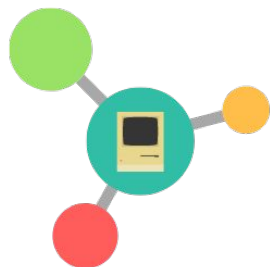


Sustainable Web Archiving at Scale: An Introduction



DPOE-N

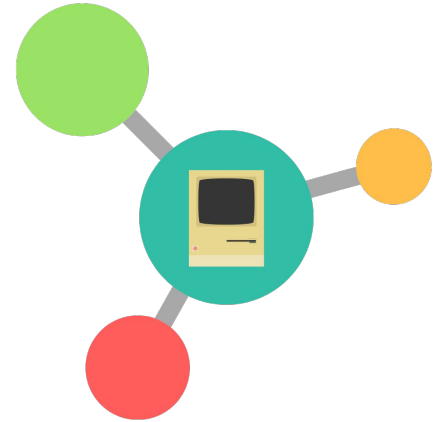
Digital Preservation Outreach
& Education Network

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

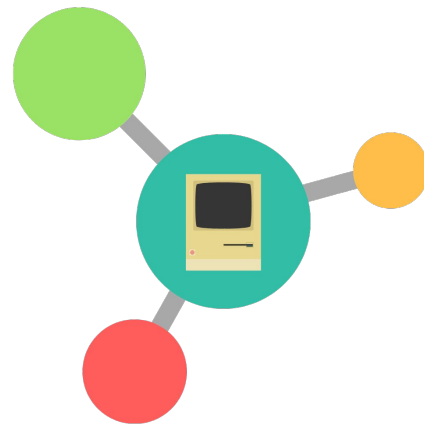


Day Two Agenda

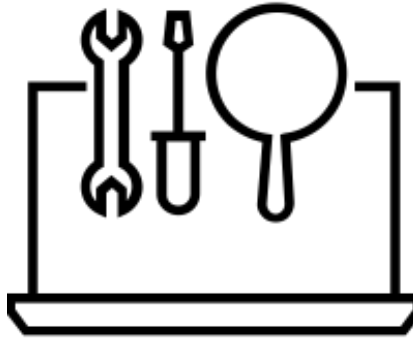
Tools + Technology

Overview, demos and hands-on exercises:

- Tools and replay mechanisms
- Internet Archive + Archive-It
- Conifer/Webrecorder
- Other tools

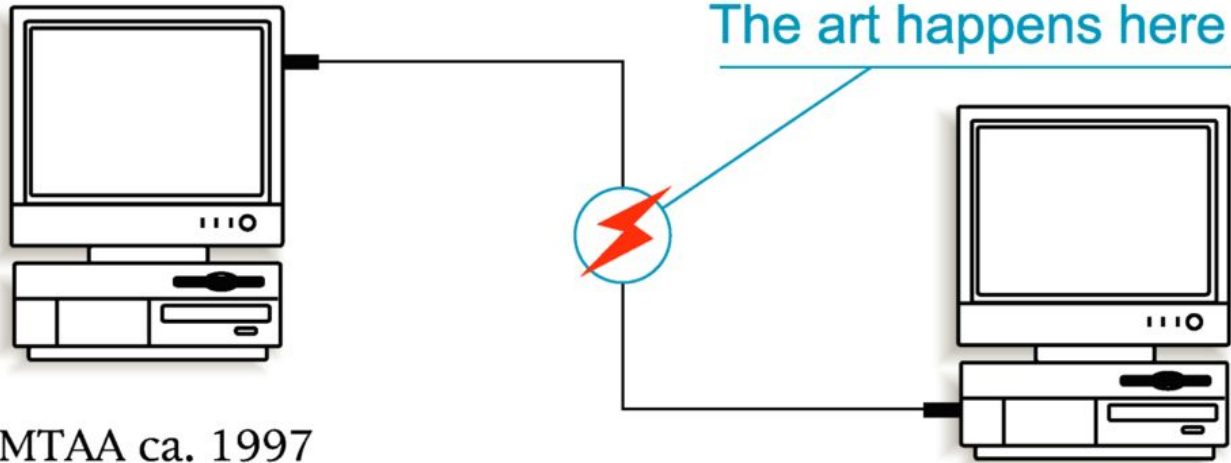


Introduction to Tools & Terms



Technology Stack: A Moving Target

Simple Net Art Diagram



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

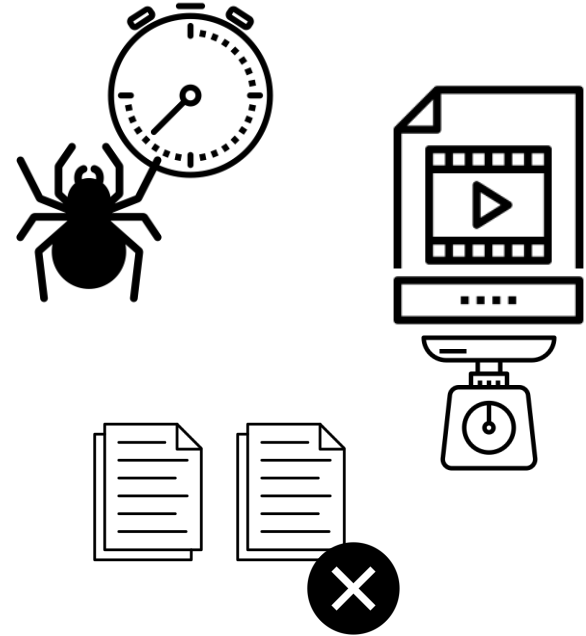
- Web crawls stored to local directories
- Command line, higher technical threshold



