## Descriptive Metadata for Web Archiving

# Review of Harvesting Tools

Mary Samouelian and Jackie Dooley

OCLC

# Descriptive Metadata for Web Archiving: Review of Harvesting Tools

**Mary Samouelian**
Harvard University

**Jackie Dooley**
OCLC Research

February 2018

OCLC Research
Dublin, Ohio 43017 USA
www.oclc.org

ISBN: 978-1-55653-014-2
doi: 10.25333/C37H0T
OCLC Control Number: 1021288422

**ORCID iDs**
Mary Samouelian https://orcid.org/0000-0002-0238-5405
Jackie Dooley https://orcid.org/0000-0003-4815-0086

**Please direct correspondence to:**
OCLC Research
oclcresearch@oclc.org

**Suggested citation:**
Samouelian, Mary, and Jackie Dooley. 2018. *Descriptive Metadata for Web Archiving: Review of Harvesting Tools*. Dublin, OH: OCLC Research. doi:10.25333/C37H0T.

## ACKNOWLEDGMENTS

The OCLC Research Library Partnership Web Archiving Working Group used several subgroups to make our range of research investigations possible. The Tools Subgroup defined the scope of work, selected and analyzed the 11 tools, and prepared the grids that articulate the nature of each tool in some detail.

In addition to Mary Samouelian, the members of the Tools Subgroup were:

- Jason Kovari, Cornell University
- Dallas Pillen, University of Michigan
- Lily Pregill, Getty Research Institute
- Matthew McKinley, California Digital Library

Several individuals contributed helpful comments during the external review period. Jefferson Bailey (Internet Archive) and Dragan Espenschied (Rhizome) provided extensive input about both the tools they manage and other aspects of the draft report. WAM member Deborah Kempe (Frick Art Reference Library) reviewed the draft thoroughly and provided helpful suggestions.

Special thanks are due to the developers and managers of the 11 tools that we analyzed, particularly those who provided feedback on our draft analyses. They are the dedicated experts who truly make web archiving possible.

Finally, generous colleagues in OCLC Research were indispensable contributors. Program Officer Dennis Massie provided wise counsel and ubiquitous support to the working group throughout the project. Karen Smith-Yoshimura, Roy Tennant, and Bruce Washburn all contributed insightful comments. Profuse thanks also are due to those who efficiently shepherd every publication through the production process: Erin M. Schadt, Jeanette McNicol and JD Shipengrover.

## CONTENTS

## INTRODUCTION

The OCLC Research Library Partnership Web Archiving Metadata Working Group (WAM) was formed to recommend descriptive metadata best practices for archived web content.[1] When the group began its work early in 2016, we discovered that metadata practitioners had high hopes that it would be possible to extract descriptive metadata from harvested content. This report offers our objective analysis of 11 tools in pursuit of an answer to that question.

We reviewed selected web harvesting tools to determine their descriptive metadata functionalities. The question we sought to answer was this: Can web harvesting tools automatically generate descriptive metadata that supports the discoverability of archived web resources? Auto-generation of descriptive metadata for archived web resources could result in significant gains in the efficiency of data entry and thus help enable metadata production at scale.

Our intent was twofold: 1) provide the web archiving community with a description of each relevant tool's overall purpose and metadata-related capabilities, and 2) inform WAM's overarching objective of preparing best practice recommendations for web archiving descriptive metadata based on an understanding of user needs.

What is descriptive metadata? Its ultimate purpose is resource discovery. It describes the content of a resource, associates various access points and describes its relationship to other resources.[2] Archives, libraries and museums rely on descriptive metadata to enable users to locate, distinguish and select resources of all types. Metadata creation has often been found to be the most expensive activity in preparing resources for use.

In today's library and archives environment, descriptive metadata often is repurposed for use in multiple discovery systems. In the context of archived websites, this may include library catalogs, archival finding aids, and standalone platforms for delivering web content.

At the outset, we were skeptical that the level of automated metadata generation needed to support the discoverability of web archive resources would be feasible due to an inherent limitation: only bare-bones descriptive metadata is recorded in the headers of most web pages. This prediction was borne out by our reviews. Nevertheless, the analysis of tools appended to the report can serve as a useful resource to understand the landscape of harvesting tools available for web archiving.

