Sustainable Web Archiving at Scale: An Introduction

June 21-22, 2021

DPOE-N
Digital Preservation Outreach
& Education Network

Glossary of Terms

Please note that many web archiving terms are related to their respective tools. Items marked with an asterisk (*) are from the Internet Archive's Glossary of Archive-It and Web Archiving Terms. Items from Conifer User Guide are marked with (^{MC}). Items from Webrecorder Project are marked with ($^{+}$).

Active*

Description of a collection or seed that can have, or be included in, scheduled crawls.

Archive*

Verb. The process of copying digital information into a repository for storage, preservation, and access purposes. In web archiving, often synonymous with capture.

Archive-It*

Archive-It is the leading web archiving application used by a wide range of organizations, including academic, federal, state or local libraries, archives, and other cultural heritage institutions to create, store, and provide access to collections of web content.

In addition to the core functionality of capturing and preserving web-based content, the Archive-It web application allows users to add, import, and export descriptive metadata, and allows for public browsing and full-text search via archive-it.org. Archive-It also provides APIs and other tools to facilitate external integrations with local websites and repositories or third-party discovery or preservation storage services.

Autopilot^{MC}

Autopilot can perform actions on the current web page loaded in Conifer, similar to a human user. It can be activated via the Autopilot button on the top right during capture. When the button is solid green, a specialized behavior is available for the current web page. A white button indicates that only the default behavior will be presented instead.

A specialized behavior performs actions specific to the currently presented website.

The default behavior tries to perform generally useful actions: scrolling down and triggering play embedded media.

Brozzler*

A distributed web crawler that uses a real browser (chrome or chromium) to fetch pages and embedded urls and to extract links. It also uses youtube-dl to enhance media capture capabilities.

Collection*^{MC}

In Archive-It, a group of archived web documents curated around a common theme, topic, or domain.

In Conifer, a collection typically contains multiple captured sessions (see Sessions).

Conifer^{MC}

Conifer is a web archiving service that creates a high-fidelity interactive copy of any web page that you browse manually, including content revealed by your interactions such as playing video and audio, scrolling, clicking buttons, and so forth. Conifer is an online service based on Webrecorder software.

Crawl*

A web archiving (or "capture") operation that is conducted by an automated agent, called a crawler, a robot, or a spider. Crawls identify materials on the live web that belong in your collections, based upon your choice of seed URLs and scope. Crawl can also reference the archived content associated with the action.

Crawl budget*

The amount of data that may be collected at a given subscription level.
Crawl frequency

The rate at which you set your seeds to be crawled. The frequency is on a per seed basis and can be set to one time, twice daily, daily, weekly, monthly, bi-monthly, quarterly, semiannual, or annual.

Crawler*

Explores the web and collects data about its contents. A crawler can also be configured to capture web-based resources. It starts a capture process from a seed list of entry-point URLs (EPUs).

Curator*

Anyone responsible for building a collection or collections of web-based resources, including those who specify seed lists for specific crawls.

Data De-duplication*

If a document in a collection hasn’t changed since it was last captured, the data from that document will not be captured a 2nd time. This is indicated in the All and New data columns in reports.

This function also exists for the ArchiveWeb.page tool (see Page-oriented Archiving).[+]

Directory*

Segments of a host domain in which individual files and/or further directories can be found. Similar to how folders are used to organize content in your computer's local structure. Most, but not all websites use a directory structure. More information: http://www.linfo.org/directory.html.

Document*

Any file with a unique URL - html, image, PDF, video, etc.

Domain*

The root of a host name, for example: .com, .gov, .org, etc.
Dublin Core

The metadata standard used by Archive-It. This standard has 15 fields that can be used to describe any kind of digital artifact, in this case an archived web page. More information: http://www.dublincore.org/documents/dces/.

Dynamic*

Description of web-based content created automatically by software at the web server end. May be (a) personalized for the user based on identification via login or based on cookies stored on the user's computer, (b) tailored to fulfill a specific request made by the user, or (c) code-generated (e.g., using php, jsp, asp, or xml). Information used for personalization or tailoring of pages may be retrieved in real-time from a database or other data store.

