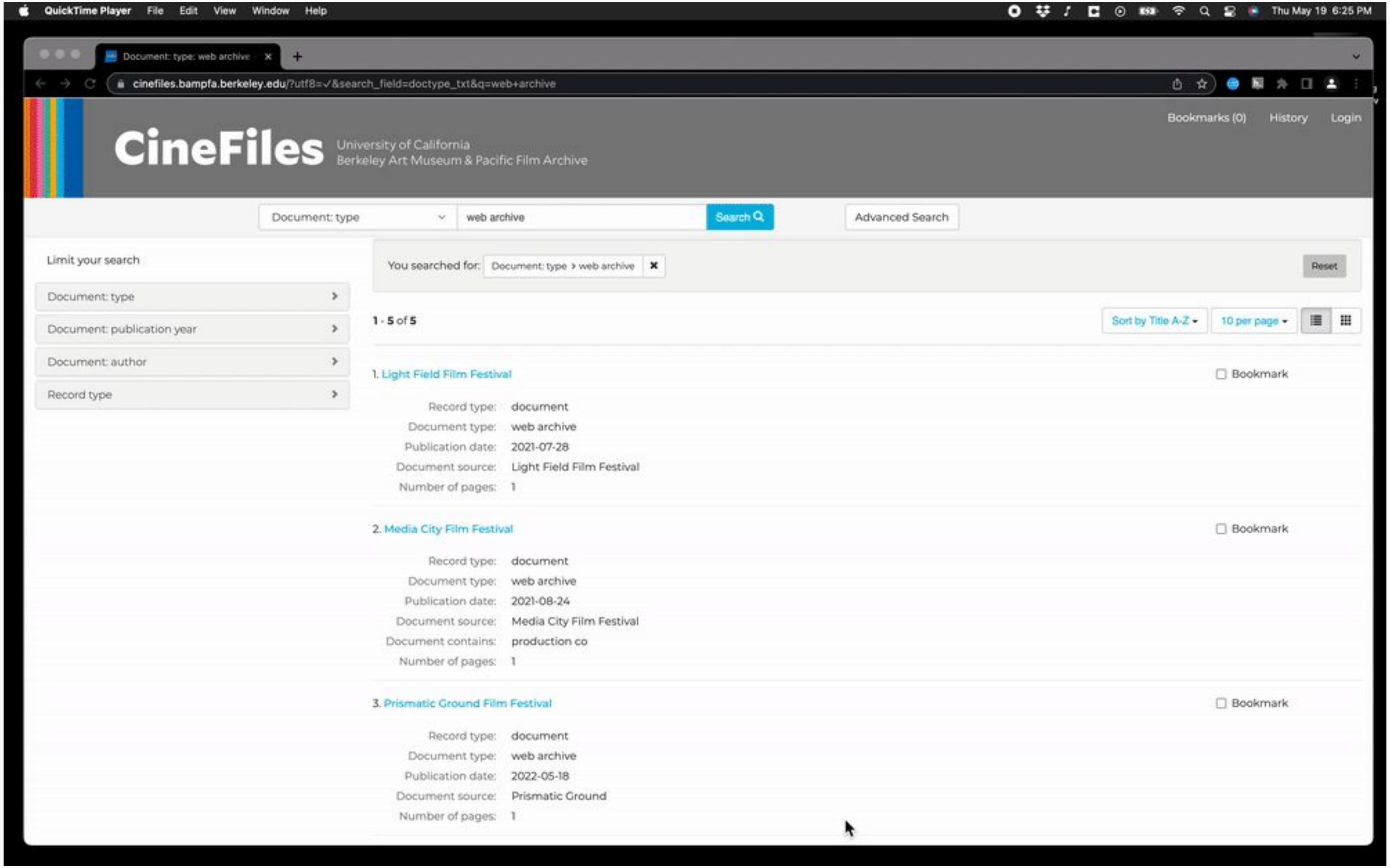
Web archives in digital repositories

Simple integration and reducing software maintenance footprint

Ed Summers
Stanford

Ilya Kreymer
Webrecorder

https://bit.ly/c4l-webarch


QuickTime Player File Edit View Window Help
Thu May 19 6:25 PM
Document: type: web archive
cinefiles.bampfa.berkeley.edu/?utf8=✓&search_field=doctype_txt&q=web+archive
Bookmarks (0)
History
Login
Document: type
Search
Advanced Search
Reset
Sort by Title A-Z
10 per page
Document: type
Document: publication year
Document: author
Record type

