Awesome Web Archiving

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 scale of the Web. Ever-evolving Web standards require continuous evolution of archiving tools to keep up with the changes in Web technologies to ensure reliable and meaningful capture and replay of archived web pages.

Contents

Training/Documentation

- Introductions to web archiving concepts:
  - What is a web archive? - A video from the UK Web Archive YouTube Channel

Channel

    - Wikipedia's List of Web Archiving Initiatives
    - Glossary of Archive-It and Web Archiving Terms
    - The Web Archiving Lifecycle Model - The Web Archiving Lifecycle Model is an attempt to incorporate the technological and programmatic arms of the web archiving into a framework that will be relevant to any organization seeking to archive content from the web. Archive-It, the web archiving service from the Internet Archive, developed the model based on its work with memory institutions around the world.
    - Training materials: module for beginners (8 sessions)
    - UNT Web Archiving Course 2022
    - Continuing Education to Advance Web Archiving (CEDWARC)
- The WARC Standard:
    - The warc-specifications community HTML version of the official specification and hub for new proposals.
    - The offical ISO 28500 WARC specification homepage.
- For researchers using web archives:
    - GLAM Workbench: Web Archives - See also this related blog post on 'Asking questions with web archives'.
    - Archives Unleashed Toolkit documentation
    - Tutorial for Humanities researchers about how to explore Arquivo.pt

Resources for Web Publishers

These resources can help when working with individuals or organisations who publish on the web, and who want to make sure their site can be archived.

- Stanford Libraries' Archivability pages
- The Archive Ready tool, for estimating how likely a web page will be archived successfully.

Tools & Software

This list of tools and software is intended to briefly describe some of the most important and widely-used tools related to web archiving. For more details, we recommend you refer to (and contribute to!) these excellent resources from other groups:

- Comparison of web archiving software

- Awesome Website Change Monitoring

Acquisition

- ArchiveBox - A tool which maintains an additive archive from RSS feeds, bookmarks, and links using wget, Chrome headless, and other methods (formerly `Bookmark Archiver`). *(In Development)*
- archivenow - A Python library to push web resources into on-demand web archives. *(Stable)*
- ArchiveWeb.Page - A plugin for Chrome and other Chromium based browsers that lets you interactively archive web pages, replay them, and export them as WARC data. Also available as an Electron based desktop application.
- Auto Archiver - Python script to automatically archive social media posts, videos, and images from a Google Sheets document. Read the article about Auto Archiver on bellingcat.com.
- Browsertrix Crawler - A Chrome based high-fidelity crawling system, designed to run a complex, customizable browser-based crawl in a single Docker container.
- Brozzler - A distributed web crawler (爬虫) that uses a real browser (Chrome or Chromium) to fetch pages and embedded urls and to extract links. *(Stable)*
- Cairn - A npm package and CLI tool for saving webpages. *(Stable)*
- Chronicler - Web browser with record and replay functionality. *(In Development)*
- crau - crau is the way (most) Brazilians pronounce crawl, it's the easiest command-line tool for archiving the Web and playing archives: you just need a list of URLs. *(Stable)*
- Crawl - A simple web crawler in Golang. *(Stable)*
- crocoite - Crawl websites using headless Google Chrome/Chromium and save resources, static DOM snapshot and page screenshots to WARC files. *(In Development)*
- DiskerNet - A non-WARC-based tool which hooks into the Chrome browser and archives everything you browse making it available for offline replay. *(In Development)*
- F(b)arc - A commandline tool and Python library for archiving data from Facebook using the Graph API. *(Stable)*
- freeze-dry - JavaScript library to turn page into static, self-contained HTML document; useful for browser extensions. *(In Development)*
- grab-site - The archivist's web crawler: WARC output, dashboard for all crawls, dynamic ignore patterns. *(Stable)*

