Sustainable Web Archiving at Scale: An Introduction

DPOE-N
Digital Preservation Outreach & Education Network

June 21-22, 2021
Day One Recap

Introduction to Web Archiving: Concepts, Policy and Practice

- What is web archiving?
- Collecting objectives/scope
- Copyright, permissions & ethics
- Capacity considerations & description/metadata
- Exercises
Day Two Agenda

Tools + Technology

Overview, demos and hands-on exercises:

- Tools and replay mechanisms
- Internet Archive + Archive-It
- Conifer/Webrecorder
- Other tools
Introduction to Tools & Terms
Technology Stack: A Moving Target

Simple Net Art Diagram

MTAA ca. 1997
Open Source vs. Subscription Services

Open Source
- Free to use
- Technical expertise
  - IT staff
  - Training
- Professional & enthusiast communities

Subscription
- Monthly/Annual costs
- Automated software updates
- IT support/Terms of Service
Community

Wide Adoption
- Less likely abandoned
- More likely to be upgraded
- Shared standards & expertise

Low Adoption
- Less incentive to continue support
- May be isolated from community of practice
Community

- Webrecorder Monthly Community Calls
- Archive-It Quarterly Calls/User Group Meetings
- ART|WARC
- Society of American Archivists, Web Archiving Section
Local vs. Cloud-based

Archive-it / Conifer
- Automatic software updates, web hosting

ArchiveWeb.page
- Web archives stored in Chrome browser, offline/p2p

ReplayWeb.page
- HTTP, S3, IPFS, and GoogleDrive storage
Local vs. Cloud-based

ArchiveWeb.page App / ReplayWeb.page App

- WARC and WACZ files stored locally, offering a layer of security for sensitive materials

Heritrix / Browsertrix Crawler

- Web crawls stored to local directories
- Command line, higher technical threshold
Storage

Capacity Considerations

- Frequency
- Social media
- Video & embedded media
- Data de-duplication
Automated, Semi-automated, & Manual

Automated

Archive-It
  - User-friendly web crawl scheduling & scoping

Heritrix
  - Command line tool, several simultaneous crawls
  - Respects robots.txt exclusions
Automated, Semi-automated, & Manual
Symmetrical Archiving

Conifer & ArchiveWeb.page/ReplayWeb.page

- User-friendly, browser-based
- Manual high-fidelity collecting
- Time intensive
- Auto-pilot: autoscroll, video replay
Automated, Semi-automated, & Manual

Symmetrical Archiving + Scalability

Browsertrix Crawler
- Browser-based high fidelity crawler
- Supports custom browser behavior

Archive-It: Umbra and Brozzler
- Rich-media and dynamic web behaviors and content
Significant Properties

- Aesthetic properties
- Custom behaviors
- Embedded media as information
- Interactive features of data-driven applications
Remote/Emulated Browsers

Conifer
- Pre-configured remote browsers
- Chrome/Firefox, Java/Flash

OldWeb.today
- Emulated legacy browsers
- No capture capabilities
Playback/Rendering/Access

Archive-It: Wayback
- Seed URLs as access point
- Time-stamped links

Conifer / ArchiveWeb.page
- Browser-based replay, identical to capture
- Conifer: Collections, lists, descriptions
- ArchiveWeb.page: full text search
Dead Ends/Patching

Archive-It / Conifer
- Patching tools unite missing content

Conifer
- Importing from open web archives

ArchiveWeb.page
- Page-level archiving
- Export selected pages as single WARC/WACZ
Questions?
Internet Archive & Archive-It: Overview/Demo
Internet Archive (archive.org)
Wayback Machine (archive.org/web)
Search the history of over 555 billion web pages on the Internet.

WayBackMachine

Enter URL or keywords

https://www.dpoenetwork/workshops/

Latest
Show All

Hrm.
The Wayback Machine has not archived that URL.

This page is available on the web!
Help make the Wayback Machine more complete!
Save this URL in the Wayback Machine

Click here to search for all archived pages under https://www.dpoenetwork/workshops/.
Save Page Now (web.archive.org/save)
Saving page https://www.dpoe.network/workshops/

The same snapshot had been made 1 minute and 10 seconds ago. We only allow new captures of the same URL every 30 minutes.

