



Tool Academy: Web Archiving

Nicholas Taylor
[@nullhandle](#)

Digital Cultural Heritage DC Meetup
December 20, 2012

"cobwebbed screw driver" by Flickr user Colby Gutierrez-Kraybill under CC BY 2.0

what does a web archive look like?

W/ARC (web archive container format)



“Buckets and Buckets” by Flickr user [Josh Kenzer](#) under [CC BY-NC-SA 2.0](#)

flat files, directory tree



“Files” by Flickr user [Artform Canada](#) under [CC BY-NC-ND 2.0](#)

A photograph showing several rectangular crab traps made of black mesh and reinforced with thick blue and white ropes. The traps are set in a field of tall green grass and weeds. One trap in the foreground is partially open, revealing its interior. A black plastic tarp is visible on the ground to the left. The text "CAPTURE TOOLS" is overlaid in white on a black rectangular background in the lower-left quadrant.

CAPTURE TOOLS

HTTrack

<http://www.httrack.com/>

- small-scale website copier
- recreates website structure as filesystem hierarchy
- Windows GUI or CLI
- *nix local web service or CLI



Heritrix

<https://webarchive.jira.com/wiki/display/Heritrix/Heritrix>

- web-scale archival crawler
- WARC output
- configure and run through web service
- Java app, runs best on *nix

The logo for Heritrix, featuring the word "HERITRIX" in a bold, white, sans-serif font. The letters are slightly shadowed, giving it a 3D appearance as if it's floating above a dark surface.

Wget

http://archiveteam.org/index.php?title=Wget_with_WARC_output

- retrieve Internet-accessible files
- supports WARC output
- CLI utility

WARCreate

<http://matkelly.com/warcreate/>

- archive single webpage(?) to WARC
- Chrome extension
- no production release yet
- may eventually bundle a self-contained Wayback Machine



Warrick

<http://warrick.cs.odu.edu/>

- reconstruct website from web archives
- uses Memento protocol
- web service or downloadable Perl script



ArchiveFacebook

<https://addons.mozilla.org/en-US/firefox/addon/archivefacebook/>

- archive an individual (authenticated) Facebook profile
- Firefox add-on





REPLAY

REW

REPLAY TOOLS



Wayback Machine

<https://github.com/internetarchive/wayback>

- replay web resources stored in WARC and ARC files
- web service provided by Internet Archive
- also, downloadable software package
- Java app (Tomcat), runs best on *nix

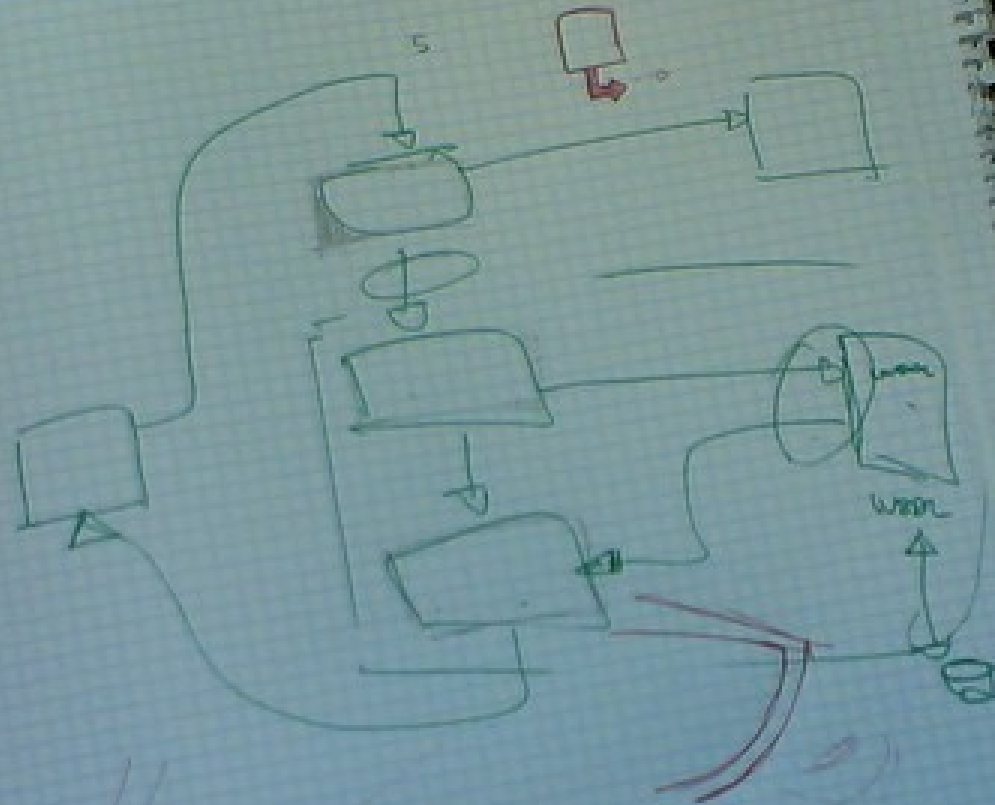


MementoFox

<https://addons.mozilla.org/en-us/firefox/addon/mementofox/>

- federated discovery of web archive resources
- uses **Memento** protocol
- utility limited by paucity of aggregated indexes