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

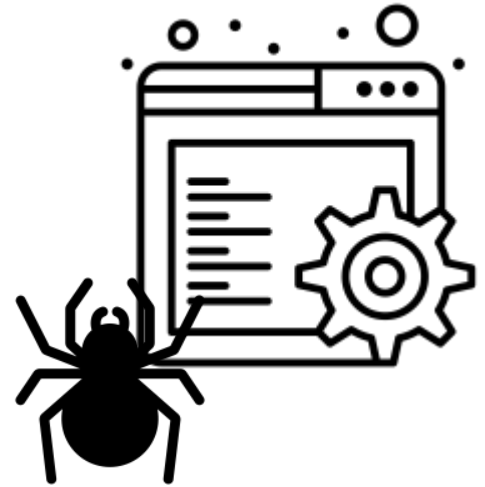
Symmetrical Archiving + Scalability

Browsertrix Crawler

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

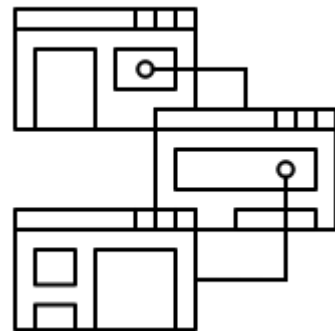
Archive-It: Umbra and Brozzler

- Rich-media and dynamic web behaviors and content



Significant Properties

- Aesthetic properties
- Custom behaviors
- Embedded media as information
- Interactive features of data-driven applications



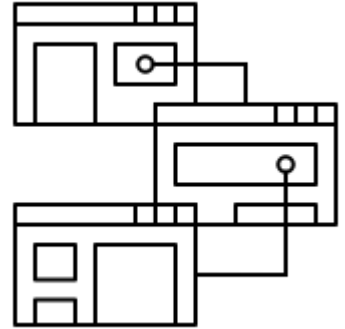
Remote/Emulated Browsers

Conifer

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

OldWeb.today

- Emulated legacy browsers
- No capture capabilities



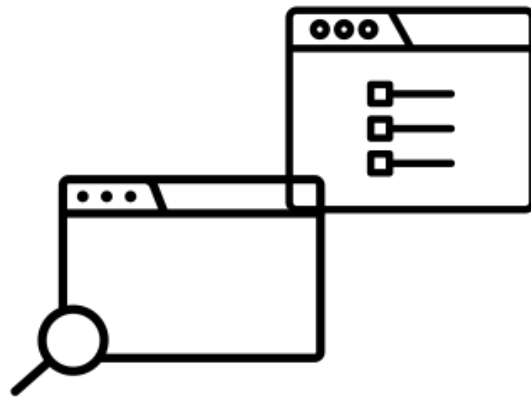
Playback/Rendering/Access

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

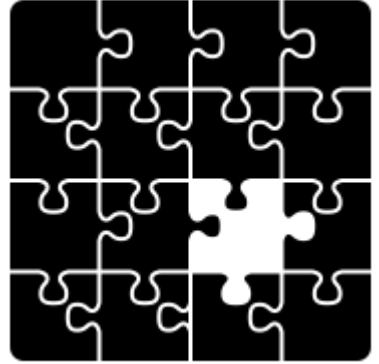
- Patching tools unite missing content

Conifer

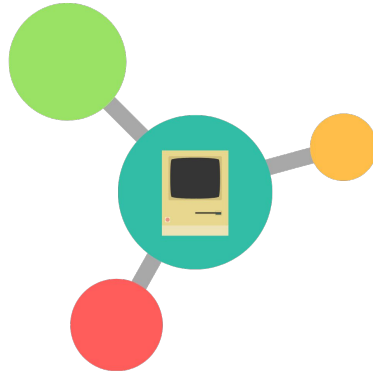
- Importing from open web archives

ArchiveWeb.page

- Page-level archiving
- Export selected pages as single WARC/WACZ




Questions?



Internet Archive & Archive-It: Overview/Demo



Internet Archive (archive.org)


 INTERNET ARCHIVE


WEBBOOKSVIDEODATA SOFTWAREIMAGES

SIGN UP | LOG IN | UPLOAD


ABOUTBLOGPROJECTSHELPDONATECONTACTJOBSVOLUNTEERPEOPLE


Search the history of over 555 billion web pages on the Internet.








Internet Archive is a non-profit library of millions of free books, movies, software, music, websites, and more.


 554B


 30M


 6.7M


 13M

 2.2M

 659K

 3.8M

 226K

 1.0M

GO

Advanced Search

Announcements

Welcome to the Webspace Jam


Filecoin Foundation Grants 50,000 FIL to the Internet Archive

Author and Open Source Advocate VM Brasseur: Internet Archive


'Legitimately Useful' for Lending and Preservation of Her Work

[More announcements](#)


Top Collections at the Archive




American Libraries




The LibriVox Free Audiobook...



Canadian Libraries

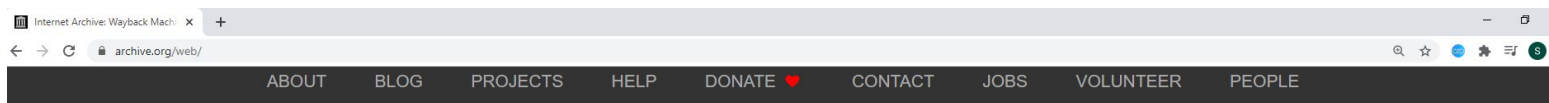


Electric Sheep



Live Music Archive

Wayback Machine (archive.org/web)



Explore more than 555 billion [web pages](#) saved over time

BROWSE HISTORY

Find the Wayback Machine useful?

[DONATE](#)



Tools

[Wayback Machine Availability API](#)
Build your own tools.

[WordPress Broken Link Checker](#)
Banish broken links from your blog



Subscription Service

Archive-It enables you to capture, manage and search collections of digital content without any technical expertise or hosting facilities. [Visit](#)



Save Page Now

SAVE PAGE

Capture a web page as it appears

Search the history of over 555 billion web pages on the Internet.