A snapshot was captured. Visit page: /web/20210412165535/https://www.dpoe.network/workshops/

https://www.dpoe.network/workshops/
https://www.dpoe.network/wp-includes/css/dist/block-library/style.min.css?ver=5.7
https://www.dpoe.network/wp-content/plugins/contact-widgets/assets/css/font-awesome.min.css?ver=4.7.0
https://www.dpoe.network/wp-content/plugins/otter-blocks/vendor/codeinwp/gutenberg-

Downloaded elements: 41

Return to Save Page Now
Workshops

Sustainable Web Archiving at Scale: An Introduction

**Free Workshop** Thursday, April 15th & Friday, April 16th: 2:00-5:00pm EST

Apply now!

**Description:**
As cultural production and communication have moved online, the need for archivists to document history in real time has become increasingly clear and urgent. Yet the scale and complexity of online information and media can be daunting, even as web archiving tools and practices continue to evolve. And while large institutions are able to devote staff to keeping pace with new technologies and their attendant ethical concerns, small-to-mid sized organizations frequently rely on archivists who already handle many other responsibilities for this work. How does one design a web archiving...

**Details:**
This web archiving virtual workshop is being hosted by the Digital Preservation Outreach & Education Network (DPOE-N) in partnership with the Pratt Institute School of Information. It is being offered **tuition-free**, thanks to generous support from the Andrew W. Mellon Foundation.

The application for this program is now closed.

The deadline to apply was Thursday, March 25.
Live Web Proxy Crawls
Content crawled via the Wayback Machine Live Proxy mostly by the Save Page plugin.
Now feature on web.archive.org.
Liveweb proxy is a component of Internet Archive’s wayback machine project.
The liveweb proxy captures the content of a web page in real time, archives it.

Collection: Live Web Proxy Crawls
Content crawled via the Wayback Machine Live Proxy mostly by the Save Page plugin.
Now feature on web.archive.org.
Liveweb proxy is a component of Internet Archive’s wayback machine project.
The liveweb proxy captures the content of a web page in real time, archives it.

TIMESTAMPS
https://www.dpoenetwork/wp-content/plugins/ultimate-social-media-icons/images/icons_theme/default/default_facebook.png - 1 month 4 days
https://stats.wp.com/e-202115.js - 16 hours 55 minutes
https://connect.facebook.net/en_US/sdk.js#hash=1fe2e418acb320f5bb379a5f27352ff5&ua=modern_es6 - 1 hour 20 minutes
https://www.dpoenetwork/wp-content/uploads/2021/03/workshop-banner@2x-1024x178.png + 11 minutes 23 seconds
https://connect.facebook.net/en_US/sdk.js + 8 minutes 55 seconds
https://www.dpoenetwork/wp-includes/js/comment-reply.min.js?ver=5.7 + 5 seconds
https://www.dpoenetwork/wp-includes/js/wp-embed.min.js?ver=5.7 + 5 seconds
https://www.dpoenetwork/wp-includes/js/wp-emoji-release.min.js?ver=5.7 + 5 seconds
https://www.dpoenetwork/wp-content/uploads/2021/03/Sumitra.png + 4 seconds
https://www.dpoenetwork/wp-content/uploads/2021/03/Anigo.png + 4 seconds
https://www.dpoenetwork/wp-content/plugins/ultimate-social-media-icons/js/shuffle/modernizr.custom.min.js?ver=5.7 + 4 seconds
https://www.dpoenetwork/wp-content/plugins/ultimate-social-media-icons/js/shuffle/modernizr.custom.min.js?ver=5.7 + 4 seconds
https://www.dpoenetwork/wp-content/plugins/ultimate-social-media-icons/js/custom.js?ver=5.7 + 4 seconds
https://www.dpoenetwork/wp-content/themes/neve/assets/js/build/modern/frontend.js?ver=2.7.3 + 4 seconds
https://www.dpoenetwork/wp-includes/js/jquery/ui/core.min.js?ver=1.12.1 + 4 seconds
https://www.dpoenetwork/wp-content/plugins/ultimate-social-media-icons/js/shuffle/random-shuffle.min.js?ver=5.7 + 4 seconds
https://www.dpoenetwork/wp-content/plugins/ultimate-social-media-icons/images/icons_theme/default/default_twitter.png + 1 seconds
Wayback Machine browser extensions, apps and add-ons

Chrome Extension
Firefox Add On
Safari Extension
iOS App
Android App
Archive-It (archive-it.org)
DPOE-N Workshops

Current Subscription

0 bytes
Archived

Data Budget Usage

Current Subscription Details
Data Budget:
New Documents: 0
New Crawled Data: 0 bytes
New Uploaded Data: 0 bytes
Subscription started: Apr 6, 2021
Subscription ends: Jul 6, 2021
Member Since: Apr 6, 2021