dynamic ignore patterns. *(Stable)*

- Heritrix - An open source, extensible, web-scale, archival quality web crawler. *(Stable)*
  - Heritrix Q&A - A discussion forum for asking questions and getting answers about using Heritrix.
  - Heritrix Walkthrough *(In Development)*
- html2warc - A simple script to convert offline data into a single WARC file. *(Stable)*
- HTTrack - An open source website copying utility. *(Stable)*
- monolith - CLI tool to save a web page as a single HTML file. *(Stable)*
- Obelisk - Go package and CLI tool for saving web page as single HTML file. *(Stable)*
- Scoop - High-fidelity, browser-based, single-page web archiving library and CLI for witnessing the web. *(Stable)*
- SingleFile - Browser extension for Firefox/Chrome and CLI tool to save a faithful copy of a complete page as a single HTML file. *(Stable)*
- SiteStory - A transactional archive that selectively captures and stores transactions that take place between a web client (browser) and a web server. *(Stable)*
- Social Feed Manager - Open source software that enables users to create social media collections from Twitter, Tumblr, Flickr, and Sina Weibo public APIs. *(Stable)*
- Squidwarc - An open source, high-fidelity, page interacting archival crawler that uses Chrome or Chrome Headless directly. *(In Development)*
- StormCrawler - A collection of resources for building low-latency, scalable web crawlers on Apache Storm. *(Stable)*
- twarc - A command line tool and Python library for archiving Twitter JSON data. *(Stable)*
- WAIL - A graphical user interface (GUI) atop multiple web archiving tools intended to be used as an easy way for anyone to preserve and replay web pages; Python, Electron. *(Stable)*
- Warcprox - WARC-writing MITM HTTP/S proxy. *(Stable)*
- WARCreate - A Google Chrome extension for archiving an individual webpage or website to a WARC file. *(Stable)*
- Warcworker - An open source, dockerized, queued, high fidelity web archiver based on Squidwarc with a simple web GUI. *(Stable)*
- Wayback - A toolkit for snapshot webpage to Internet Archive, archive.today, IPFS and beyond. *(Stable)*

IPFS and beyond. *(Stable)*

- Waybackpy - Wayback Machine Save, CDX and availability API interface in Python and a command-line tool *(Stable)*
- Web2Warc - An easy-to-use and highly customizable crawler that enables anyone to create their own little Web archives (WARC/CDX). *(Stable)*
- Web Curator Tool - Open-source workflow management for selective web archiving. *(Stable)*
- WebMemex - Browser extension for Firefox and Chrome which lets you archive web pages you visit. *(In Development)*
- Webrecorder - Create high-fidelity, interactive recordings of any web site you browse. *(Stable)*
- Wget - An open source file retrieval utility that of version 1.14 supports writing warcs. *(Stable)*
- Wget-lua - Wget with Lua extension. *(Stable)*
- Wpull - A Wget-compatible (or remake/clone/replacement/alternative) web downloader and crawler. *(Stable)*

Replay

- InterPlanetary Wayback (ipwb) - Web Archive (WARC) indexing and replay using IPFS.
- OpenWayback - The open source project aimed to develop Wayback Machine, the key software used by web archives worldwide to play back archived websites in the user's browser. *(Stable)*
- PyWb - A Python (2 and 3) implementation of web archival replay tools, sometimes also known as 'Wayback Machine'. *(Stable)*
- Reconstructive - Reconstructive is a ServiceWorker module for client-side reconstruction of composite mementos by rerouting resource requests to corresponding archived copies (JavaScript).
- ReplayWeb.Page - A browser-based, fully client-side replay engine for both local and remote WARC files.
- warc2html - Converts WARC files to static HTML suitable for browsing offline or rehosting.

Search & Discovery

- Mink - A Google Chrome extension for querying Memento aggregators while browsing and integrating live-archived web navigation. *(Stable)*
- playback - A toolkit for searching archived webpages from Internet Archive