CineFiles University of California Berkeley Art Museum & Pacific Film Archive

web archive

Limit your search

You searched for: Document: type › web archive ✖

1 - 5 of 5

1. Light Field Film Festival □ Bookmark

Record type:	document
Document type:	web archive
Publication date:	2021-07-28
Document source:	Light Field Film Festival
Number of pages:	1

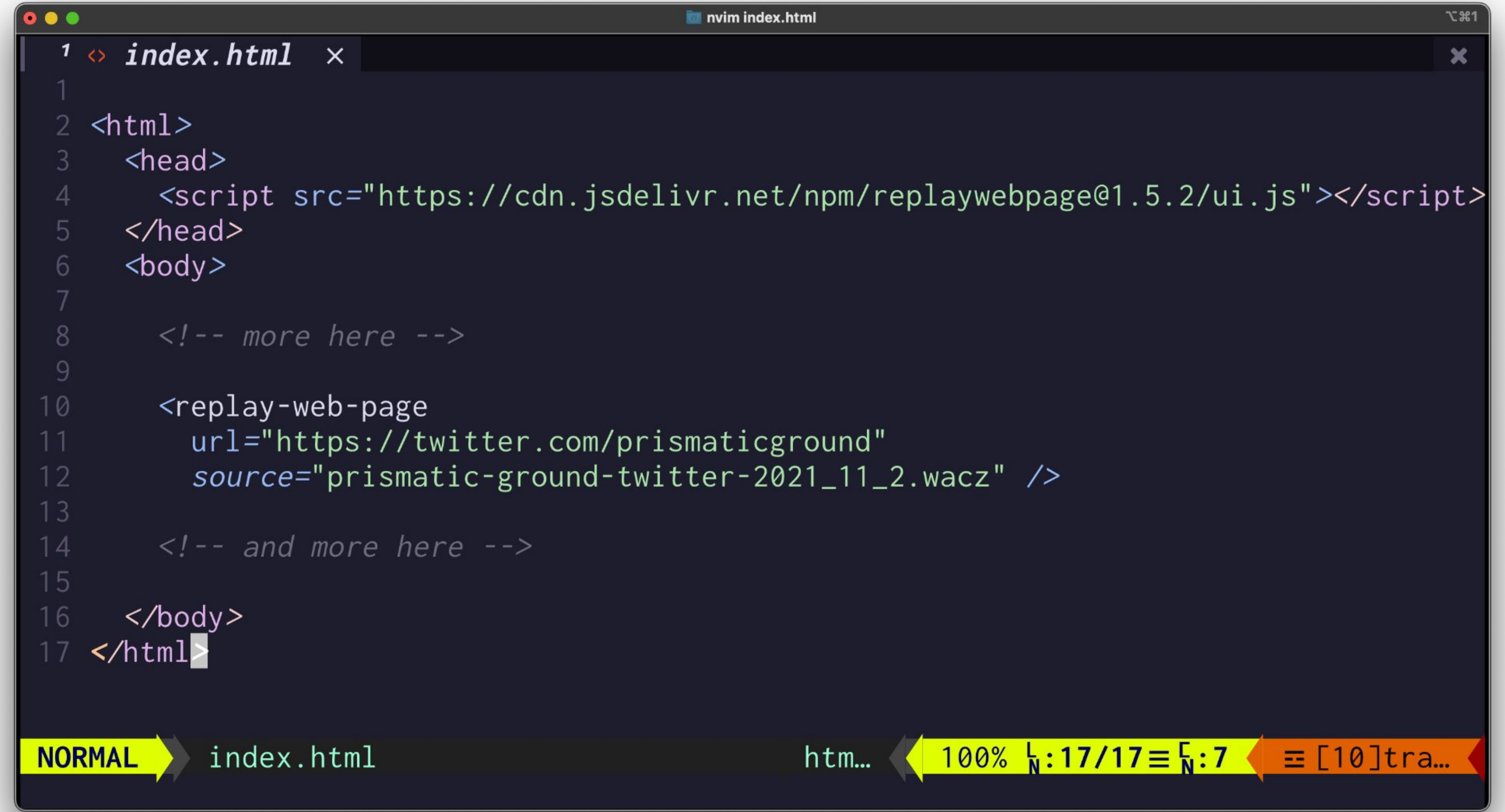
2. Media City Film Festival □ Bookmark

Record type:	document
Document type:	web archive
Publication date:	2021-08-24
Document source:	Media City Film Festival
Document contains:	production co
Number of pages:	1

3. Prismatic Ground Film Festival □ Bookmark

Record type:	document
Document type:	web archive