WayBackMachine

 enter URL or keywords

INTERNET ARCHIVE

WayBackMachine

<https://www.dpoe.network/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.dpoe.network/workshops/>.

Save Page Now (web.archive.org/save)







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.




SIGN UP | LOG IN

BLOGPROJECTSHELPDONATE ♥CONTACTJOBSVOLUNTEERPE

INTERNET ARCHIVE

WayBackMachine

Save Page Now

☒ Save error pages (HTTP Status=4xx, 5xx)

SAVE PAGE

Capture a web page as it appears now for use as a trusted citation in the future.
[Sign in](#) to use extra features: "Save outlinks", "Save screen shot" and "My web archive".

DONATE

INTERNET ARCHIVE



Saving page <https://www.dpoe.network/workshops/>

✓ Done!

The same snapshot had been made 1 minutes 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://web/20210412165535/https://www.dpoe.network/workshops/)

```
https://www.dpoe.network/workshops/  
https://fonts.googleapis.com/css?family=Poppins:400,600&ver=5.7  
https://www.dpoe.network/wp-content/plugins/otter-blocks/vendor/codeinwp/gutenberg-  
animation/assets/css/style.css?ver=5.7  
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/ultimate-social-media-icons/css/sfsi-style.css?ver=5.7  
https://www.dpoe.network/wp-content/plugins/otter-blocks/vendor/codeinwp/gutenberg-blocks/build/style.css?  
ver=1.6.3  
https://www.dpoe.network/wp-content/plugins/otter-blocks/vendor/codeinwp/gutenberg-
```

Downloaded elements: 41

[Return to Save Page Now](#)

← → ↻ web.archive.org/web/"https://www.dpo.e.network/workshops/

INTERNET ARCHIVE Explore more than 555 billion [web pages](#) saved over time


[DONATE](#) **WayBackMachine**

×

Results: 50 100 500

[Calendar](#) · [Collections](#) ^{beta} · [Changes](#) ^{beta} · [Summary](#) · [Site Map](#)

Saved **1 time** [April 12, 2021](#).



2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 **2021**

JAN							FEB							MAR							APR									
					1	2					1	2	3	4	5	6					1	2	3							
3	4	5	6	7	8	9		7	8	9	10	11	12	13			7	8	9	10	11	12	13	4	5	6	7	8	9	10
10	11	12	13	14	15	16		14	15	16	17	18	19	20			14	15	16	17	18	19	20	11	12	13	14	15	16	17
17	18	19	20	21	22	23		21	22	23	24	25	26	27			21	22	23	24	25	26	27	18	19	20	21	22	23	24



Workshops

Sustainable Web Archiving at Scale: An Introduction

Free Workshop

Thursday, April 15th & Friday, April 16th: 2:00-5:00pm EST
Apply now!

Description:

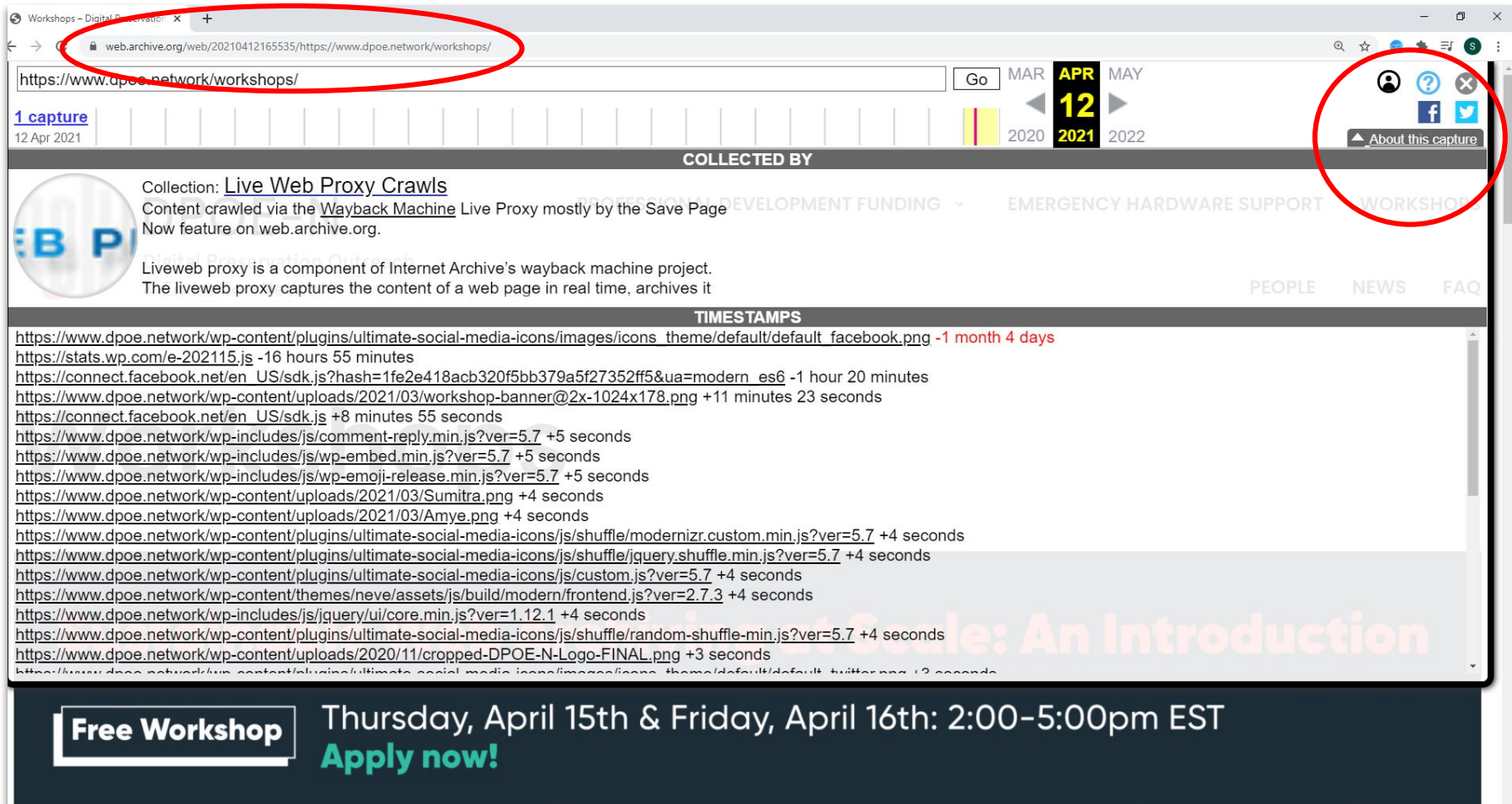
As cultural production and communication have moved online, the need for archivists to document history in real time has become increasingly clear and urgent. Yet the scale and complexity of online information and media can be daunting, even as web archiving tools and practices continue to evolve. And 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.