- playback - A toolkit for searching archived webpages from Internet Archive, archive.today, Memento and beyond. *(In Development)*
- SecurityTrails - Web based archive for WHOIS and DNS records. REST API available free of charge.
- Tempas v1 - Temporal web archive search based on Delicious tags. *(Stable)*
- Tempas v2 - Temporal web archive search based on links and anchor texts extracted from the German web from 1996 to 2013 (results are not limited to German pages, e.g., Obama@2005-2009 in Tempas). *(Stable)*
- webarchive-discovery - WARC and ARC full-text indexing and discovery tools, with a number of associated tools capable of using the index shown below. *(Stable)*
    - Shine - A prototype web archives exploration UI, developed with researchers as part of the Big UK Domain Data for the Arts and Humanities project. *(Stable)*
    - SolrWayback - A backend Java and frontend VUE JS project with freetext search and a build in playback engine. Require Warc files has been index with the Warc-Indexer. The web application also has a wide range of data visualization tools and data export tools that can be used on the whole webarchive. SolrWayback 4 Bundle release contains all the software and dependencies in an out-of-the box solution that is easy to install.
    - Warclight - A Project Blacklight based Rails engine that supports the discovery of web archives held in the WARC and ARC formats. *(In Development)*
    - Wasp - A fully functional prototype of a personal web archive and search system. *(In Development)*
    - Other possible options for builting a front-end are listed on in the `webarchive-discovery` wiki, here.

Utilities

- ArchiveTools - Collection of tools to extract and interact with WARC files (Python).

- cdx-toolkit - Library and CLI to consult cdx indexes and create WARC extractions of subsets. Abstracts away Common Crawl's unusual crawl structure. *(Stable)*
- Go Get Crawl - Extract web archive data using Wayback Machine and Common Crawl. *(Stable)*
- gowarcserver - BadgerDB-based capture index (CDX) and WARC record server, used to index and serve WARC files (Go).

used to index and serve WARC files (Go).

- har2warc - Convert HTTP Archive (HAR) -> Web Archive (WARC) format (Python).
- httpreserve.info - Service to return the status of a web page or save it to the Internet Archive. Returns JSON via browser or command line via CURL using GET (Golang Package). *(Stable)*
- HTTPreserve Workbench - Tool and API to describe the status of a web page encoded in a simple JSON output describing current status, and earliest and latest links on wayback.org. Save a web page to the Internet Archive. Audit lists of URIs and output a CSV with the data described above (Golang). *(In Development)*
- httrack2warc - Convert HTTrack archives to WARC format (Java).
- MementoMap - A Tool to Summarize Web Archive Holdings (Python). *(In Development)*
- MemGator - A Memento Aggregator CLI and Server (Golang). *(Stable)*
- node-cdxj - CDXJ file parser (Node.js). *(Stable)*
- OutbackCDX - RocksDB-based capture index (CDX) server supporting incremental updates and compression. Can be used as backend for OpenWayback, PyWb and Heritrix. *(Stable)*
- py-wasapi-client - Command line application to download crawls from WASAPI (Python). *(Stable)*
- The Archive Browser - The Archive Browser is a program that lets you browse the contents of archives, as well as extract them. It will let you open files from inside archives, and lets you preview them using Quick Look. WARC is supported (macOS only, Proprietary app).
- The Unarchiver - Program to extract the contents of many archive formats, inclusive of WARC, to a file system. Free variant of The Archive Browser (macOS only, Proprietary app).
- tikalinkextract - Extract hyperlinks as a seed for web archiving from folders of document types that can be parsed by Apache Tika (Golang, Apache Tika Server). *(In Development)*
- wasapi-downloader - Java command line application to download crawls from WASAPI. *(Stable)*
- Warchaeology - Warchaeology is a collection of tools for inspecting, manipulating, deduplicating and validating WARC-files. *Stable*
- warcdedupe - WARC deduplication tool (and WARC library) written in Rust. (In Development)
- WarcPartitioner - Partition (W)ARC Files by MIME Type and Year. *(Stable)*

- warcPartitioner - Partition (W)ARC Files by MIME Type and Year. (Stable)
- warcrefs - Web archive deduplication tools. *Stable*
- webarchive-indexing - Tools for bulk indexing of WARC/ARC files on Hadoop, EMR or local file system.
- wikiteam - Tools for downloading and preserving wikis. *(Stable)*

WARC I/O Libraries

- FastWARC - A high-performance WARC parsing library (Python).
- HadoopConcatGz - A Splitable Hadoop InputFormat for Concatenated GZIP Files (and `*.warc.gz`). *(Stable)*
- jwarc - Reading and write WARC files with a typesafe API (Java).
- Jwat - Libraries and tools for reading/writing/validating WARC/ARC/GZIP files (Java). *(Stable)*
- node-warc - Parse WARC files or create WARC files using either Electron or chrome-remote-interface (Node.js). *(Stable)*
- Sparkling - Internet Archive's Sparkling Data Processing Library. *(Stable)*
- Unwarcit - Command line interface to unzip WARC and WACZ files (Python).
- Warcat - Tool and library for handling Web ARChive (WARC) files (Python). *(Stable)*
- warcio - Streaming WARC/ARC library for fast web archive IO (Python). *(Stable)*
- warctools - Library to work with ARC and WARC files (Python).
- webarchive - Golang readers for ARC and WARC webarchive formats (Golang).