Publication date:	2022-05-18
Document source:	Prismatic Ground
Number of pages:	1

https://bit.ly/c4l-webarch1



```html

<html>
  <head>
    <script src="https://cdn.jsdelivr.net/npm/replaywebpage@1.5.2/ui.js"></script>
  </head>
  <body>

    <!-- more here -->

    <replay-web-page
      url="https://twitter.com/prismaticground"
      source="prismatic-ground-twitter-2021_11_2.wacz" />

    <!-- and more here -->

  </body>
</html>
```

What is a ***Digital Repository***?

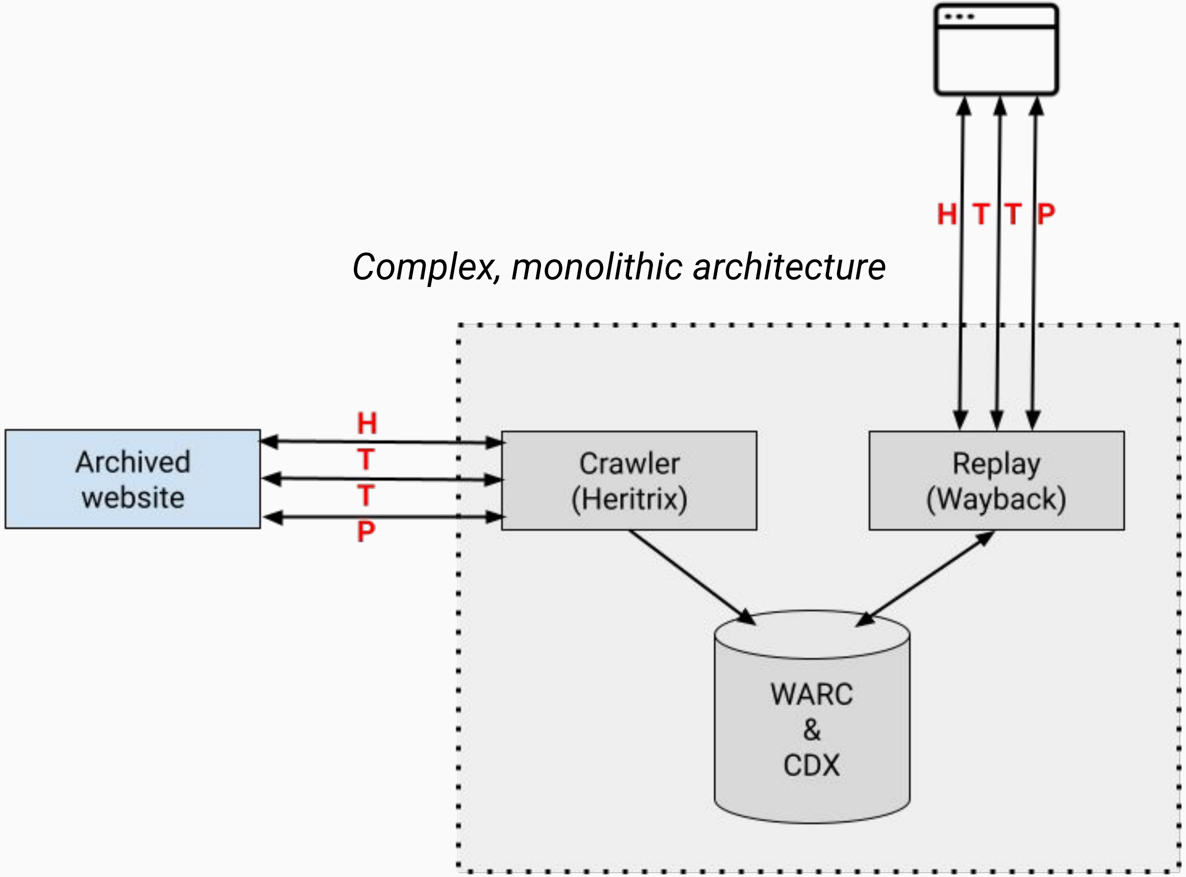
"An institutional repository is an archive for collecting, preserving, and disseminating digital copies of the intellectual output of an institution, particularly a ***research institution***."