This report is one of a complementary trio being issued simultaneously to document the work of the WAM Working Group. Its siblings are *Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group*[3] and *Descriptive Metadata for Web Archiving: Literature Review of User Needs.*[4]

# Methodology

We began by examining two lists of web archiving tools as the starting point for identifying relevant tools: one compiled by the International Internet Preservation Consortium (IIPC)[5] and the other by WAM member Rosalie Lack.[6] We filtered the lists to retain only those tools that harvest or replay web content, are actively under development and/or are actively supported, and appeared to include descriptive metadata capture features.[7]

Given the open-source nature of tools in this realm, it is challenging to provide continued support and development for particular tools. Some of the tools on our initial list are no longer supported and so were discarded from consideration.

Ultimately, we analyzed these 11 tools:

- Archive-It
- Heritrix
- HTTrack
- Memento
- Netarchive Suite
- SiteStory
- Social Feed Manager
- Wayback Machine
- Web Archive Discovery
- Web Curator Tool
- Webrecorder

We developed seven criteria for evaluating each tool to ensure consistency in our approach:

1. What is the basic purpose of the tool and its core functionalities? (e.g., capture, display and/or administrative layer)
2. What objects/files can it take in and generate? (i.e., the atomic unit that the tool creates or alters, such as Mementos, WARCs (Web ARChives) or PDFs)
3. In which metadata profiles does it record?
4. Which descriptive elements are automatically generated?
5. Which descriptive elements can be created or edited by the user?
6. Which descriptive data elements can be exported for use outside of the tool?
7. What relation does it have to other tools? (e.g., Heritrix gathers metadata that is embedded in a WARC file, some of which is used by Archive-It.)

We did not investigate the tools' capabilities for generating technical or preservation metadata.

# Brief Description of Each Tool

A summary of each tool follows to highlight basic functionality, including what we learned about metadata capabilities.

After preparing a description of each tool's purpose and functionality, we contacted its owner to obtain feedback and thus ensure the accuracy of our descriptions. We are grateful to all who responded; their suggestions significantly improved upon our work. Our full review of the tools is in the appendix.

**Archive-It:** This is a widely used subscription web archiving service from the Internet Archive (IA) that harvests websites using a variety of capture technologies including IA's open-source web crawler Heritrix. WARC files[8] are preserved in the IA digital repository and can be downloaded by users for preservation in their own repositories.[9] The "grab title" feature at the seed level automatically scrapes title metadata from each page. Archive-It provides 16 Dublin Core metadata fields from which users can choose, as well as the ability to add custom fields that can be added manually at the collection, seed and document levels.[10]

**Heritrix:** A widely used, open-source web crawler developed by the IA, Heritrix is one of the principal capture tools used by IA and by many others for harvesting websites. It produces WARCs but does not allow for the input or generation of additional descriptive metadata within the tool.[11]

**HTTrack:** An open-source capture tool that uses an off-line browser utility to download a website to a directory, generates a folder hierarchy and saves content that mirrors the original website structure. HTTrack produces basic log files and stores some technical metadata in its cache, but it does not generate WARCs or ARCs. It also does not allow for input of any descriptive metadata.[12]

**Memento group:** Open-source tooling built on Memento protocols, such as MediaWiki Memento Extension and Memento TimeTravel, that facilitates access to archived websites through http content negotiation based on capture date. Current tooling built on this protocol focuses on browsing of archived content based on URL and capture date. At this time, it lacks functionality to capture descriptive metadata for enhanced discovery.[13]

**NetarchiveSuite:** This open-source software is used to plan, schedule and run web crawls. NetArchiveSuite consists of several modules, including a harvester module (which uses Heritrix) for defining, scheduling and running crawls; an archive module that serves as a preservation repository for harvested material; and an access module for viewing harvested material. NetArchiveSuite produces WARCs.[14]

**SiteStory:** An open-source transactional tool that captures every version of a resource as it is requested by a web browser, thus replicating the user's experience. The resulting archive is effectively representative of a web server's entire history, although versions of resources that are never requested by a browser are not included. The only metadata automatically generated is the "http" metadata associated with the web server. SiteStory creates Mementos of each resource served by the web server it is associated with irrespective of the media of the resource. These Mementos can be off-loaded to WARC files and then uploaded into an instance of the Internet Archive's Wayback software.[15]

**Social Feed Manager:** An open-source web application, still under active development, that allows users to create collections of data from social media platforms, including Twitter, Tumblr and Flickr. The application harvests the data, including images and web pages linked from or embedded in the social media content, and uses Heritrix via public APIs. Consistent descriptive metadata such as creator and

date are inherently embedded in social media posts and are automatically generated. Other descriptive metadata pertaining to selection of a collection (including title and a description) or set of collections must be manually generated.[16]

