Web archives in digital repositories

Simple integration and reducing software maintenance footprint

Ed Summers
Stanford

Ilya Kreymer
Webrecorder

<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
Web archiving is \textit{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 \texttt{<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 Archive Collection (or Crawl) Zipped**

- 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/](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](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.

SIGN-UP FOR EARLY ACCESS

https://browsertrix.cloud
Browsertrix Cloud running a crawl with multiple browsers
Some more examples ... we would love to see more!


https://sucho.org
● https://webrecorder.net
● https://forum.webrecorder.net
● https://github.com/webrecorder
● https://specs.webrecorder.net