Wayback Machine

browser extensions, apps and add-ons



Chrome Extension

Firefox Add On

Safari Extension

iOS App

Android App



Archive-It (archive-it.org)

The screenshot shows the Archive-It website interface. At the top, there's a navigation bar with the Archive-It logo, links for HOME, EXPLORE, LEARN MORE, and CONTACT US, and a 'Login' button. Below this is a banner area with a welcome message and a link to a webinar. The main section is titled 'Explore Collections' and features three featured collections: 'Ukraine Conflict', 'Everglades Explorer', and 'North Carolina State Government Web Site Archive'. Each collection has a thumbnail image and a brief description.

archive-it.org

HOME EXPLORE LEARN MORE CONTACT US

The leading web archiving service for collecting and accessing cultural heritage on the web
Built at the Internet Archive

Welcome to Archive-It!
Attend a live informational webinar and demo to learn more about the service

Contact Us to sign up for an upcoming session:
Apr 22 2021, 11:00 AM PDT
May 06 2021, 11:00 AM PDT

Explore Collections Find a Collection by Name Search Show All Collections

Organising Euromaidan. The biggest protest in Ukraine's recent history
From Euromaidan to the Russian Revolution, this collection documents the history of protest in Ukraine. The Russian Revolution was the first time in Russian history that the Russian people organized mass protests and the struggle for democracy. The protests led to the Russian Revolution and the establishment of the Soviet Union. This is the first time in Russian history that the Russian people organized mass protests and the struggle for democracy. This is the first time in Russian history that the Russian people organized mass protests and the struggle for democracy.

Ukraine Conflict
By Internet Archive Global Events

This collection seeks to document conflict in Ukraine as it progresses. Content includes news outlets, social media, blogs, and government websites. Sites are written in English,...

Everglades Explorer
FIU Libraries

Everglades Explorer – EAPRA (Assorted PDF & Report Archive)
By Florida International University Libraries

An archive of digital government and non-government organization (NGO) documents and reports, representing the Greater Everglades watershed and adjacent...

N.C. PROJECT GREEN
GOVERNMENT & POLICY DOCUMENTS & REPORTS & RESEARCH & HISTORY

North Carolina State Government Web Site Archive
By North Carolina State Archives and State Library of North Carolina

The North Carolina State Government Web Site Archives allows free and open access to North Carolina state agency web sites dating back to 1996. Access this collection using the link...



Home

Collections

Crawls

Archives

ARS

Help Center

Welcome, Sumitra Duncan ▾

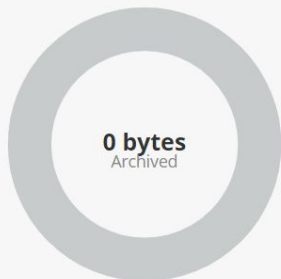
Home

DPOE-N Workshops

0 bytes archived since Apr 6, 2021

InstaCrawl

Current Subscription



Data Budget Usage

▸ Current Subscription Details

▸ Past Subscription Totals

Active Collection List (0 Active Collections)

Type to Filter Active Collections

Create a Collection

📄 Download Collection List

Collection Name

Data (this period) ▾

Docs (this period)

Active Seeds

Last Crawl

No Results

? Help



Home

Collections

Crawls

Archives

ARS

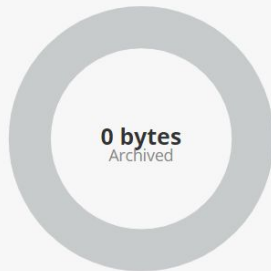
Help Center

Welcome, Sumitra Duncan

Home

DPOE-N Workshops

Current Subscription



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

Active Collection List (0 Active Collections)

[Download Collection List](#)

Create a Collection

Collection Name	Data (this period) ▼	Docs (this period)	Active Seeds	Last Crawl
No Results				

Create A Collection

Please name your collection below. After creating your collection, you'll be taken to the Collection page where you can complete setting up settings and seeds.