Analysis

- Archives Research Compute Hub - Web application for distributed compute analysis of Archive-It web archive collections. *(Stable)*
- ArchiveSpark - An Apache Spark framework (not only) for Web Archives that enables easy data processing, extraction as well as derivation. *(Stable)*
- Archives Unleashed Notebooks - Notebooks for working with web archives with the Archives Unleashed Toolkit, and derivatives generated by the Archives Unleashed Toolkit. *(Stable)*
- Archives Unleashed Toolkit - Archives Unleashed Toolkit (AUT) is an open-source platform for analyzing web archives with Apache Spark. *(Stable)*
- Common Crawl Columnar Index - SQL-queryable index, with CDX info plus language classification. *(Stable)*
- Common Crawl Web Graph - A host or domain-level graph of the web, with ranking information. *(Stable)*

ranking information. *(Stable)*

- Common Crawl Jupyter notebooks - A collection of notebooks using Common Crawl's various datasets. *(Stable)*
- Tweet Archvies Unleashed Toolkit - An open-source toolkit for analyzing line-oriented JSON Twitter archives with Apache Spark. *(In Development)*
- Web Data Commons - Structured data extracted from Common Crawl. *(Stable)*

Quality Assurance

- Chrome Check My Links - Browser extension: a link checker with more options.
- Chrome link checker - Browser extension: basic link checker.
- Chrome link gopher - Browser extension: link harvester on a page.
- Chrome Open Multiple URLs - Browser extension: opens multiple URLs and also extracts URLs from text.
- Chrome Revolver - Browser extension: switches between browser tabs.
- FlameShot - Screen capture and annotation on Ubuntu.
- PlayOnLinux - For running Xenu and Notepad++ on Ubuntu.
- PlayOnMac - For running Xenu and Notepad++ on macOS.
- Windows Snipping Tool - Windows built-in for partial screen capture and annotation. On macOS you can use Command + Shift + 4 (keyboard shortcut for taking partial screen capture).
- WineBottler - For running Xenu and Notepad++ on macOS.
- xDoTool - Click automation on Ubuntu.
- Xenu - Desktop link checker for Windows.

Curation

- Zotero Robust Links Extension - A Zotero extension that submits to and reads from web archives. Source on GitHub. Supercedes leonkt/zotero-memento.

Community Resources

Other Awesome Lists

- Web Archiving Community
- Awesome Memento
- The WARC Ecosystem
- The Web Crawl section of COPTR

Blogs and Scholarship

- IIPC Blog
- Web Archiving Roundtable - Unofficial blog of the Web Archiving Roundtable of the Society of American Archivists maintained by the members of the Web Archiving Roundtable.
- The Web as History - An open-source book that provides a conceptual overview to web archiving research, as well as several case studies.
- WS-DL Blog - Web Science and Digital Libraries Research Group blogs about various Web archiving related topics, scholarly work, and academic trip reports.
- DSHR's Blog - David Rosenthal regularly reviews and summarizes work done in the Digital Preservation field.
- UK Web Archive Blog

Mailing Lists

- Common Crawl
- IIPC
- OpenWayback
- WASAPI

Slack

- IIPC Slack - Ask @netpreserve for access.
- Archives Unleashed Slack - Fill out this request form for access to a researcher group of people working with web archives.
- Archivers Slack - Invite yourself to a multi-disciplinary effort for archiving projects run in affiliation with EDGI and Data Together.

Twitter

- @NetPreserve - Official IIPC handle.
- @WebSciDL - ODU Web Science and Digital Libraries Research Group.
- #WebArchiving
- #WebArchiveWednesday

Releases