Elasticsearch*

An open source search engine utilized by Archive-It to make archived websites text searchable. More information: https://www.elastic.co/guide/en/elasticsearch/guide/current/index.html.

Flash[MC]

Flash, like various web technologies before it, was a popular way to build websites, games, and other types of interactive content online. Flash slowly fell out of favor and was ultimately deprecated. The Chrome web browser is scheduled to remove support for Flash websites on December 31st, 2020. As long as a Flash site remains online it will still be accessible and able to be archived using Conifer’s remote pre-configured browsers.

ArchiveWeb.page embeds the Ruffle emulator, allowing users to archive and replay Flash-based works. Ruffle is automatically enabled on pages that have Flash.[+]

Hadoop*

A computing framework used by Archive-It in order to process, index, and distribute storage of our partners' archived data.

Heritrix*

The name of Internet Archive's open-source, extensible, web-scale, and archival-quality web crawler project. An archaic word for heiress (woman who inherits). More information: http://crawler.archive.org/.

Host*

Where web content is stored, or a single networked machine, as usually designated by its Internet host name (ex. archive.org). The host name can be identical to a URL's domain name, but not always.

Importing^{MC}

There are many web archives which preserve and provide web content and make it publicly accessible. With Conifer, you may import pages from a public web archive into your own collections.

Inactive*

Description of a collection or seed that does not undergo regular, scheduled crawling, but which may at the partner's discretion remain publicly visible/searchable.

Internet Archive*

A non-profit digital library seeking to provide universal access to all knowledge. Archive-It is a subscription service of the Internet Archive: https://archive.org.

Lists^{MC}

In Conifer, lists can be used to organize entry points into a collection. Privacy settings allow you disable public access to a collection, share entry points to all pages of a collection, or just specific lists.

MIME*

Stands for Multipurpose Internet Mail Extensions. This is a specification for formatting non-text content to be sent over the Internet. A MIME file can be just about any kind of non-text file, ex: gif, jpg, html, etc. Archive-It provides a MIME report of all the different types of files archived during each crawl.

One Page*

A crawling protocol that directs our crawler to only archive a given seed URL as a single web page, to include content necessary to render that page faithfully, but not to include any links to other web pages.

One Page Plus*

A crawling protocol which captures one document external to your crawl's default scope if there is a link to it from an in-scope page scheduled to be crawled.

Page-oriented Archiving[+]

In ArchiveWeb.page, the smallest unit is the page. The extension archive keeps track of which resources are loaded from which page. This allows for individual pages to be downloaded, and deleted, as necessary, and will help ensure archived pages are accurately replayed. Resources shared across multiple pages are automatically deduplicated to save storage.

This is a bit different than in Conifer and Webrecorder Desktop, where the smallest unit was a session and individual pages could not be deleted or separated. Support for removing individual pages was an oft-requested feature, and this is now available in ArchiveWeb.page.

Peer-to-peer (IPFS)[+]

The ArchiveWeb.page extension includes experimental peer-to-peer sharing of web archives, using IPFS. This features allows users to share a web archive collection from directly from their browser.

ReplayWeb.page has been updated to support loading web archives directly from IPFS, allowing shared archives from the extension to be quickly shared with others, without having to download and send full WACZ files.

Patch Crawl*

In Archive-It, a crawl to capture and patch in documents that were not captured in your original crawl.

Patching

The process of manually patching in missing content using the patching tools that are available via either Archive-It or Conifer. This can also be accomplished with ArchiveWeb.page by archiving the missing content separately and exporting extant and newly captured content in the same WARC/WACZ file.

Persistent name*

A unique name assigned to a web-based resource that will remain unchanged regardless of movement of the resource from one location to another or changes to the resource's URL. Persistent names are resolved by a third party that maintains a map of the persistent name to the current URL of the resource.