Cancel **Create**

DPOE-N Workshop Test Collection

Created: Apr 12, 2021 by Sumitra-dpoe

Updated: Apr 12, 2021 by Sumitra-dpoe

Overview

Seeds

Crawls

Collection Scope

Metadata

Wayback QA

Upload WARCs

Welcome to your new collection

Take a Tour

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

[Guided Tour](#)

Add Seeds

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

[Add Seeds](#)

Delete Collection

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

[Delete Collection](#)

Collection Settings

Public

Active

[Save](#)

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

DPOE-N Workshop Test Collection

Created: Apr 12, 2021 by Sumitra-dpoe

Updated: Apr 12, 2021 by Sumitra-dpoe

Overview

Seeds

Crawls

Collection Scope


Metadata

Wayback QA

Upload WARCs

Seed List (0 Seeds)

Type to Filter Seeds

 Download Seed List

Run Crawl

Edit Settings

Add Metadata


Edit Groups

Delete Seeds


Add Rules


Manage Seed Groups


Add Seeds


☒ 


Seed URL

Group 

Status 

Frequency 

Type 

Access 

Last Crawl

Captures

Wayback

No Results

CollectionsCrawlsArchivesARS

Home / Collections / DPOE-N Workshop Test Collection / Seeds

DPOE-N Workshop Test Collection

Updated: Apr 12, 2021 by Sumitra-dpoe

OverviewSeeds

Seed List (0 Seeds)

Type to Filter Seeds

Run CrawlEdit Settings

☒ Seed URL Group

No Results

Download Seed List

Seed GroupsAdd Seeds

CapturesWayback

Add Seeds

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

✓https://www.dpoe.network/

Access:Public

Frequency:One-Time

Seed Type:Standard

Cancel

Add Seeds

DPOE-N Workshop Test Collection

Created: Apr 12, 2021 by Sumitra-dpoe

Updated: Apr 12, 2021 by Sumitra-dpoe

Overview

Seeds

Crawls

Collection Scope

Metadata

Wayback QA

Upload WARCs

Seed List (1 Seed)

 Download Seed List

Run Crawl

Edit Settings

Add Metadata

Edit Groups

Delete Seeds


Add Rules


Manage Seed Groups


Add Seeds




Seed URL

Group 

Status 

Frequency 

Type 

Access 

Last Crawl

Captures

Wayback



<https://www.dpoe.network/>

Active

One-Time

Standard


Public

[Wayback >](#)

Adding Metadata to Seed URLs

The screenshot shows a web application interface with a dark navigation bar at the top containing links for Home, Collections, Crawls, Archives, and ARS. On the right side of the navigation bar are links for Help Center and Welcome, S. Below the navigation bar, the main content area displays the URL **https://www.dpoenetwork/** with a pencil icon to its right. To the right of the URL are three dropdown menus: Active, Public, and One-Time. Below the URL bar, a status bar shows 'Created: Apr 12, 2021', 'Updated: Apr 12, 2021', and 'In collection DPOE-N Workshop Test Collection'. A tabbed interface below the status bar includes tabs for Settings, Metadata (which is selected and underlined), Crawling History, Notes, and Seed Scope. The main content area under the 'Metadata' tab contains a message: 'No metadata. Click "Edit" to add metadata.' To the right of this message is a button labeled 'Edit', which is circled in red. Above the 'Edit' button is a button labeled 'Import from WorldCat'.

Home Collections Crawls Archives ARS Help Center Welcome, S

https://www.dpoenetwork/  Active Public One-Time

Created: Apr 12, 2021 Updated: Apr 12, 2021 In collection **DPOE-N Workshop Test Collection**

Settings **Metadata** Crawling History Notes Seed Scope

No metadata. Click "Edit" to add metadata.

Import from WorldCat **Edit**

https://www.dpoe.network/ 

Active ▼ Public ▼ One-Time ▼

Created: Apr 12, 2021

Updated: Apr 12, 2021

In collection [DPOE-N Workshop Test Collection](#)

Settings

Metadata

Crawling History

Notes

Seed Scope

Done

Title

Digital Preservation Outreach & Education Network

Save

Grab Title

Creator

Save

Subject

Save

Description

Save

Publisher

Save

Contributor

Save

Initiating Crawls

Home / Collections / DPOE-N Workshop Test Collection / Seeds

DPOE-N Workshop Test Collection

Created: Apr 12, 2021 by Sumitra-dpoe **Updated:** Apr 12, 2021 by Sumitra-dpoe






Overview **Seeds** Crawls Collection Scope Metadata Wayback QA Upload WARCs

Seed List (1 Seed)

Type to Filter Seeds

[Download Seed List](#)

[Run Crawl](#) [Edit Settings](#) [Add Metadata](#) [Edit Groups](#) [Delete Seeds](#) [Add Rules](#) [Manage Seed Groups](#) [Add Seeds](#)

<input checked="" type="checkbox"/> Seed URL	Group 	Status 	Frequency 	Type 	Access 	Last Crawl	Captures	Wayback
<input checked="" type="checkbox"/> https://www.dpoe.network/		Active	One-Time	Standard	Public			Wayback >

DPOE-N Workshop Test Collection

Created: Apr 12, 2021 by
Sumitra-dpoe

Updated: Apr 12, 2021 by
Sumitra-dpoe

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

Documents

Data Limit

1

GB

Time Limit

1 Hour

Crawl PDFs Only

☐

Crawling Technology

☒ Standard

☐ Brozzler

1 selected seed will be crawled.

Cancel

Crawl

DPOE-N Workshop Test Collection

Created: Apr 12, 2021 by Sumitra-dpoe

Updated: Apr 12, 2021 by Sumitra-dpoe

Select two crawls to compare.