Past Subscription Totals
Creating New Collections
Welcome to your new collection

**Take a Tour**
New to Archive-It? Take a moment to learn more about the collection area.

**Add Seeds**
Your collection is empty. Add some seeds to your collection and get crawling!

**Delete Collection**
Created this collection by mistake? Empty collections can be deleted.

**Collection Settings**
- Public
- Active

**Scheduled Crawls**
- Frequency
- Active Seeds
- Next Crawl
- Last Crawl
- Time Limit
- Data Limit
- Doc. Limit
- Edit Schedule

No Scheduled Crawls.
Adding Seeds to a Collection
Add Seeds

Enter one seed URL per line below to add them to this collection.

- https://www.dpoе.network/

Access: Public
Frequency: One-Time
Seed Type: Standard

[Cancel] [Add Seeds]
<table>
<thead>
<tr>
<th>Seed URL</th>
<th>Status</th>
<th>Frequency</th>
<th>Type</th>
<th>Access</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="https://www.dpoenetwork/">https://www.dpoenetwork/</a></td>
<td>Active</td>
<td>One-Time</td>
<td>Standard</td>
<td>Public</td>
</tr>
</tbody>
</table>
Adding Metadata to Seed URLs
Initiating Crawls
Run Crawl

Please select options below for a test crawl or one-time crawl of the selected seeds.

Crawl Type
- One-Time Crawl
- Test Crawl

Doc. Limit: Whole Number, e.g. 10000

Data Limit: 1

Time Limit: 1 Hour

Crawl PDFs Only

Crawling Technology
- Standard
- Brozzler

1 selected seed will be crawled.
DPOE-N Workshop Test Collection

Select two crawls to compare.

<table>
<thead>
<tr>
<th>Overview</th>
<th>Seeds</th>
<th>Crawls</th>
<th>Collection Scope</th>
<th>Metadata</th>
<th>Wayback QA</th>
<th>Upload WARC</th>
<th>Scheduled Crawl</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Crawl Reports</td>
<td>Current Crawls</td>
<td></td>
<td></td>
<td></td>
<td>Test Crawls</td>
</tr>
</tbody>
</table>

Test Crawl List (1 Test Crawl)

Type to Filter Test Crawls

<table>
<thead>
<tr>
<th>Crawl ID</th>
<th>Started</th>
<th>Completed</th>
<th>Status</th>
<th>New Data</th>
<th>Docs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1397678</td>
<td>Apr 12, 2021</td>
<td>In progress (new)</td>
<td>0 bytes</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Crawl Reports
# DPOE-N Workshop Test Collection

**Test Crawl:** 1397678  |  **Started:** April 12, 2021 11:24 AM Eastern Daylight Time  |  **Completed:** April 12, 2021 12:24 PM Eastern Daylight Time

## Scoping Rules for this Crawl
- **Crawl Limits:** [2]  
- **Collection Scope Rules:** [4]  
- **Seed Scope Rules:** [6]

### Seed List (1 Seed)

<table>
<thead>
<tr>
<th>Seed URL</th>
<th>Seed Type</th>
<th>Seed Status</th>
<th>Docs</th>
<th>New Docs</th>
<th>Data</th>
<th>New Data</th>
<th>Seed</th>
<th>Wayback Link</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="https://www.dpoe.network/">https://www.dpoe.network/</a></td>
<td>Standard</td>
<td>Crawled</td>
<td>995</td>
<td>383</td>
<td>33.6 MB</td>
<td>15.3 MB</td>
<td>Seed</td>
<td>Wayback &gt;</td>
</tr>
</tbody>
</table>
Digital Preservation Outreach & Education Network

We support digital preservation education and outreach in the nation's libraries, archives and museums.
https://archive-it.org/organizations/2027
Questions?
Conifer

- Open source, high-fidelity service
- Online
- Free up to 5 GB
- Service plans for 40+ GB
Conifer: Background

- Previously the Webrecorder.io service
- Designed with net artists in mind
- High-fidelity capture of dynamic web
- Conifer now under Rhizome only
Conifer: Sessions

- Individual interactions w the web at time of capture
- Replicates any captured experience in any order
- Bound set: exactly what was clicked on or interacted with—nothing else
- Accessed by URL but individual URLs cannot be removed
- Sessions are basis for Collections
# Conifer: Sessions