–Wikipedia

What is a ***Web Archive***?

“Web archiving is the process of collecting portions of the World Wide Web to ensure the information is preserved in an archive for future researchers, historians, and the public. Web archivists typically employ web crawlers for automated capture due to the massive size and amount of information on the Web.”

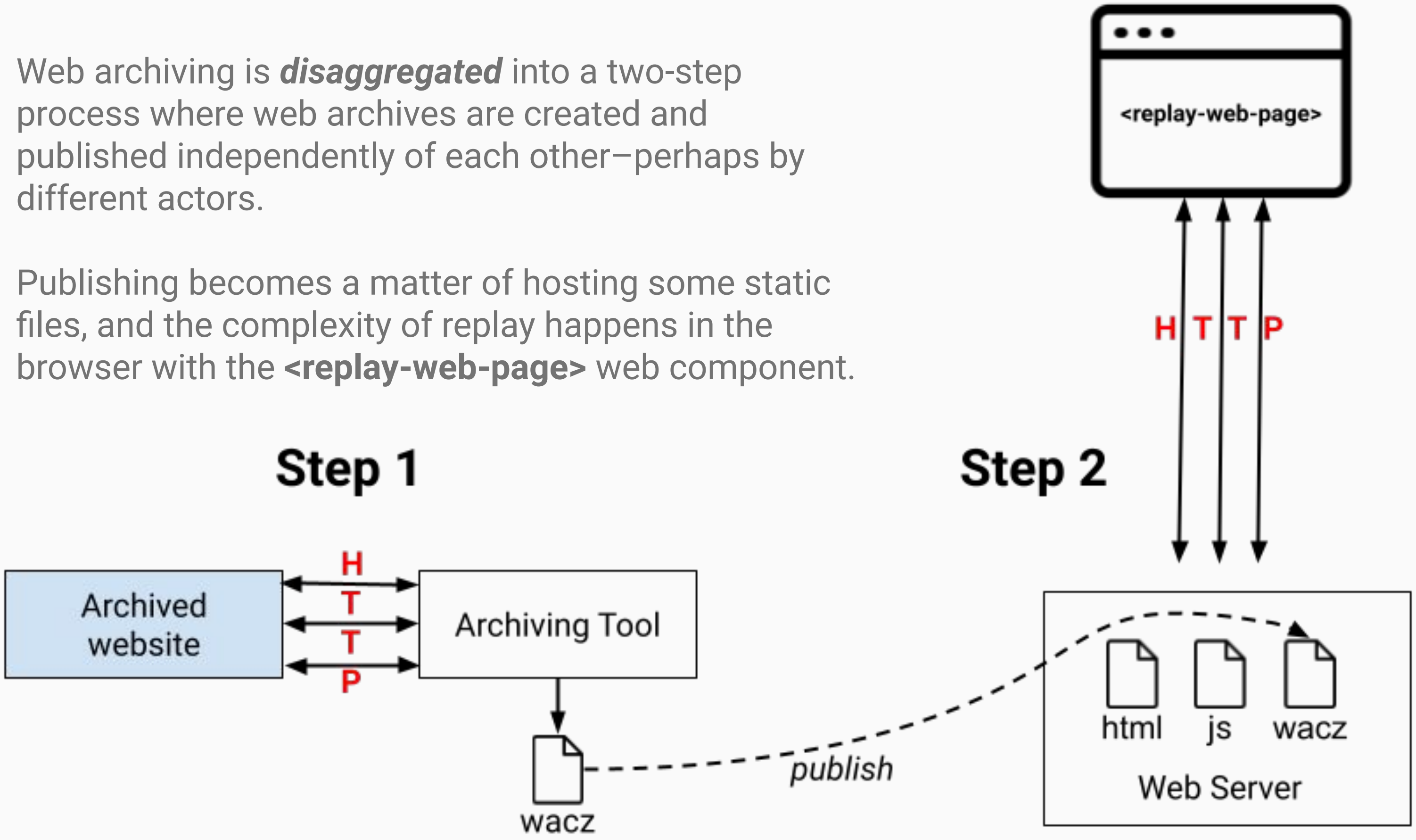
–Wikipedia



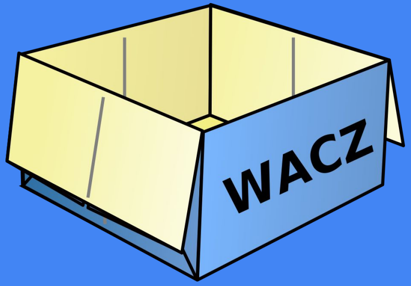
Making Web Archives Portable with *WACZ*

Web archiving is ***disaggregated*** into a two-step process where web archives are created and published independently of each other–perhaps by different actors.

Publishing becomes a matter of hosting some static files, and the complexity of replay happens in the browser with the **<replay-web-page>** web component.



What is the **WACZ** format?



```
WACZ
├── datapackage.json (Manifest)
├── datapackage-digest.json (Signatures)
├── archive
│   └── data.warc.gz (Raw Web Archives)
├── indexes
│   └── index.cdx.gz (Indexes)
└── pages
    └── pages.jsonl (Page Metadata)
```

Web **A**rchive **C**ollection (or **C**rawl) **Z**ipped

- Bundles web archive data (WARC) + metadata
- Indexed (CDX), random-access readable
- Extends *Frictionless Data Package*
- Can include cryptographic signatures
- Just a Zip File!
- https://specs.webrecorder.net/wacz/latest/

