Webrecorder

Web archiving for all!

FOSDEM Lightning Talk

Ilya Kreymer
Webrecorder Lead Developer

ABOUT WEBRECORDER

Webrecorder project builds tools to specializing in a ‘user-driven’ form of web archiving, where the user is able to direct the archiving process through their browser.

Started in 2014, an independent project since 2020

https://webrecorder.net/

Webrecorder

KEY TOOLS:

Replayweb.page
Fully Browser-based web archive replay system (uses wabac.js replay system)

Archiveweb.page
Brand new High-Fidelity Chrome Browser-Based Web Archiving

pywb
Core Python web archiving system, used by many GLAM institutions

Browsertrix-Crawler
Automated High-Fidelity Browser-Based Crawling System

About Webrecorder

The Webrecorder project is focused on advancing open source software development and research in the following key areas:

- FOSS web archiving tools to create and view web archives
- Highest-fidelity capture and replay
- Integrate with existing archival systems
- Exploring intersection of web archiving and software emulation
- Empower anyone to create, use and share web archives
- Making web archiving more accessible via decentralized and p2p technologies

Web archiving for all!

https://github.com/webrecorder

https://github.com/oldweb-today

https://webrecorder.net/

https://archiveweb.page/

https://replayweb.page/

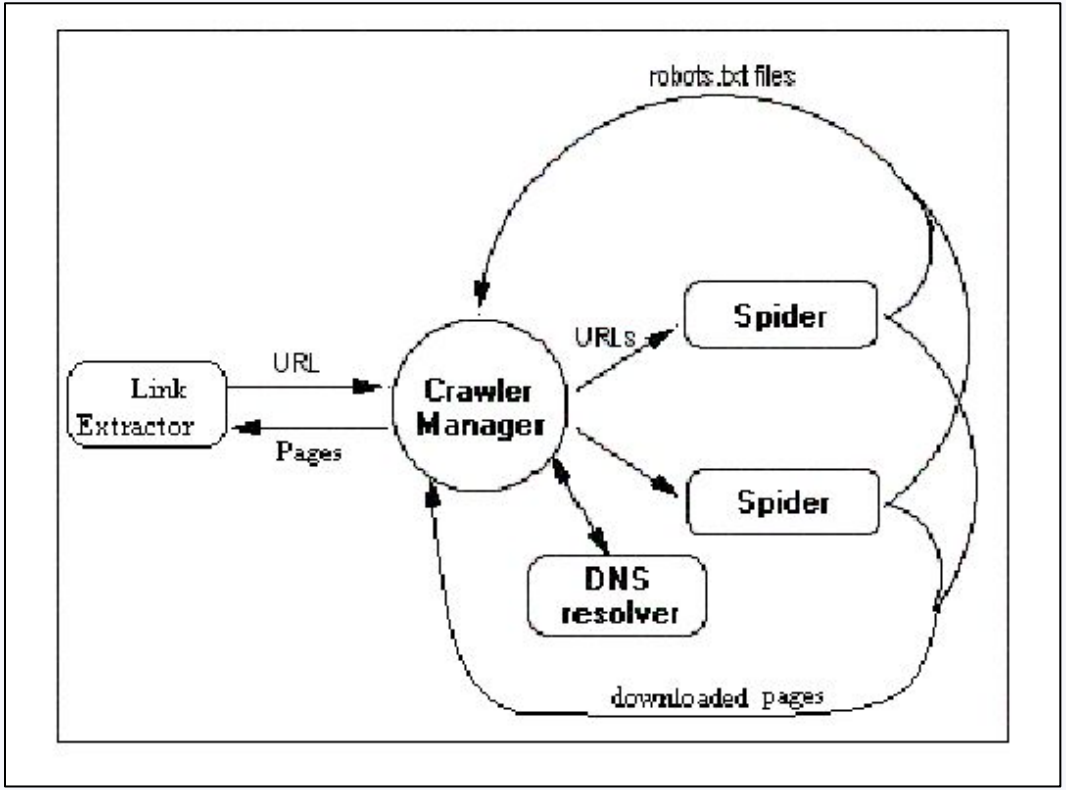


Image from: DESIGNING AND IMPLEMENTATION OF" REGIONAL CRAWLER" AS A NEW STRATEGY FOR CRAWLING THE WEB

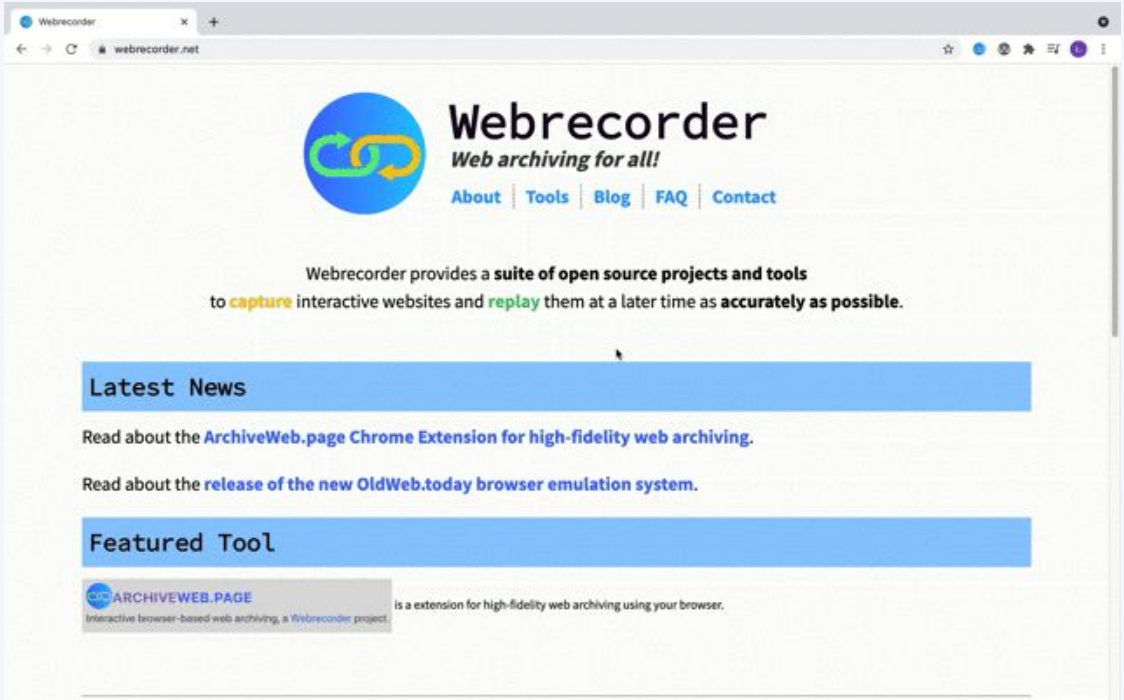
What is web archiving?