WORKFLOW TOOLS

Web Curator Tool

<http://webcurator.sourceforge.net/>

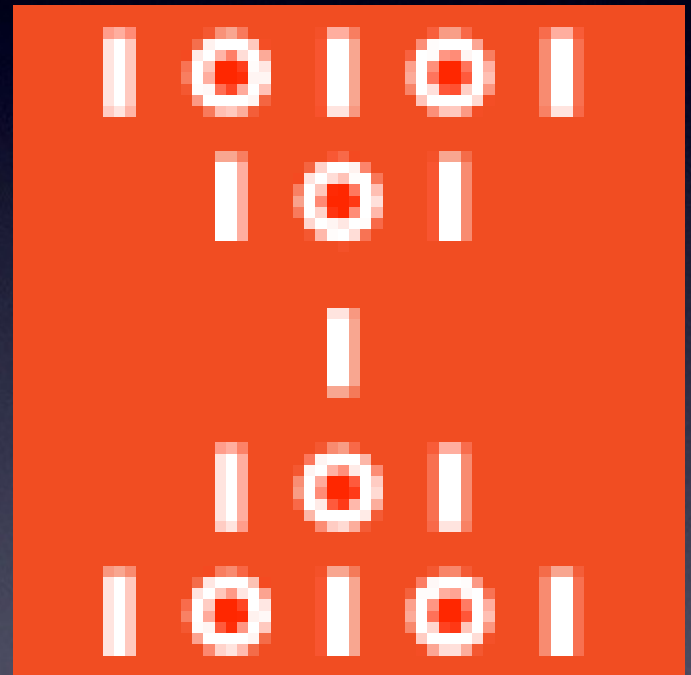
- permissioning, job scheduling, harvesting, quality review, storing descriptive metadata
- coupled with Heritrix v1.0
- Java app (Tomcat)



NetarchiveSuite

<https://sbforge.org/display/NAS/NetarchiveSuite>

- job scheduling, data transfer to preservation system, proxy replay
- built for national domain crawls
- coupled with Heritrix v1.0
- Java app (JMS)



CINCH

<http://cinch.nclive.org/Cinch/>

- batch retrieval of Internet-accessible documents and transfer to preservation system
- web service for NC state government
- also, downloadable software package
- runs on *nix





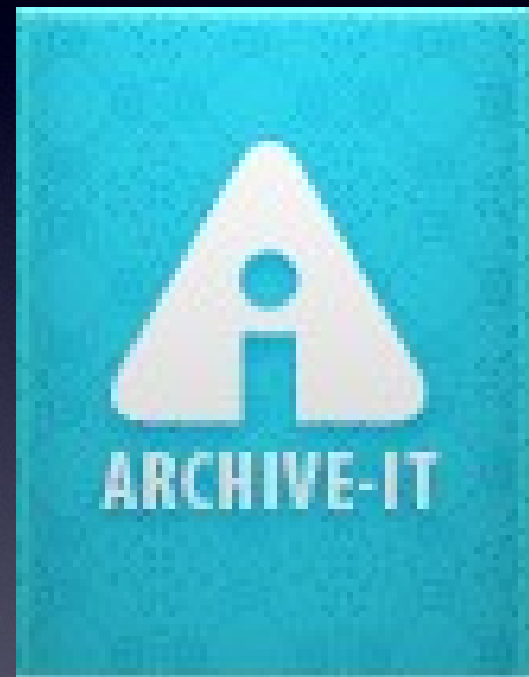
HOSTED SERVICES

"Services" by Flickr user [spodzone](#) under [CC BY-NC-ND 2.0](#)

Archive-It

<http://www.archive-it.org/>

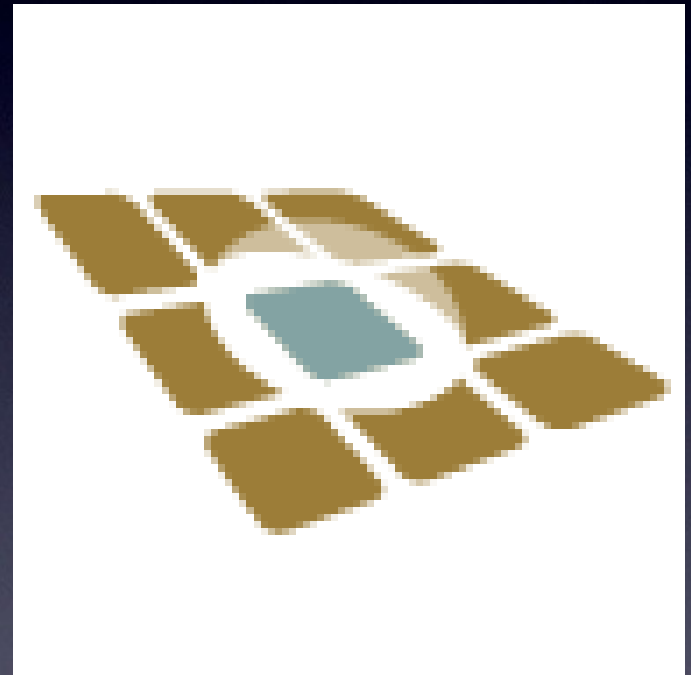
- integrated web archiving platform
- uses Heritrix and Wayback Machine
- contract service provided by Internet Archive



Web Archiving Service

<http://webarchives.cdlib.org/was>

- integrated web archiving platform
- uses Heritrix and Wayback Machine
- contract service provided by California Digital Library



FILE UTILITIES



"Begin at the Beginning" by Flickr user kate e. did under CC BY-NC-SA 2.0

HTTrack2Arc

<http://code.google.com/p/httrack2arc/>

- convert HTTrack output to ARC format
- CLI Java utility

warc-tools

<http://code.hanzoarchives.com/warc-tools>

- parse and re-write WARC files
- convert ARC files to WARC files
- no production release yet
- CLI Python utilities

Web Archive Transformation (WAT) Utilities

- extract metadata from WARC_s for data analysis
- read data from local, http, or hdfs-accessible W/ARC_s
- output JSON

<https://webarchive.jira.com/wiki/display/lre+search/Web+Archive+Transformation+%28WAT%29+Specification,+Utilities,+and+Usage+Overview>

thank you!

Nicholas Taylor
@nullhandle