Overview

Seeds

Crawls

Collection Scope

Metadata

Wayback QA

Upload WARCs

Crawl Reports

Current Crawls

Test Crawls

Scheduled Crawls

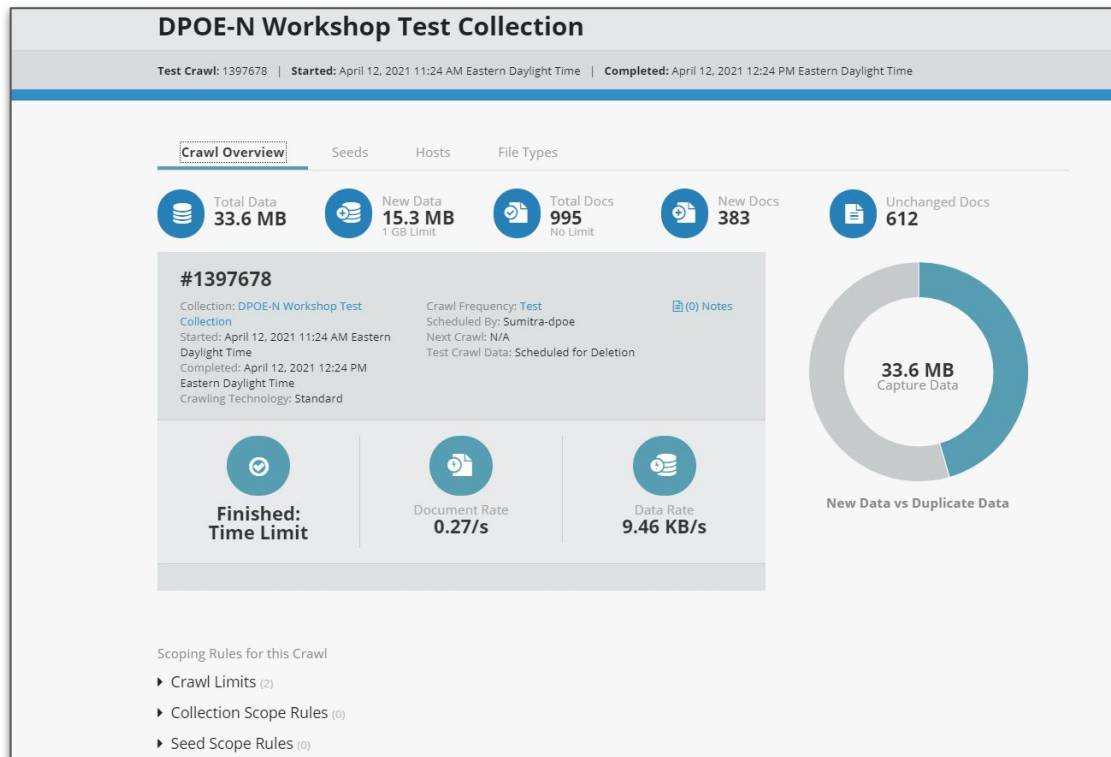
Test Crawl List (1 Test Crawl)

Type to Filter Test Crawls

 [Download Test Crawl List](#)

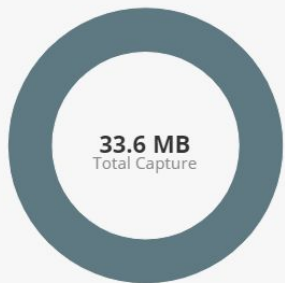
Crawl ID	Started ▼	Completed	Status	New Data	Docs
<input type="checkbox"/> 1397678	Apr 12, 2021		In progress (new)	0 bytes	0

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



Data by Seed



Top Seed by Document Count

Scoping Rules for this Crawl

- ▶ Crawl Limits (2)
- ▶ Collection Scope Rules (0)
- ▶ Seed Scope Rules (0)

Seed List (1 Seed)

[Download Seed List](#)

Seed URL	Seed Type	Seed Status	Docs	New Docs	Data ▾	New Data	Seed	Wayback Link
https://www.dpoenetwork/	Standard	Crawled	995	383	33.6 MB	15.3 MB	Seed >	Wayback >

You are viewing a temporarily archived web page, collected at the request of [DPOE-N Workshops](#) using [Archive-It](#). This page was captured in a test crawl on 15:24:33 Apr 12, 2021, run in the [DPOE-N Workshop Test Collection](#) collection. If you would like this crawl to become part of your collection permanently, you will need to save it from the [crawl report](#). The information on this web page may be out of date. See [All versions](#) of this archived page. Found 0 archived media items out of 0 total on this page.



PROFESSIONAL DEVELOPMENT FUNDING ▾

EMERGENCY HARDWARE SUPPORT

WORKSHOPS

PEOPLE

NEWS

FAQ

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

Explore >> DPOE-N Workshops



DPOE-N Workshops

Archive-It Partner Since: Apr, 2021

Organization Type: [Other Institutions](#)

Organization URL:

Narrow Your Results

There are no further ways to narrow your results.

Sites and collections from this organization are listed below. Narrow your results at left, or enter a search query below to find a collection, site, specific URL or to search the text of archived webpages.

Collections

Sites

Search Page Text

Page 1 of 1 (1 Total Results)

Sort By: Collection Name (A-Z) | [Collection Name \(Z-A\)](#)

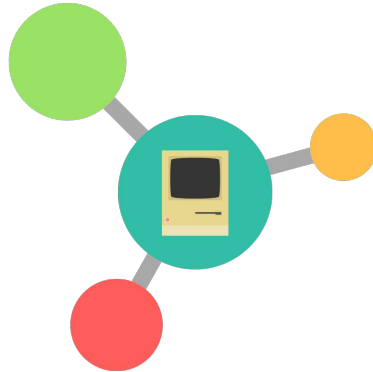
DPOE-N Workshop Test Collection

Archived since: Apr, 2021

No description.

Page 1 of 1 (1 Total Results)

Questions?