**Wayback:** Released by the Internet Archive in 2005, Wayback is software that replays archived webpages. It consists of a group of interrelated applications "under the hood" that index and retrieve captured web content, render WARC and ARC files, and display the content in a web-based user interface. The next-generation OpenWayback was launched in 2013. The Wayback Machine can be searched only by URL, but the Beta Wayback Machine, released in October 2016, expands search capabilities, including the ability to perform a keyword search against an index of terms contained within the archive. Descriptive elements automatically generated during archiving are username (labeled "creator"), title, capture date/time, archive file format, title of collection (if defined by user) and URL.[17]

**Web Archive Discovery:** This open-source tool enables full-text search within web archives by indexing WARC or ARC files, parsing their contents and posting the results in JSON format to an Apache SOLR server. The resulting records can then be parsed either by using a basic front-end interface that is included with the tool or by implementing a custom interface atop the SOLR instance. The only descriptive metadata captured are crawl date, crawl year and Wayback date.[18]

**Web Curator Tool:** This open-source workflow management tool is designed to enable non-technical users to manage the web archiving process. Workflow can be customized to include a variety of specialized tasks to support web acquisition and description, including permission/authorization, selection/scoping/scheduling, harvesting, quality review and archiving. Harvest date is automatically captured, while all other descriptive metadata must be added to by the user.[19]

**Webrecorder:** An open-source tool that captures the exact sequence of navigation through a series of web pages or digital objects, thus preserving the user's experience. The recording produces WARC files with a high level of detail that can then be replayed using Webrecorder. Descriptive metadata elements include username (labeled "creator"), title, capture date/time, archive file format, title of collection (if included in the metadata) and all URLs the user visited during a recording session. Descriptive metadata is automatically generated during archiving and embedded in a downloadable WARC file. Future development plans include the ability to add user-created metadata and support for page-level annotations.[20]

# Analysis

As predicted, we found that capabilities for storing descriptive metadata elements and automatic metadata generation capabilities vary widely and tend to be minimal. This could possibly be changing as new tools are developed and existing ones enhanced. The owners of several tools that are in early stages of development are actively eliciting feedback to determine which elements would be valuable to their users.

Many tools generate WARC files, since this ISO standard is widely used by the web archiving community. Inherent relationships among tools and the ability to export at least some data elements (including administrative and technical metadata) are also fairly common. Widespread use of tools such as Heritrix also contribute to dependencies between tools.

We derived several generalizable conclusions:

- Most tools built for web archives focus on capturing and storing technical metadata for accurate transmission and re-creation but capture minimal descriptive metadata. This is at least partially because so little descriptive metadata exists in the captured files. It therefore must usually be created manually, either within the tool or externally.

- The hope for auto-generation may be fruitless unless or until creators of textual web pages routinely embed more metadata that can be made available for capture.

- The title of a site (as recorded in its metadata) and the date of capture are routinely recorded, but it may not be possible to extract them automatically. Titles are sometimes unhelpful, such as “home page” or “title.”

- Not all tools define descriptive metadata in the same way.

Clearly, development of strong metadata functionality in tools to enable automatic generation of descriptive metadata warrants further development, given the need expressed by librarians and archivists. Factors discussed above, particularly the lack of metadata embedded in source files and a completely appropriate focus on capturing and storing technical and preservation metadata for accurate re-creation and transmission, lead us to ask, however, whether the harvesting stage is the most appropriate part of the web archiving process during which to add descriptive metadata for most types of content. Given the lack of metadata features in current web archiving tools, perhaps there is a clearer path forward for approaches that leverage external services and APIs.

That said, it is critical that members of the library, archives and museum community who seek to build web archives become sufficiently knowledgeable to assist tool builders in addressing needs for generating descriptive metadata for discovery of archive web resources. Ongoing development of new tools with new functionalities demonstrates a vibrant community that continues to enable more efficient and effective harvesting of web content and generation of metadata.