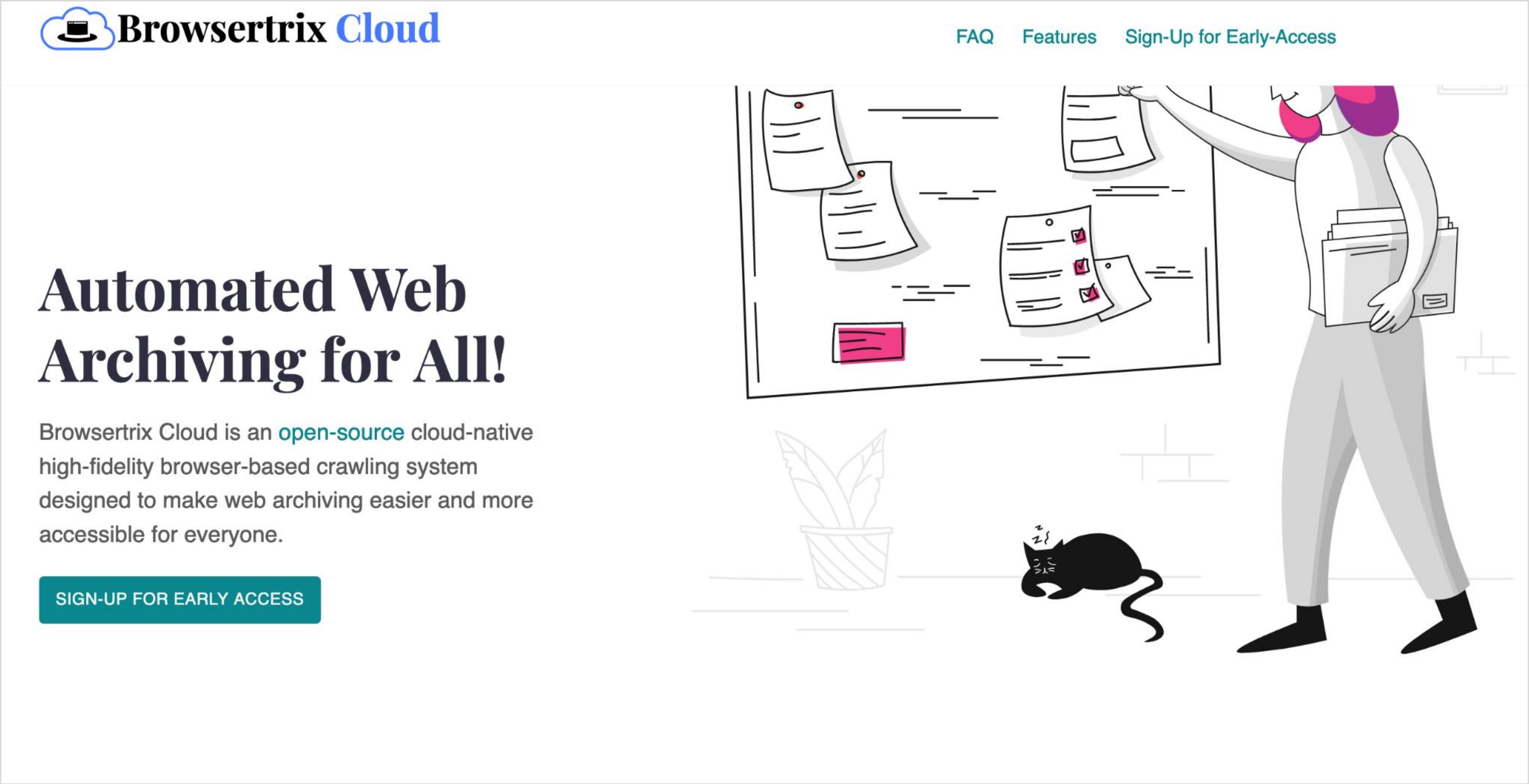
Publishing Web Archives

1. Allow repository users to add ***.wacz*** files like they can other types of files (e.g. pdf, mp4, mp3, jpg, docx, etc)
2. Add a ***<script>*** element for the ReplayWebPage JavaScript library to your web archive item views.
3. Use the ***<replay-web-page />*** web component in item views for web archives that points at the WACZ file.

See https://replayweb.page/docs/embedding for more details

Creating WACZ Files

1. py-wacz - a CLI tool for bundling WARC files into a WACZ
2. archiveweb.page - a Chrome extension and Electron app for interactively creating web archives
3. express.archiveweb.page - create a snapshot of a single public web page in any browser
4. browsertrix-crawler - a command line browser-based crawler (Docker)

