Title
Web Archiving Service (WAS) for Library Liaisons

Permalink
https://escholarship.org/uc/item/1sr2w7vs

Authors
Tsang, Daniel C
McKinley, Matthew J

Publication Date
2014-10-27
Web Archiving Service
Collect, manage & preserve websites
A SERVICE OF THE UNIVERSITY OF CALIFORNIA CURATION CENTER — UC3

• Capture content from one or more websites
• Set scope & frequency
• Analyze results
• Sophisticated search & browse interface
Why use WAS?

**Bibliographers**
- Capture(s) as collection development
- Preserve ephemeral or at-risk sites

**Researchers**
- Capture(s) as class learning tool
- Capture(s) for research use
Example WAS Archives

http://webarchives.cdlib.org/a/CivilandHumanRights
Using WAS: Capture

- Seed URLs
- Scope
  - Capture Linked Pages?
- Scheduling (frequency)
  - Daily, Weekly, Monthly, Custom
- Descriptive Data
Using WAS: Managing Sites

- Limiting Results
  - Date of first capture
  - Tag
  - Frequency
  - Keyword (search term)

- Tagging Sites
Using WAS: Comparing, Reporting & QA

Compare captures for Academic Budget (University of California, Irvine) Web Archives

Earlier: 01/17/12 12:24 PM
Scope: Host site only, Files: 29, Duration: 45s

Change Summary
Changed: 1 files (2.5% by size)
New: 23 files (50.1% by size)
Missing: 20 files (40.2% by size)

- Crawl Report
  "Crawl" refers to a specific capture. This brief report provides total job size (in bytes), duration, number of files, and whether the size or duration limits were reached.

- Crawl Log
  The crawl log is the most detailed account of capture activity, providing a separate line of information for every URL attempted. This includes a timestamp for the moment the capture was attempted, a status code indicating whether capture was successful or encountered errors, the document size, the URL of the document, a discovery path code explaining how the document was captured and more.

- Hosts Report
  A useful report if you selected "host + linked pages" as your capture scope. This report tells you every other host name involved in your capture results and how many files each host provided.

- Mimetype Report
  A list of document formats found and their frequency.

- Processors Report
  An activity report for each Hartrix processor used in this capture.

- Response Code Report
  A list of response codes returned during this capture and their frequency.

- Seeds
  The list of starting URLs you selected for capturing.

- Seeds Report
  A list indicating the status of each seed in your capture.

Failed captures
View only the captures that returned only a few files. You may want to check these sites for the presence of a robots.txt file; it is possible that the site owner is preventing capture. A lower than expected number of results may also indicate an error in the seed URL or scope settings.

Captures with fewer than 10 files
These captures still had files remaining to capture when they finished. You may wish to narrow the scope of your capture.

Redirected seed URLs
List the sites that send the crawler to a new URL for the home page. These sites may have out of date seed URLs that are at risk of failing in the future.

Sites with curatorial notes
See all sites where a curator has added a note. (These can be used to document conversations with site owners). You can also look up words in curatorial notes from the Manage Sites screen. Select the "Curator note" radio button for the keyword limit.

Find Recent Captures
Use the date selection tool on the right-hand side of the screen or the Manage Sites or View Capture screens to zero in on sites captured since the last time you did QA on your archive.
WAS Rights Management

• **Section 108 of the Copyright Act**
  • “...Section 108 provides limited exceptions for libraries and archives to make copies in specified instances for preservation, replacement and patron access.”
    - http://www.section108.gov/about.html

• **Section 108 Study Group Report, 2008**
  • “...among its recommendations was that libraries should not have to consult website owners in advance to archive **freely available** public content”
    - http://was.cdlib.org/docs/was_rights_management.pdf

• **That said**: YOU decide to make Capture(s) public/private
Learn more about WAS:

http://webarchives.cdlib.org/learning_center

- Videos
- User Guides
- Rights Management Documents
Selection from 133 sites captured in the Orange County, Web Archive

OC Metro Web Archive
OC Partnership Web Archive
OC Political Web Archive
OC Politics Blog Web Archive
OC Pride Web Archive
OC Talking Points Web Archive
OC Weekly Best of Web Archive
OC Weekly Flipbook Web Archive
OC Weekly in PDF Web Archive
OC Weekly Special Editions Web Archive
OC Weekly Web Archive
Occupy Orange County (Santa Ana) Web Site
Occupy Orange County Web Archive
Occupy UCI! Web Archive
Activity (15 min.)

1. Identify a site to capture (2 min.)

2. Explore site (3 min.)

3. Discuss & Decide on (5 min.):
   - **Scope** (WHAT to capture: 1 URL or many? Links?)
   - **Frequency** (HOW OFTEN to capture: daily? Weekly? Monthly?)

4. Share your site w/ Scope & Frequency decisions (5 min.)
WAS & Digital Scholarship Services

Bibliographers!

Are You or Your Faculty interested in collection building or research using WAS? Schedule a Digital Scholarship Consultation at:

www.lib.uci.edu/dss

Feedback Google Form: http://goo.gl/forms/Jyd41UZu27