## APPENDIX: FULL REVIEW OF TOOLS

### Archive-It

URL: https://www.archive-it.org/

Tool Owner: Lori Donovan

| Criteria for Evaluating Tool | Description |
|---|---|
| **What is the basic purpose of the tool and its core functionalities?** | Archive-It is a subscription web archiving service from the Internet Archive (IA) and has multiple purposes. It is a capture tool using IA's open-source web crawler Heritrix; it is an administrative tool that allows partners to collect, catalog and manage its archived collections using a web application; and it preserves partners' web archives by storing (and having the ability to download) its WARC files that contain web-captured content in IA's digital repositories. The public-facing website displays the archived websites using the Wayback Machine and allows visitors to browse and search the content of partners' public collections. The metadata in the public-facing website is both browsable and searchable. |
| **What objects/files can it take in and generate?** | The tool generates WARC files and WAT (Web Archive Transformation) files, which can be used to create data analysis reports based on very large data sets; LGA (Longitudinal Graph Analysis) files; and WANE files, which provide researchers the named entities from each text resource in a web archive. |
| **In which metadata profiles does it record?** | Dublin Core Metadata Element Set |
| **Which descriptive metadata elements are automatically generated?** | The "Grab Title" feature at the seed level automatically scrapes title metadata from the URL's title meta tag. Additionally, preservation and administrative metadata (including capture date, URL, response codes, etc.) is captured automatically during crawling and is available in the WARC files and via cdx (Wayback index files). Archive-It provides 16 standard metadata fields plus customized metadata fields. A user can manually add metadata at the collection, seed and document levels. This includes descriptive metadata (title, creator, subject, description, publisher, contributor, relation and collector), administrative metadata (date, identifier, coverage, rights, language) and technical metadata (type, format). Customized elements are not in Dublin Core. |

| | |
|---|---|
| **Which descriptive metadata elements can be created/edited by the user?** | Adding metadata at the collection level is currently manual. Partners have the capability to import a spreadsheet to make bulk changes to metadata at the seed or document levels by using an .ODS (OpenDocument Spreadsheet) file. |
| **Which descriptive metadata elements can be exported and used outside of the tool?** | A user can make collection and seed-level metadata available to harvesters, including OCLC's WorldCat catalog via an “opt-in” OAI-PMH metadata XML feed. By selecting an export metadata checkbox at the collection level, the collection will be included the next time the feed is picked up. Metadata at the seed and document level can also be exported via .ODS. |
| **What relation does the tool have to other tools?** | Archive-It uses the following tools: Heritrix (web crawler), NutchWax (allows for archived websites to be text searchable), SOLR (provides metadata-based search for Archive-It), and Umbra (works with Heritrix; this allows a client-side script to be executed so that previously unavailable URLs can be detected for Heritrix to crawl; useful for websites with dynamic content like social media). Archive-It also integrates with several preservation and access systems including LOCKSS and DuraCloud. |

## Heritrix

URL: https://webarchive.jira.com/wiki/display/Heritrix/Heritrix

Tool Owners: Jefferson Bailey, Lori Donovan, Noah Levitt

| Criteria for Evaluating Tool | Description |
|---|---|
| **What is the basic purpose of the tool and its core functionalities?** | Heritrix is "the Internet Archive's open-source, extensible, web-scale, archival-quality web crawler project" (taken from the tool description). It is one of the capture tools used by the IA, including its Archive-It service, and is designed to respect the robots.txt exclusion directives and META robot tags, and to collect material at a measured, adaptive pace unlikely to disrupt normal website activity. |
| **What objects/files can it take in and generate?** | The tool generates WARC files. |
| **In which metadata profiles does it record?** | WARC |
| **Which descriptive metadata elements are automatically generated?** | Heritrix captures technical and structural data (e.g., the structure of a website, MIME type and content length), all of which are automatically generated. |
| **Which descriptive metadata elements can be created/edited by the user?** | Most elements are automatically generated. The user can add custom metadata in the WARC info header. |
| **Which descriptive metadata elements can be exported and used outside of the tool?** | All the metadata captured in WARCs can be parsed from the WARCs and used outside of the tool. |
| **What relation does the tool have to other tools?** | Heritrix gathers metadata that is embedded in a WARC file, some of which is used by Archive-It and the IA's Wayback Machine, the latter of which is used to playback WARCs. Heritrix is also used as the web crawler by other web archiving software packages including NetarchiveSuite. Additionally, Heritrix generates logs and reports that can be used in a variety of ways by many external tools. |