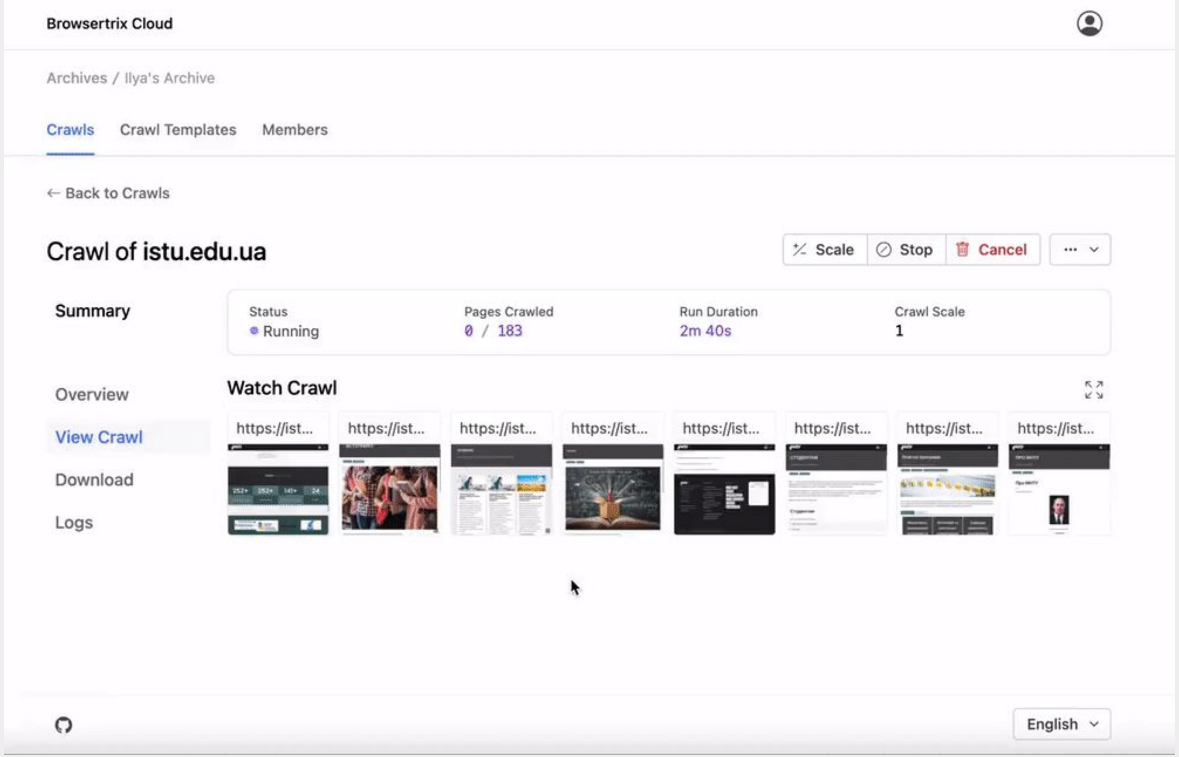


Automated Web Archiving for All!

Browsertrix Cloud is an open-source cloud-native high-fidelity browser-based crawling system designed to make web archiving easier and more accessible for everyone.

https://browsertrix.cloud

Browsertrix Cloud running a crawl with multiple browsers



Crawl of istu.edu.ua

Summary

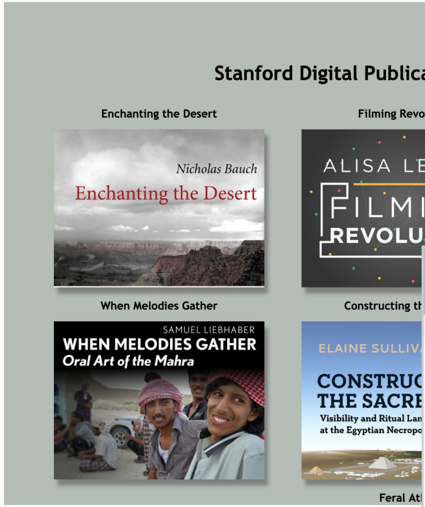
Status: Running

Pages Crawled: 0 / 183

Run Duration: 2m 40s

Crawl Scale: 1

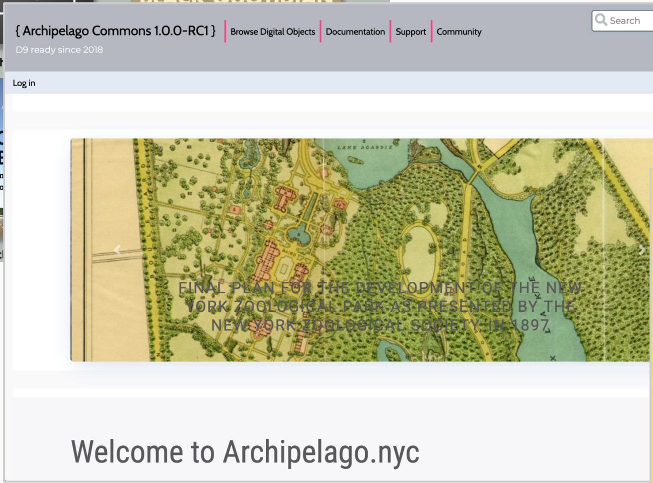
Some more examples … we would ❤️ to see more!