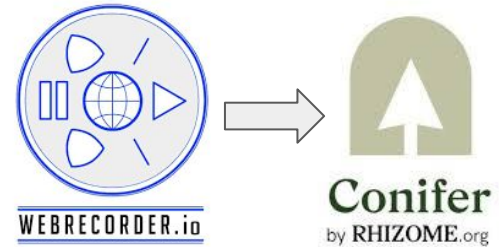
Conifer

- 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 removed
- Sessions are basis for Collections



Conifer: Sessions

Sessions

Expand All

 	Session from 4/15/2021, 10:53:15 PM	1 Pages	2mins	2.59 MB
 	Patch of	0 Pages	1min	67.32 KB
	Session from 4/12/2021, 5:30:25 PM	3 Pages	1min	415.10 MB

12
Apr
MON
2021

Delete

Download

Session Notes

Add notes about this session. Visible only to you.

edit

Session Pages (3)

National Forum on Ethics and Ar... <https://eaw.rhizome.org/>

Ethics and Archiving the Web: W... <https://vimeo.com/277336026>


Ethics and Archiving the Web: W... <https://vimeo.com/277336026>

Conifer: Collections

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



Conifer: Collection Manager



DPOE-N Workshop

amccarther

COLLECTION by amccarther

DPOE-N Workshop

This collection demonstrates the types of

+ New Session

...

Collection Cover

Private

Collection Resources

LISTS (3 Public) EDIT +

Soap Library Instagram (1)

Rhizome EAW Vimeo (1)

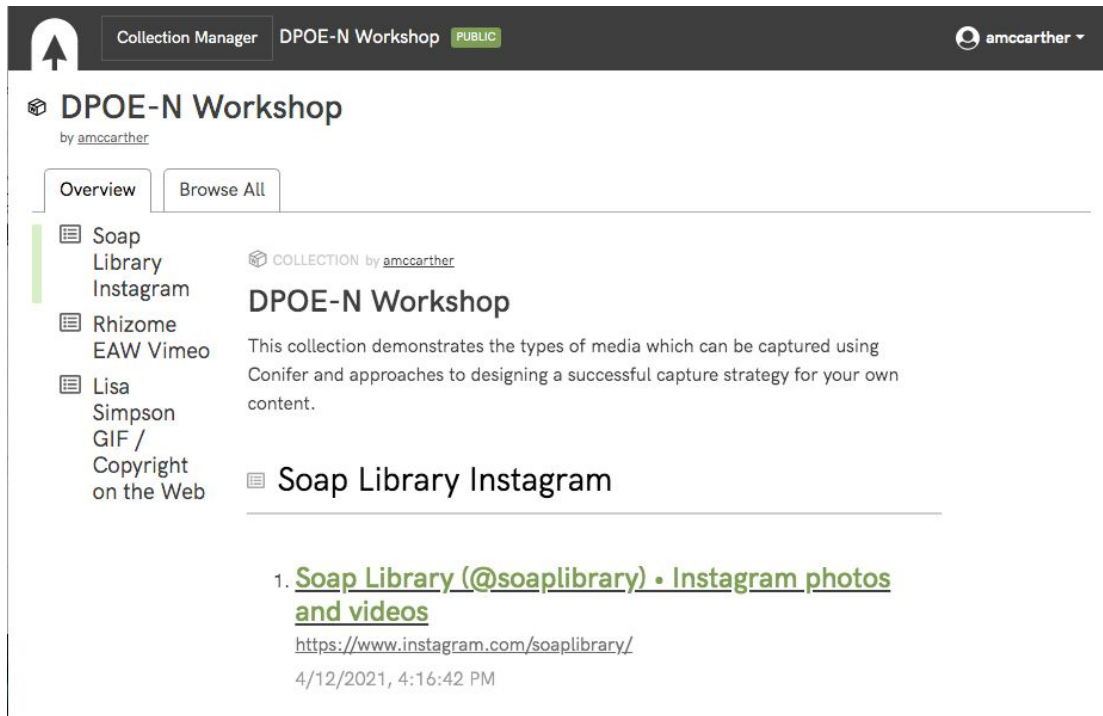
Lisa Simpson GIF / Copyri... (1)

is:page

TIMESTAMP	PAGE TITLE	URL	CAPTURE B...	SESSI...
4/15/2021, 10:56:...		https://twitter.com/ha:		fckghtsyzli
4/15/2021, 10:53:...		https://twitter.com/dp:		plxoiilp3s6
4/12/2021, 5:30:4...	Ethics and Archivir	https://vimeo.com/27:		afk767lwc
4/12/2021, 5:30:3...	Ethics and Archivir	https://vimeo.com/27:		afk767lwc
4/12/2021, 5:30:2...	National Forum on	https://eaw.rhizome.or		afk767lwc
4/12/2021, 4:57:3...	Lisa Simpson Episc	https://giphy.com/gifs/		qadurbsne
4/12/2021, 4:57:2...	Lisa Simpson Episc	https://giphy.com/gifs/		qadurbsne
4/12/2021, 4:47:43...	Ethics and Archivir	https://vimeo.com/27:		3djpdue2i
4/12/2021, 4:47:39...	Ethics and Archivir	https://vimeo.com/27:		3djpdue2i



Conifer: Collection Cover



The screenshot displays the 'Collection Manager' interface for a user named 'amccarther'. The main header shows 'Collection Manager', 'DPOE-N Workshop', and a 'PUBLIC' status tag. The user's profile 'amccarther' is in the top right. The collection title 'DPOE-N Workshop' is prominently displayed, followed by 'by amccarther'. Below the title are two tabs: 'Overview' (selected) and 'Browse All'. A left sidebar lists collection categories: 'Soap Library Instagram', 'Rhizome EAW Vimeo', and 'Lisa Simpson GIF / Copyright on the Web'. The main content area for the 'DPOE-N Workshop' collection includes a description: '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.' Below this is a section titled 'Soap Library Instagram' which lists a single item: '1. [Soap Library \(@soaplibrary\) • Instagram photos and videos](https://www.instagram.com/soaplibrary/)'. The item's URL and a timestamp '4/12/2021, 4:16:42 PM' are shown below the link.