## HTTrack

URL: https://www.httrack.com/

Tool Owner: Xavier Roche

| Criteria for Evaluating Tool | Description |
|---|---|
| **What is the basic purpose of the tool and its core functionalities?** | HTTrack is a free open-source capture tool that uses an off-line browser utility to download a website to a directory, which generates a folder hierarchy and saves content that mirrors the original website structure. HTTrack can also update an existing mirrored site, and resume interrupted downloads. |
| **What objects/files can it take in and generate?** | The objects/files generated by HTTrack reflect the objects/files present on the website being mirrored (e.g., HTML, CSS, JavaScript, JPEG, etc.) HTTrack also produces basic log files and stores some information in its cache, including the raw data, MIME type, URL and server's headers related to the capture. It does not generate WARCs or ARCs. |
| **In which metadata profiles does it record?** | Not applicable |
| **Which descriptive metadata elements are automatically generated?** | HTTrack does not allow for the input of any descriptive metadata and does not extract any technical or preservation metadata from mirrored content. The captured content would need to be run through additional tools to produce any technical or preservation metadata and any descriptive metadata would need to be added in a separate collection management or access system. |
| **Which descriptive metadata elements can be created/edited by the user?** | Descriptive metadata is not built into the tool. HTTrack is primarily used for copying/preserving websites. |
| **Which descriptive metadata elements can be exported and used outside of the tool?** | Not applicable |
| **What relation does the tool have to other tools?** | Not applicable |

### Web Archiving Tool: Memento (suite of tools)

Extension: Memento URL: https://www.mediawiki.org/wiki/Extension:Memento

Memento Time Travel URL: http://timetravel.mementoweb.org/

Memento Time Travel API URL: http://timetravel.mementoweb.org/guide/api/

About the Time Travel Service URL: http://timetravel.mementoweb.org/about/

Tool Owner: Owner did not respond to the request for feedback

| Criteria for Evaluating Tool | Description |
|---|---|
| **What is the basic purpose of the tool and its core functionalities?** | The Memento suite of tools facilitates access to archived versions of web content. The Memento protocol allows http content negotiation, such as the ability to navigate between different versions of a web resource based on the capture/crawl date. This suite of tools is more about the discovery of archived sites than crawling new ones. |
| **What objects/files can it take in and generate?** | Mementos |
| **In which metadata profiles does it record?** | Constrained RESTful Environments Link Format (RFC 6690): https://datatracker.ietf.org/doc/rfc6690/ |
| **Which descriptive metadata elements are automatically generated?** | Descriptive metadata is in tooling sitting on top of the Memento protocol and not part of protocol itself. |
| **Which descriptive metadata elements can be created/edited by the user?** | Tooling that allows for capture of descriptive metadata sits on top of the Memento protocol. |
| **Which descriptive metadata elements can be exported and used outside of the tool?** | DateTime and URL/URI |
| **What relation does the tool have to other tools?** | A user can build discovery tools on top of Memento protocol (as evidenced by Time Travel). |

## NetarchiveSuite

URL: https://sbforge.org/display/NAS/NetarchiveSuite

Tool Owner: Owner did not respond to the request for feedback

| Criteria for Evaluating Tool | Description |
|---|---|
| **What is the basic purpose of the tool and its core functionalities?** | The NetarchiveSuite is "a complete web archiving software package" that can be used "to plan, schedule and run web harvests of parts of the Internet." NetarchiveSuite is divided into several modules, including a harvester module that "handles defining, scheduling and performing harvests," an archive module that "makes it possible to setup and run a repository with replication, active bit consistency checks for bit-preservation and support for distributed batch jobs," and an access module that "gives access to previously harvested material through a proxy solution" (quoted phrases are taken from the tool description). |
| **What objects/files can it take in and generate?** | The tool generates WARC files. |
| **In which metadata profiles does it record?** | WARC |
| **Which descriptive metadata elements are automatically generated?** | Technical and structural data (e.g., the structure of a website, MIME type and content length) are both automatically generated. |
| **Which descriptive metadata elements can be created/edited by the user?** | All metadata elements are automatically generated. |
| **Which descriptive metadata elements can be exported and used outside of the tool?** | All of the metadata captured in WARCs can be parsed from the WARCs and used outside of the tool. |
| **What relation does the tool have to other tools?** | NetarchiveSuite uses Heritrix as its web crawler and the WARCs produced by NetarchiveSuite can be viewed in the IAI's Wayback software. |