"Web archiving is the process of collecting portions of the World Wide Web, preserving the collections in an archival format, and then serving the archives for access and use."
International Internet Preservation Consortium

Traditional Web Crawling

A web crawler, spider, or search engine bot downloads and indexes content from all over the Internet.

Common traditional web crawling tools:
Heritrix, wget, HTTrack



Browser-Based Web Archiving

- The browser is the tool that creates the web archive.
- Can be user-driven, users choose what to archive and what not to.
- Mostly manual, but can be partially automated.
- Focus on quality vs quantity

Browser Based Web Archiving Tools

CAPTURE

ArchiveWeb.page
https://archiveweb.page/

REPLAY

ReplayWeb.page
https://replayweb.page/

ArchiveWeb.page

GitHub:
https://github.com/webrecorder/archiveweb.page

https://archiveweb.page/

Features

- Available as Chromium Extension and Electron Desktop App

- Archive any page as you're browsing
- Store all data locally
- Export and Import to portable format (WACZ)
- Automation via Autopilot System

ReplayWeb.page

GitHub:
https://github.com/webrecorder/replayweb.page

https://replayweb.page/

Embedding Guide:
https://replayweb.page/docs/embedding

Features

- Client-side SPA for loading web archives directly in the browser
- Loads WARC, WACZ and HAR formats
- Can be embedded into other web sites as WebComponent
- Embedding Example

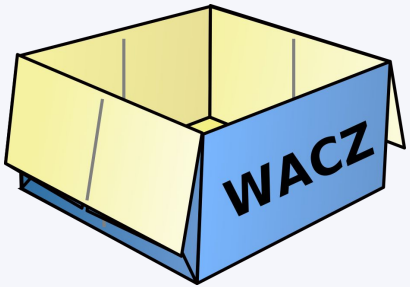
WACZ Format

Web **AR**chive **C**ollection **Z**ipped (WACZ)

A directory structure + ZIP format specification for sharing and distributing web archives.

Bundling *ISO Standard WARC files* + random access indexes + metadata.

Draft Specification:
https://webrecorder.github.io/wacz-spec/latest



Current* (and Potential) Features:

- List of entry pages to start browsing from*
- Full-text search index*
- Technical metadata*
- Descriptive metadata*

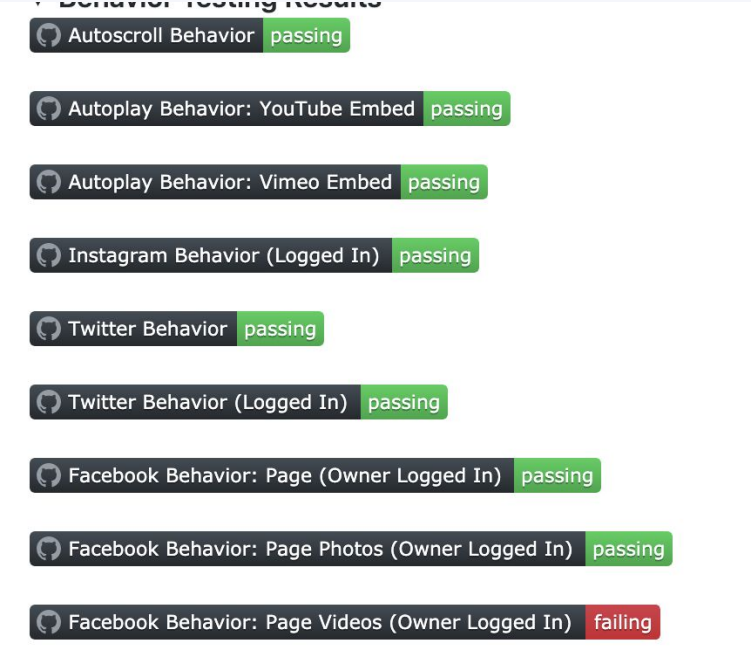
In development

- Screenshots of key pages
- Encryption
- Proof of Authenticity
- Fast access to multiple WACZ
- Crawl or capture logs

Webrecorder
Web archiving for all!

Automated Behaviors

- Behaviors will break and will need to be updated !?! :(
- Daily CI test runs each behavior on a fixed data set (social media profile, preset page)
- We can see the results of automated behaviors and respond to more quickly, users can see results and manage expectations on what works
- Behaviors available at: https://github.com/webrecorder/browsertrix-behaviors
- *Collaborators and Contributions Welcome!*



Webrecorder

Web archiving for all!

THANK YOU

info@webrecorder.net (Email)

webrecorder (GitHub)

Subscribe (RSS)

webrecorder (YouTube)

webrecorder_io (Twitter)

Discuss (Forum)

https://webrecorder.net/