<table>
<thead>
<tr>
<th>Session Notes</th>
<th>Session Pages</th>
<th>Date/Time</th>
<th>Pages</th>
<th>Duration</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Add notes about this session, visible only to you.</td>
<td>- National Forum on Ethics and Archiving the Web: Western Europe, 2021, 10:30-12:30</td>
<td>4/15/2021, 10:53:15 PM</td>
<td>1</td>
<td>2 mins</td>
<td>2.59 MB</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>1 min</td>
<td>67.32 KB</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4/12/2021, 5:30:25 PM</td>
<td>3</td>
<td>1 min</td>
<td>415.10 MB</td>
</tr>
</tbody>
</table>

**Delete**

**Download**
Conifer: Collections

- Any number of sessions.
- Sessions seamlessly interact with each other
- Public Access:
  - Title & description
  - Lists
Conifer: Collection Manager
Conifer: Collection Cover

DPOE-N Workshop
by amcCarthy

Overview Browse All

- Soap
- Library
- Instagram
- Rhizome
- EAW Vimeo
- Lisa
- Simpson
- GIF /
- Copyright
- on the Web

DPOE-N Workshop
This collection demonstrates the types of media which can be captured using Conifer and approaches to designing a successful capture strategy for your own content.

Soap Library Instagram

1. Soap Library (@soaplibrary) • Instagram photos and videos
   https://www.instagram.com/soaplibrary/
   4/12/2021, 4:16:42 PM
Conifer: Remote Pre-configured Browsers

- Older versions of Firefox and Chrome
- Java and Flash plug-ins
- May reveal content that is not compatible with modern browsers
- Test in advance
- Capture browser = Replay browser
Conifer: Pre-configured browsers
Conifer: Iterative capturing

- Individual URLs can **not** be removed from a session
- Review in advance for pages/media that may cause issues
- If an error is encountered, file a bug report and delete the session from the collection
Conifer: Patching & Importing

Patching
- Browse mode, behaves as it would offline
- Missing content prompts error message
- Patching automatically starts a new session from missing URL that is added to the collection

Importing/Extraction
- Patches content from open web archives
Conifer: Patching & Importing
Webrecorder

ReplayWeb.page + ReplayWeb.page App
- Provides a web archive replay system as a single web site (which also works offline)
- Allows users to view web archives from anywhere, including local computer or even Google Drive.

ArchiveWeb.page + ArchiveWeb.page App
- Interactive high-fidelity Chrome extension and standalone desktop app
- Allows archiving interactively as you browse
Webrecorder

Browsertrix Crawler

- Browser-based high-fidelity crawling system, designed to run a single crawl in a single Docker container.
- Combines the scalability of web crawling with the fidelity of symmetrical archiving.
- Requires basic familiarity with a command-line and Docker to run crawls.
Questions?
Additional Tools to Explore

- Social Feed Manager
- Perma.cc
- YouTube-DL
- Documenting the Now
Social Feed Manager

https://gwu-libraries.github.io/sfm-ui/
Perma.cc

Websites change. Perma Links don’t.

Perma.cc helps scholars, journals, courts, and others create permanent records of the web sources they cite.

Perma.cc is simple, easy to use, and is built and supported by libraries.

Sign up and use Perma.cc

How can my library get involved?
Perma.cc

Create an individual account

Perma.cc for individual users

Anyone can create an individual Perma.cc account, which will allow you to create records to be preserved by The Harvard Law School Library. Just complete this form to get started.

New users are able to create ten free links on a trial basis. Once you’ve used your trial, individuals not affiliated with a registrar must sign up for a paid subscription.

Many organizations qualify for free, unlimited service. To see if your organization qualifies, check out our accounts page.

To learn more about how Perma.cc works, please review our user guide.
YouTube-DL
youtube-dl.org

- Open source command line tool for harvesting video and audio from:
  - YouTube
  - 1,000+ other video hosting websites
Documenting the Now: Tools
docnow.io

DocNow
Appraise Twitter for archival and research collections.

Tweet Catalog
A catalog of publicly shared tweet ID sets. Add yours!

Hydrator
"Rehydrate" tweet ID sets into tweets with metadata.

Social Humans Labels
Labels for ethically describing and sharing social media data.

Twarc
Archive Twitter JSON using this command line tool.

Diff Engine
Track changes in news articles through RSS feeds.
Questions?
Break: 10 minutes
Hands-on exercises: Breakout rooms
Discussion/Questions
Roadmap

SCOPE

BUDGET & STAFF

COLLECTING

COLLECTING POLICY

SOFTWARE & STORAGE
Thank you!