## SiteStory

URL: http://mementoweb.github.io/SiteStory/

Tool Owner: Herbert Van de Sompel

| Criteria for Evaluating Tool | Description |
|---|---|
| **What is the basic purpose of the tool and its core functionalities?** | SiteStory is an open-source transactional web archive that continues to be actively developed. It captures every version of a resource as it is being requested by a web browser. The resulting archive is effectively representative of a web server's entire history, although versions of resources that are never requested by a browser will also never be archived. It archives resources of a web server it is associated with. As a browser requests a resource published by the server, that resource is delivered to the browser but also pushed into the archive. This tool is more fine-grained than "traditional" web archiving and what triggers it is usage of a page by a web browser. The "collection" is the web server that is being archived. |
| **What objects/files can it take in and generate?** | SiteStory creates Mementos of each resource served by the web server that it is associated with, regardless of the media of the resource. Mementos can be off-loaded to WARC files and can be uploaded into an instance of the AI's Wayback software. |
| **In which metadata profiles does it record?** | There is no metadata profile associated with SiteStory. |
| **Which descriptive metadata elements are automatically generated?** | None. The only metadata being automatically generated is the http metadata associated with the web server. |
| **Which descriptive metadata elements can be created/edited by the user?** | There is no descriptive metadata associated with SiteStory, although Van de Sompel indicated that it would be very simple to add a form to capture some descriptive metadata about the web server and other information. |
| **Which descriptive metadata elements can be exported and used outside of the tool?** | Not applicable |
| **What relation does the tool have to other tools?** | A SiteStory archive is accessible via the Memento protocol. |

## Social Feed Manager

URL: http://gwu-libraries.github.io/sfm-ui/about/overview

Tool Owner: Christie Peterson

| Criteria for Evaluating Tool | Description |
|---|---|
| **What is the basic purpose of the tool and its core functionalities?** | Social Feed Manager (SFM) is a web application that allows users to create collections of data from social media platforms including Twitter, Tumblr, Flickr and Sina Weibo (Chinese microblogging site). The application harvests social media data via public APIs, and images and web pages are linked from or embedded in the social media using Heritrix. Users can export a collection to a spreadsheet, feed data into their processing pipeline from the command line or explore data with Elasticsearch/Logstash/Kibana (ELK). Version 1 was released June 2016; it is a "work in progress" and development is "active and ongoing" (taken from the tool description). |
| **What objects/files can it take in and generate?** | Social media data from APIs is written to WARCs, as are web and media captures via Heritrix. |
| **In which metadata profiles does it record?** | The tool does not record metadata following any established metadata standard, but currently the focus is on provenance metadata. SFM is in active development and the developers are currently taking feedback from early users and testers; they have not yet committed to a metadata profile and are actively exploring what elements can be automatically generated and what can or should be entered by users. |
| **Which descriptive metadata elements are automatically generated?** | Creation (e.g., who authored a tweet, how and where it was posted, etc.), collection (e.g., info about the http request, the seeds and the WARC files) and selection (deciding which social media to collect and which not to collect, including collection name and description). |
| **Which descriptive metadata elements can be created/edited by the user?** | Metadata about social media posts and harvesting is automatically generated, while selection metadata about a collection and set of collections is human generated. This includes a title and description. The creation of the collection/collection set and all following edits and additions are documented in a change log to which comments can be appended. SFM is a collection tool, not a preservation or access tool, and as such does not support robust descriptive metadata of the type that would be used in those systems. Instead, the |

| | |
|---|---|
| | developers have focused on the elements of description that are critical at the collection phase (i.e., provenance metadata), with only the minimal descriptive metadata required for control at that point (i.e., a title). |
| **Which descriptive metadata elements can be exported and used outside of the tool?** | While metadata that is part of the social media data (i.e., creation metadata) is exportable, work is still in progress on an “export manifest” with descriptive data elements about collection and selection. |
| **What relation does the tool have to other tools?** | SFM's primary focus is on harvesting from social media APIs. It makes use of Heritrix for capture of websites and media embedded or referenced in social media. It uses the WARC format for storing harvests; WARCs may be used in a variety of other applications for analysis and playback. |

