Technology Module: Web Archiving

@digitalPOWRR

Expected Outcomes

- ✓ **Become familiarized with the practice of web archiving**

- ✓ **Become familiar with common terminology**

- ✓ **Learn about common tools/services currently available to perform this work**

- ✓ **Consider a scenario provided by the instructor concerning a use case for using WebRecorder in an archival setting**

- ✓ **Use WebRecorder to capture several websites**

Web Archiving Overview

- Process of collecting portions of the world wide web to ensure information is preserved in an archive for future researchers.

- Typically employ "web crawlers" for scheduled, automated capture.

- Internet Archive began crawling in 1996.

- Wayback Machine launched 2001 - making crawls publicly available.

- Bulk archiving requires special software for capture and use.

- Various kinds of web content can be captured, depending on particular needs.

- Web ARChive format (WARC) is now an ISO Standard, used by LOC, de facto preservation standard.

Web Archiving Software/Services Examples

Archive-It (Internet Archive's paid service)

Preservica (web archiving component built in)

ArchiveFacebook (Firefox extension)

DocNow (suite of Twitter-specific archiving tools)

Web Recorder

Web Curator Tool

Heritrix

HTTrack

NutchWAX

WAIL

WARCreate

wget

More on Software/Tools

Web archiving software can cover various aspects:
- ➢ creation of the archived content
- ➢ scheduling crawls
- ➢ indexing/searching of the content
- ➢ viewing the content
- ➢ making that content available to the public.

Some software only performs one function. IIPC divides tools into the following categories: Acquisition, Replay, Search & Discovery, Analysis, Utilities.

Most institutions who do web archiving at a large scale subscribe to Archive It or use a combination of open source tools to build their own service.

"Good Enough" Web Archiving

Before embarking on a web archiving endeavor, it's important to consider the following questions:

➢ What volume of web content do I need to archive?

➢ At what level do I need to archive content? Entire websites? A page here or there?

➢ Do I need to replicate/capture the appearance and behavior of the site? OR just scrape content from it?

➢ How often do I need to capture the content on the site/pages? Do I really need automation or can I do it manually?
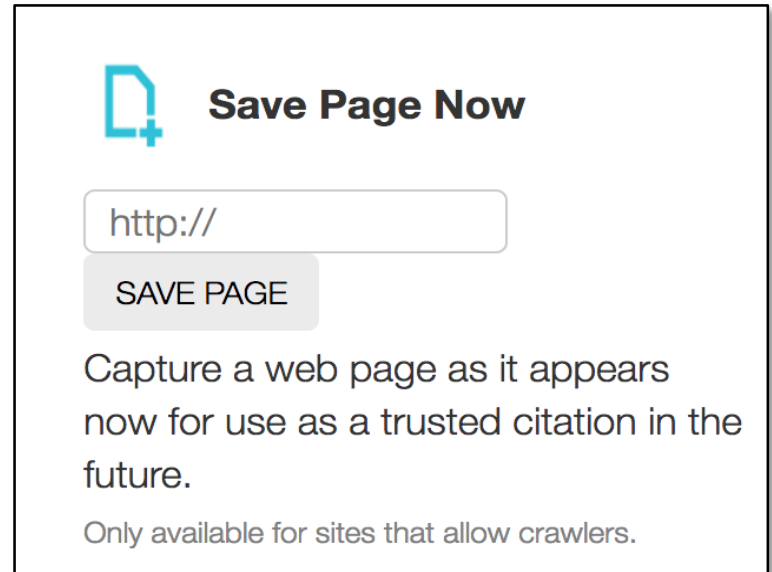
"Good Enough" Web Archiving

If you only need to save a webpage here or there, you have a couple options.

Wayback Machine's "Save Page Now" feature (stores copy in the Wayback Machine). You can save the resulting archived link for your own use in the future.

Available at *https://archive.org/web/*, or you can also download the Save Page Now browser extension for Chrome.

Archive.is is a similar service, but allows you to download a zip file of the site you saved.

Save Page Now

http://

SAVE PAGE

Capture a web page as it appears now for use as a trusted citation in the future.

Only available for sites that allow crawlers.

"Good Enough" Web Archiving

➢ You can save the page(s) as HTML or PDF/A.

➢ If you save to PDF, the formatting may be compromised, and some aspects of dynamic content will not be captured.

➢ If you save HTML, your browser will also save associated CSS and Javascript files. Can be a pain to save/organize.

➢ Apps like Sitesucker will download the entire contents of a website (including media) & replicate the directory structure for you.

