<td>0</td>
<td>1</td>
<td>938.0 MB</td>
</tr>
<tr>
<td>topics.nytimes.com</td>
<td>905</td>
<td>48.4 MB</td>
<td>880</td>
<td>48.0 MB</td>
<td>3,617</td>
<td>0</td>
<td>437.0 MB</td>
</tr>
<tr>
<td>en.wikipedia.org</td>
<td>820</td>
<td>71.3 MB</td>
<td>791</td>
<td>71.2 MB</td>
<td>66,478</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><a href="http://www.nature.org">www.nature.org</a></td>
<td>693</td>
<td>47.0 MB</td>
<td>3</td>
<td>63.7 KB</td>
<td>0</td>
<td>1.0</td>
<td>342.0 MB</td>
</tr>
<tr>
<td><a href="http://www.globalwarming.org">www.globalwarming.org</a></td>
<td>643</td>
<td>31.3 MB</td>
<td>464</td>
<td>29.2 MB</td>
<td>1,218</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><a href="http://www.wifiblogs.com">www.wifiblogs.com</a></td>
<td>510</td>
<td>20.6 MB</td>
<td>0</td>
<td>0.0 bytes</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Archive-It Tools

Quality Assurance Tool

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://civiliwarstudies.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.js

QA This Crawl Again

Issue
- Capture Issue
- Capture Issue
- Capture Issue
- Capture Issue
- Capture Issue

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

Full size screenshot
Full size proxy mode screenshot
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/Archivelt
Check out our blog: blog.archive-it.org/
Panelist Contact Information:

Claude Zachary
czachary@usc.edu

James Jacobs
jrjacob@stanford.edu

Henry Lowood
lowood@stanford.edu