## Wayback Machine

URL: http://archive.org/web/

Tool Owners: Jefferson Bailey, Lori Donovan

| Criteria for Evaluating Tool | Description |
|---|---|
| **What is the basic purpose of the tool and its core functionalities?** | The Internet Archive's Wayback Machine is a "group of interrelated applications that index captured Web content, retrieve the content and display the content in a web-based user-interface" (quotations taken from the tool description). It replays WARC files from the Internet Archive's collection of over 270 billion web pages. The public interface, also referred to as the general Wayback, can only be searched by URL on the public interface, as opposed to Wayback within the Archive-It service that supports full-text searching. Once a URL has been retrieved, the archive for that website can be browsed by date of capture. Three Wayback Machine APIs are available: Wayback Availability JSON API, Memento API and Wayback CDX Server API. |
| **What objects/files can it take in and generate?** | Wayback reads and renders WARC and ARC files. The Wayback Index is generated by parsing WARC data and creating CDX (capture index) records. The CDX points to the location within the WARC and ARC of the data being replayed and also contains other key metadata. |
| **In which metadata profiles does it record?** | CDX file format |
| **Which descriptive metadata elements are automatically generated?** | Wayback captures descriptive metadata (canonical URL, original URL), administrative metadata (date, redirect URL) and technical metadata (IP, MIME type, HTML response code, MD5 digest, file offset and WARC file name). |
| **Which descriptive metadata elements can be created/edited by the user?** | All descriptive metadata elements are automatically generated. |
| **Which descriptive metadata elements can be exported and used outside of the tool?** | The Wayback CDX Server API allows for complex querying, filtering and analysis of capture data. The following fields are publicly accessible: urlkey, timestamp, original, mimetype, statuscode, digest and length. |
| **What relation does the tool have to other tools?** | Not applicable |

## Web Archive Discovery

URL: https://github.com/ukwa/webarchive-discovery

Tool Owner: Andrew Jackson

| Criteria for Evaluating Tool | Description |
|---|---|
| **What is the basic purpose of the tool and its core functionalities?** | Web Archive Discovery provides full-text search for web archives by indexing/parsing WARC and/or ARC files and posting resulting records to an Apache SOLR server. The SOLR server can then be queried by a front-end GUI (the GUI is currently based on Drupal Sarnia, but soon to be custom built). Web Archive Discovery is intended to be middleware, and users can put their own interface on top of SOLR. This could be useful for end users with technical skills, but it is not intended to be an end-user tool. |
| **What objects/files can it take in and generate?** | WARC, ARC |
| **In which metadata profiles does it record?** | The tool parses WARC into JSON format for SOLR querying. |
| **Which descriptive metadata elements are automatically generated?** | Web Archive Discovery captures descriptive metadata (including crawl_date, crawl_year, wayback_date and content), administrative metadata (including source_file, url, server, host, content_type and hash) and technical metadata (including content, url, url_type, resourcename, content_type_droid and content_type_tika). |
| **Which descriptive metadata elements can be created/edited by the user?** | All metadata elements are automatically generated. |
| **Which descriptive metadata elements can be exported and used outside of the tool?** | All metadata elements can be extracted and exported. |
| **What relation does the tool have to other tools?** | "While search is the primary goal, the fact that we are going through and parsing every byte means that this is a good time to perform any other analysis or processing of interest. Therefore, we have been exploring a range of additional content properties to be exposed via the Solr index. These include format analysis (Apache Tika and DROID), some experimental preservation risk scanning, link extraction, metadata extraction and so on" (taken from the tool description). |

## Web Curator Tool

URL: http://dia-nz.github.io/webcurator/

Tool Owner: Ben O'Brien

| Criteria for Evaluating Tool | Description |
|---|---|
| **What is the basic purpose of the tool and its core functionalities?** | The Web Curator Tool (WCT) is an open-source workflow management application designed to allow non-technical users to manage the selective web archiving process. Harvesting workflow comprises a series of specialized tasks to support web acquisition and description including: permission/authorization, selection/scoping/scheduling, harvesting, quality review and archiving. WCT does not attempt to be a digital repository, access tool, catalog or records management system. |
| **What objects/files can it take in and generate?** | WARC, ARC |
| **In which metadata profiles does it record?** | Optional basic Dublin Core Metadata Schema for "Description" element. |
| **Which descriptive metadata elements are automatically generated?** | Harvest date is automatically calculated and inserted into the dc:date field. All other descriptive metadata must be added to "target" web resource by a user. |
| **Which descriptive metadata elements can be created/edited by the user?** | Web resource name, owner, annotation/notes and Basic Dublin Core metadata in the descriptive field. |
| **Which descriptive metadata elements can be exported and used outside of the tool?** | All metadata is added to WARC/ARC object and will be included when submitted to a digital repository, where it may then be accessed: "When a Target Instance is submitted to an archive, its Target metadata is included in the SIP" (taken from the tool description). |
| **What relation does the tool have to other tools?** | Employs Heritrix to crawl and uses WARC as "atomic unit." Can optionally integrate with Wayback machine and Rosetta DPS. For Rosetta DPS, a plugin adds DNX metadata, some user input and some auto-generated that is used by Rosetta. |