➢ This can be time consuming if you're doing a lot of sites.

➢ Using HTML versions of websites can be cumbersome.

Potential "Good Enough" Solution: Webrecorder

➢ Developed by Rhizome, the born-digital art organization.

➢ Free and easy to use

➢ It doesn't "crawl" - it *records* - records the dynamic web, live as your view it.

➢ This means that **anything** you want to save in your archive needs to be opened/played, it does not open/play automatically. This includes videos.

➢ They will host 5 gb of your recordings/WARCs for free, but you must create free account. OR you can download the WARCs you create locally.

➢ They also have a free Web Archive Viewer app, that plays your WARCs.

Why Webrecorder?

➢ The Archivist does not anticipate needing to archive a tremendous amount of web content at this time.

➢ The Archivist feels comfortable enough creating web archives "on the fly" - when special events come, or on a manageable schedule (like, archiving major portions of their Uni website every semester.)

➢ The Archivist anticipates needing to crawl social media sites.

➢ The Archivist does not want to install anything that requires sysadmin knowledge.

➢ The Archivist does not have a budget to pay for the Archive It service.

➢ The Archivist likes the idea of creating a WARC which will ensure she can later use it in 3rd party applications. She also likes that the WARC will contain multiple pages/sites relating to a particular event

Web Archiving Case Study

The Naropa University Archivist is contacted by a staff member from the University's internal Development office, looking for information on alumni donations made for the 40th Anniversary that was celebrated in 2014. The Archivist looks in the usual places to find mention of the event (news releases, etc), but is unable to locate anything.

After some head scratching, she feels a sinking feeling upon realizing that the Communications office had stopped sending the Archives formal press releases, and instead published this information on their website directly, removing it after a period of 6 months.

The archivist realizes that the campus website has become a documentary black hole.

What can the archivist do to start to plug this gaping hole?

Naropa University 40th Anniversary URL's to capture

1. http://www.naropa.edu/about-naropa/events/40.php

2. http://www.naropa.edu/media/press-releases/press-2014/naropa-university-day.php

3. https://www.facebook.com/NaropaUniversity/photos/a.126003793681.106597.54736648681/

   10152056106173682/?type=3&hc_ref=PAGES_TIMELINE

4. https://www.buddhistdoor.net/news/naropa-university-celebrates-40th-anniversary

5. http://www.dailycamera.com/news/boulder/ci_26522937/boulders-naropa-celebrates-40-

   years-contemplative-education

6. https://dc.shambhala.org/2014/11/30/radical-compassion-report-naropas-40th-anniversary/

7. https://www.poets.org/poetsorg/stanza/celebration-naropas-40th-anniversary

8. http://litseen.com/jack-kerouac-school-of-disembodied-poetics-40th-anniversary/

9. https://www.centerforthehumanities.org/programming/naropa-at-40

10. http://www.beatdom.com/naropa-turns-40/

11. https://twitter.com/search?q=naropa%2040th%20anniversary&src=typd

12. https://www.facebook.com/NaropaUniversity/

Before Using the Tool – Do Some Scoping!

Regardless of the tool you use to create web archives, it's important to know exactly what you are capturing, and how much is there.

Scoping is the term most frequently used when talking about what we tell a crawler to capture and what not to capture.

With Web Recorder, YOU are essentially the crawler. Web Recorder will only record what you show it. Therefore, you want to know if there is embedded content you should be capturing, or if there are internal links you should be following.

Using Webrecorder

1. Open Chrome or Firefox
2. Go to https://webrecorder.io/
3. Choose a name for your recording
4. Paste in the first URL from our list in the "URL to Record" box
5. Press "Record"
6. When you get to the page, scroll through it. Mouse over any animations you want to capture.
7. Press play on any videos, sound bites, or navigate through any photo galleries. Remember, whatever you click on gets recorded and added to your archive.
8. When you're ready to add the next URL, click on "Temporary Collection" in the upper left corner. Click "NEW" under Recordings. Repeat until you're done!

Viewing Your Web Archive

1. Click on "Temporary Collection", and then click the "Download Collection" button in the upper left corner.
2. Your browser will download a WARC file of all the material you recorded.
3. Open Webrecorder Player
4. Click on "Load Web Archives"
5. Select your unzipped WARC file
6. Have a look around - see what you recorded and what you didn't.
7. If you were going to save the file locally, you can also rename your WARC to something more meaningful to you and add it to your own local storage.

Technology Module: Web Archiving

QUESTIONS?