Collection Manager DPOE-N Workshop PUBLIC amccarther

DPOE-N Workshop

by amccarther

Overview Browse All

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

COLLECTION by amccarther

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/)
<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

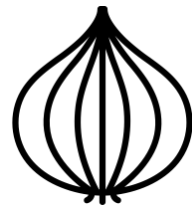


Conifer: Pre-configured browsers

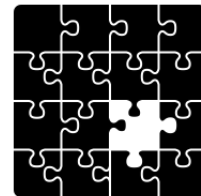
The screenshot displays the Conifer web interface. At the top, a dark navigation bar contains a logo on the left, the text "Collection Manager" and "DPOE-N Workshop" with a "PUBLIC" status tag in the center, and a user profile "amccarther" on the right. Below this is a light-colored header bar representing a browser window. It includes a "Chrome v76" dropdown menu with a blue arrow pointing to it, a URL bar showing a path to "http://www.newmuseum.org/", a timestamp "Mon, 10 Oct 2011 00:38:46 GMT" and "Internet Archive" dropdown, and an "Extract" button. The main content area features a central box with the text "Create a new capture" and "Ready to add a new capture to your collection DPOE-N Workshop". On the right side, there is a logo consisting of a stylized tree with an upward arrow, and the text "Conifer by RHIZOME.org".

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

The screenshot displays the Conifer web interface. At the top, a dark header bar contains a logo on the left, the text "Collection Manager" and "DPOE-N Workshop" in the center, and a "PUBLIC" status indicator and a user profile "amccarther" on the right. Below the header, a simulated browser interface is shown. It includes a Chrome v76 logo, a URL bar with the address "http://www.newmuseum.org/", a timestamp "Mon, 10 Oct 2011 00:38:46 GMT", and a source "Internet Archive". A green "Extract" button is positioned to the right of the browser simulation, with a blue arrow pointing towards it. In the center of the page, a light gray box contains the text "Create a new capture" and "Ready to add a new capture to your collection DPOE-N Workshop". On the right side, there is a logo consisting of a stylized tree with an upward-pointing arrow, with the text "Conifer by RHIZOME.org" below it.

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



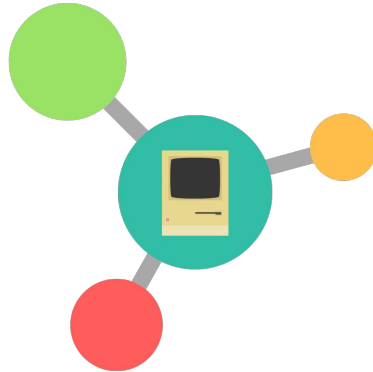
Webrecorder

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

Perma.cc ∞

Social
Feed
Manager

SOCIAL
HUMANS



YouTube - DL

twarc.py

DocNow

4172545 501064209648848896 50106
2 501064211997274114 50106421230
64220415229953 50106422084765696
96499201 501064225142624256 5010
92 50106422888364546 5010642286
064231941570561 5010642321886659
107897344 501064235175399425 501
808 501064237863559168 501064238
1064242347638787 501064241106132
6671593473 501064246730301445 50
4944 501064250484596736 50106425
01864253478936576 50106425365510
56775663616 501064257115482240 5

DIFF
ENGINE

Social Feed Manager

<https://gwu-libraries.github.io/sfm-ui/>



Perma.cc

Perma.cc ∞

About Perma.cc

Guide

Blog

Sign up

Log in

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

Individual

Anyone can create an account and start creating Perma Links.

Libraries >

Libraries play a critical role in powering and supporting Perma.cc.

Journals >

Over 150 academic law journals prevent link rot with Perma.cc.

Faculty >

Faculty use Perma.cc to prevent link rot in their scholarship.

Courts >

Courts care about the accuracy, integrity and reliability of the citations in their opinions.

Law Firms >

Law firms use Perma.cc to prevent link rot in their court filings and marketing materials.

Create an individual account

First name

Last name

Email address

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

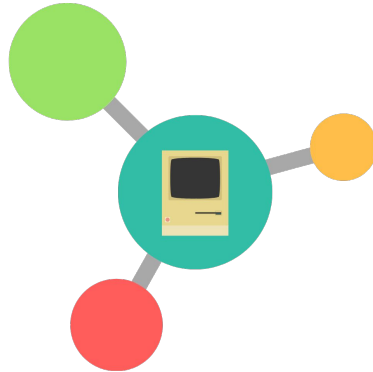
Archive Twitter JSON using this command line tool.



Diff Engine

Track changes in news articles through RSS feeds.

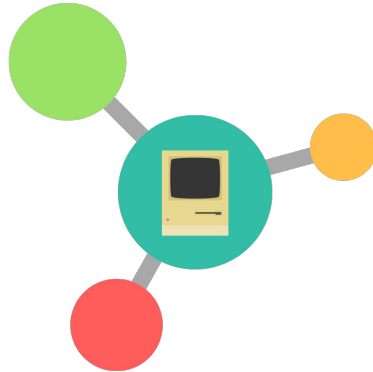
Questions?



Break: 10 minutes

Hands-on exercises: Breakout rooms

Discussion/Questions



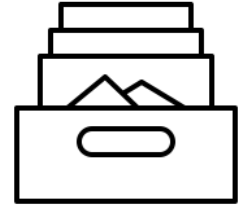
Roadmap



SCOPE



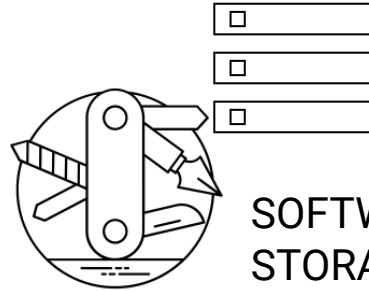
BUDGET & STAFF



COLLECTING



COLLECTING
POLICY



SOFTWARE &
STORAGE

Thank you!