## Webrecorder

URL: https://webrecorder.io/

Tool Owner: Dragan Espenschied

| Criteria for Evaluating Tool | Description |
|---|---|
| **What is the basic purpose of the tool and its core functionalities?** | Webrecorder is a free service that is "a human-centered archival tool to create high-fidelity, interactive, contextual archives of social media and other dynamic content, such as embedded video and complex JavaScript." Unlike other crawlers, Webrecorder archives web content through interactive browsing, capturing the exact sequence of navigation through a series of web pages or digital objects and preserving the unique experience of an individual user at a moment in time. This approach places control of the archive in the hands of the curator. The tool uses the same software to capture and replay the site; this approach is called symmetrical web archiving. New developments in 2016 include the ability to upload external WARC (or ARC) files into a user's Webrecorder collection and the inclusion of the full metadata about the collection included on export of the WARC file. These enhancements allow WARCs to more easily move between systems. |
| **What objects/files can it take in and generate?** | WARC |
| **In which metadata profiles does it record?** | Named fields and JSON |
| **Which descriptive metadata elements are automatically generated?** | All data elements are automatically generated during archiving. The descriptive elements are username (labeled "creator"), title, capture date/time, archive file format, title of collection (if defined by user) and URL. |
| **Which descriptive metadata elements can be created/edited by the user?** | User-created metadata is a future development which may possibly support for page-level annotations. |
| **Which descriptive metadata elements can be exported and used outside of the tool?** | All generated metadata is embedded in downloadable WARC file. |
| **What relation does the tool have to other tools?** | Not applicable |

## NOTES

1. http://oc.lc/wam.
2. For a thorough introduction to descriptive and other types of metadata, including technical, preservation, rights and structural, see: Riley, Jenn. 2017. *Understanding Metadata: What Is Metadata, and What Is It For?* Baltimore, Maryland: NISO Press. http://www.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf.
3. Dooley, Jackie, and Kate Bowers. 2018. *Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group.* Dublin, OH: OCLC Research. doi:10.25333/C3005C.
4. Venlet, Jessica, Karen Stoll Farrell, Tammy Kim, Allison Jai O'Dell, and Jackie Dooley. 2018. *Descriptive Metadata for Web Archiving: Literature Review of User Needs*. Dublin, OH: OCLC Research. doi:10.25333/C33P7Z.
5. International Internet Preservation Consortium. 2017. "Tools and Software." Web Archiving. Accessed 6 January 2017. http://www.netpreserve.org/web-archiving/tools-and-software.
6. Formerly at the California Digital Library, now University of California, Los Angeles.
7. While we did not rely on Harvard Library's Environmental Scan of Web Archiving, its analysis of web archiving tools merits a look; see Truman, Gail. 2016. *Web Archiving Environmental Scan*. Cambridge, MA: Harvard Library. https://dash.harvard.edu/handle/1/25658314.
8. International Organization for Standardization Technical Committee 46, Subcommittee 4. 2006. "Information and Documentation—The WARC File Format." Draft (ISO/DIS 28500) distributed for review and comment. http://bibnum.bnf.fr/WARC/WARC_ISO_28500_version1_latestdraft.pdf.
9. WARCs contain extensive technical and structural metadata about a web capture. Their structure is based on International Standards Organization, "ISO 28500:2009: Information and Documentation—WARC File Format." http://www.iso.org/iso/catalogue_detail.htm?csnumber=44717. Further information is available in the Library of Congress's documentation of digital file formats. https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml.
10. Archive-It: https://support.archive-it.org/hc/en-us/articles/208332603-Add-edit-and-manage-your-metadata.
11. Heritrix: https://webarchive.jira.com/wiki/display/Heritrix/Heritrix.
12. HTTrack: https://www.httrack.com.
13. Memento group: http://timetravel.mementoweb.org/about/.
14. NetarchiveSuite: https://sbforge.org/display/NAS/NetarchiveSuite.
15. SiteStory: https://mementoweb.github.io/SiteStory/.
16. Social Feed Manager: https://social-feed-manager.readthedocs.io/en/master/.
17. Wayback: http://archive.org/web/web.php.
18. Web Archive Discovery: https://github.com/ukwa/webarchive-discovery/wiki.
19. Web Curator Tool: http://webcurator.sourceforge.net/.
20. Webrecorder: https://webrecorder.io/.

For more information about our work related to digitizing library collections, please visit: **oc.lc/digitizing**

OCLC

6565 Kilgour Place
Dublin, Ohio 43017-3395

T: 1-800-848-5878
T: +1-614-764-6000
F: +1-614-764-6096
www.oclc.org/research

ISBN: 978-1-55653-014-2
DOI:10.25333/C37H0T
RM-PR-215938-WWAE 1709