Quality Assurance (QA)*

Generally, quality assurance is any process by which web archiving activities are verified to have captured the intended content and that the desired significant properties are intact. Quality assurance involves utilizing the same tools that are used for harvest to patch in any missing files and to mitigate any replay issues encountered.

Regular Expression (Regex)*

Patterns used to match character combinations in strings in order to find and replace strings that take a defined format.

Remote Browser^{MC}

Remote browsers are pre-configured browsers running in the cloud, with computing resources shared across all Conifer users.

Repository*

The physical storage location and medium for one or more digital archives. A repository may contain an active copy of an archive (i.e. one that is accessed by end users) or a mirror copy of an archive for disaster recovery.

robots.txt*

Files that a site owner can add to their site to keep crawlers from accessing all or parts of it. In some cases a web developer may add these to a site without the owner's knowledge.

Scope*

What the crawler will capture and what it won't. Scoping refers to options for telling the crawler how much or how little of a seed URL to capture. Archive-It options include seed and collection level scoping.

Seed*

An item in Archive-It with a unique ID number. The Seed URL tells the crawler where to go on the live web and acts as an access point to archived content.
Seed Type

A crawling protocol that tells the crawler how many links to follow off of a seed URL. Options are Standard, Standard Plus, One Page, or One Page Plus.

Seed URL*

The starting point URL for a crawler and access point to archived collections.

Sessions^{MC}

Instead of thinking about the web as remote files, Conifer works with sessions. During a session, requests sent by the browser and responses from the web are captured while you are interacting with sites. Sessions are the smallest unit of data in Conifer.

SOLR*

Open source search platform that enabled metadata-based search for Archive-It.

Standard*

Seed Type: A crawling protocol that directs our crawler to archive your seed URLs with its default scoping rules.

Crawling Technology: Heritrix (H3) and Umbra

Standard Plus*

A crawling protocol that directs our crawler to archive your seed URLs as it would with default scoping rules and the additional ability to include any otherwise "Out of Scope" external content directly linked from those seed URLs.

Sub-domain*

A directory named before the root web address, for example crawler.archive.org, in which crawler is the sub-domain.

Umbra*

A browser-based technology that Archive-It uses to navigate the web more as human viewers experience it during the crawl process.

URL*

Stands for Uniform Resource Locator. The location of a resource on the web.
WARC File

An open source format developed by the Internet Archive and an ISO standard (CD 28500) for web archives. Made of disaggregated (coming from different hosts) WARC records. Typically 1GB or lower in data volume each.

WARC Record*

Represents the capture of a distinct URL within a larger WARC File container. Records the archive date, content type, and archive length, as well as the raw byte stream.

Wayback Machine*

Internet Archive's general/global web archive. The Wayback Machine is a piece of software that makes archived websites visible as if they were on the live web.

Web archive*

A collection of web-published materials that an institution has either made arrangements for or has accepted long-term responsibility for preservation and access in keeping with an archive's user access policies. Some of these materials may also exist in other forms but the web archive captures the web versions for posterity.

Web Archive Collection Zipped (WACZ) Format^{+}

The WACZ format is an experimental web archiving format in development by the Webrecorder project. The goal of this spec is to provide a portable format for web archives, to address key social and technical issues:

- Social: to provide an interoperable way to share web archive collections, including any data necessary to make web archives useful to humans.
- Technical: to provide an efficient way to load small amounts of data from a remotely hosted web archive on static storage, without downloading the entire collection.

Web Archiving Service*

Enables curators to build collections of web-published materials that are stored in either local and/or remote repositories. The service includes a set of tools for selection, curation, and preservation of the archives. It also includes repositories for storage, preservation services (e.g., replication, emulation, and persistent naming), and administrative services (e.g., templates for collection strategies, content provider agreements, repository provider agreements). Archive-It is a web archiving service.

Webrecorder

Webrecorder is a suite of open source projects and tools to capture high-fidelity interactive websites and replay them at a later time as accurately as possible. Formerly Webrecorder.io.

Website*

A website is a collection of related web resources, usually as grouped by some common addressing – as when all resources on a single host, or group of related hosts, are considered a 'website'.