Stanford Digital Publication Web Archives

- Enchanting the Desert
- Filming Revolution
- Black Quotidian
- When Melodies Gather
- Constructing th
- Feral At

https://bit.ly/c4l-webarch2



{ Archipelago Commons 1.0.0-RC1 } Browse Digital Objects | Documentation | Support | Community

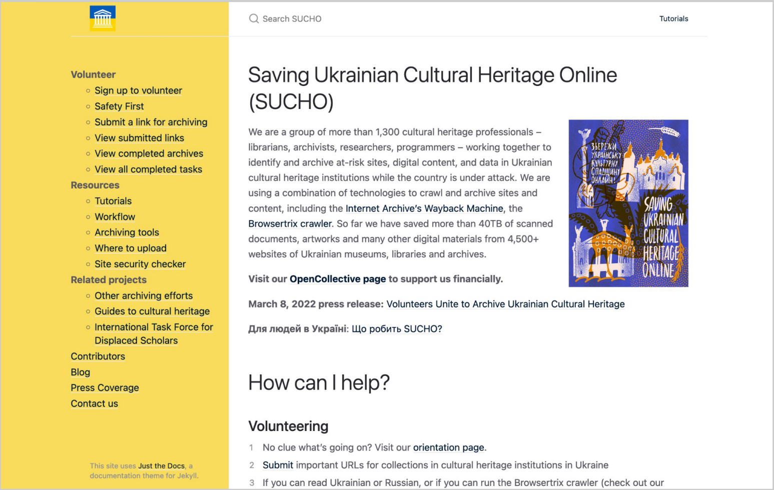
D9 ready since 2018

Search
Search
Log in
FINAL PLAN FOR THE DEVELOPMENT OF THE NEW YORK ZOOLOGICAL PARK AS PRESENTED BY THE NEW YORK ZOOLOGICAL SOCIETY IN 1897

Welcome to Archipelago.nyc

https://bit.ly/c4l-webarch3

https://sucho.org



Volunteer
- Sign up to volunteer
- Safety First
- Submit a link for archiving
- View submitted links
- View completed archives
- View all completed tasks

Resources
- Tutorials
- Workflow
- Archiving tools
- Where to upload
- Site security checker

Related projects
- Other archiving efforts
- Guides to cultural heritage
- International Task Force for Displaced Scholars

Contributors
Blog
Press Coverage
Contact us

This site uses Just the Docs, a documentation theme for Jekyll.

Saving Ukrainian Cultural Heritage Online (SUCHO)

We are a group of more than 1,300 cultural heritage professionals – librarians, archivists, researchers, programmers – working together to identify and archive at-risk sites, digital content, and data in Ukrainian cultural heritage institutions while the country is under attack. We are using a combination of technologies to crawl and archive sites and content, including the Internet Archive's Wayback Machine, the Browsertrix crawler. So far we have saved more than 40TB of scanned documents, artworks and many other digital materials from 4,500+ websites of Ukrainian museums, libraries and archives.

Visit our OpenCollective page to support us financially.

March 8, 2022 press release: Volunteers Unite to Archive Ukrainian Cultural Heritage

Для людей в Україні: Що робить SUCHO?

How can I help?

Volunteering

1. No clue what's going on? Visit our orientation page.
2. Submit important URLs for collections in cultural heritage institutions in Ukraine
3. If you can read Ukrainian or Russian, or if you can run the Browsertrix crawler (check out our

- https://webrecorder.net
- https://forum.webrecorder.net
- https://github.com/webrecorder
- https://specs.webrecorder.net