# Table of Contents

Overview .......................................................................................................................... 2
Mission .............................................................................................................................. 2
Scope ............................................................................................................................... 3
Selection ........................................................................................................................... 4
Acquisition ....................................................................................................................... 5
Arrangement ..................................................................................................................... 6
Description ....................................................................................................................... 7
Maintenance and Re-Appraisal ....................................................................................... 8
Presentation and Access .................................................................................................. 9
Appendix A: Glossary ....................................................................................................... 10
Appendix B: References .................................................................................................... 11
Overview

Web archiving is the process of crawling websites for the purpose of copying and saving their content. Ideally, websites should be harvested and saved in their original formats so that they can be rendered as closely as possible to the time it was published on the live web. While systems perform automated crawls, archivists and librarians make appraisal, curatorial, and collection management decisions to implement an organizational strategy that adds context and intellectual control to these activities.

The constant evolution of many websites creates an ephemeral nature to these information resources. Furthermore, much documentation that used to be shared on paper resources are now shared online. Active, controlled participation in web archiving allows institutions to provide access to legacy information that may not be saved otherwise and ensures the preservation of online primary resources.

Mission

The Northwestern University Libraries' Web Archives preserves and provides access to websites, web pages, and/or web content of enduring value for the Distinctive Collections unit, including:

- University Archives
- Charles Deering McCormick Library of Special Collections
- Melville J. Herskovits Library of African Studies
- Music Library
- Transportation Library

Web archiving supports the curation of unique resources at Northwestern University Libraries, enriches learning by packaging historic websites in a meaningful manner, and establishes and/or enhances collaborative web archiving activities with internal and external partners.
Scope

The Northwestern University Libraries’ Web Archives includes web content which can be categorized within the collecting scope of the University Archives, Charles Deering McCormick Library of Special Collections, Melville J. Herskovits Library of African Studies, Music Library, or Transportation Library.

Web archives can be classified into two categories:

1. Archival collection

Includes websites, web pages, and/or web content which are included, in part or whole, of an archival collection’s *fonds* by one of the Distinctive Collections units: University Archives, Charles Deering McCormick Library of Special Collections, Melville J. Herskovits Library of African Studies, Music Library, or Transportation Library. Examples include, but are not limited to, websites, social media accounts, blogs, etc., created, managed, and/or curated by the creator(s) of the archival collection.

2. Curated cultural heritage

Includes web content which has been strategically curated as a set of cultural heritage resources grouped thematically. Subject specialists within the University Archives, Melville J. Herskovits Library of African Studies, Music Library, or Transportation Library researches, identifies, and curates a selection of websites central to a topic within their respective collecting scope. Going beyond web content produced by creator(s) of an archival *fonds*, Curators build databases of web archives which are intrinsically valuable and document the cultural heritage of their collecting areas.

A curatorial approach to packaging individual websites into a themed collection enhances scholarship and access for researchers as it bundles websites within the context of related, primary resources. Simultaneously, it aids in the preservation of otherwise ephemeral web content. Examples include, but are not limited to, at risk web content within the subject areas of the four Distinctive Collections units.
Selection

Northwestern University Libraries may need to selectively archive web content based on potential resource constraints, such as cost associated with storage space and the maximum space available per annum. Additional evaluations during the Acquisition and Re-Appraisal phases of web archiving will be conducted to avoid these resource constraints.

Selection procedures are followed in accordance with a collection’s scope:

1. Archival collection

Upon donation to one of the Distinctive Collections units, Curators evaluate the archival collection and work with the donor, as necessary, to research and determine associated web content for the collection.

For existing collections, whether processed or not, Curators, processors, and/or subject specialists research live web content related to archival collections which need to be archived.

Selection Criteria:
- Materials reflect the collecting scope of the respective Distinctive Collections unit
- Materials whose provenance originates from a rare/archival collection
- Materials whose creator(s) are the creator(s) of a rare/archival collection, in part or whole
- Materials whose subject matter directly complements the provenance and/or creator(s) of a rare/archival collection

2. Curated cultural heritage

Periodically, subject specialists of the Distinctive Collections units evaluate and curate live web content into new and existing collections for web archiving. Additionally, they may pool resources with external partners to form collaborative web archives around similar subject areas.

Selection criteria:
- Materials reflect the collecting scope of the respective Distinctive Collections unit
- Materials are deemed at risk and/or relevant to the preservation of cultural heritage
- Materials can be logically grouped by subject, event, location, and/or creator(s) to strengthen the context of the archived content
- Upon initial curation, materials are not already being captured as part of an existing Distinctive Collections unit’s archival collection(s)
- Upon initial curation, materials are not being actively archived by another institution with similar access capabilities
- Materials may support existing primary resources, both traditional and born-digital, collected by a Distinctive Collections unit
Acquisition

Northwestern University Libraries subscribes to the Archive-It web archiving service, which automates the crawling efforts, storage, and preservation of archived web content for the Distinctive Collections unit.

Distinctive Collections staff involved in web archiving tasks determine capture levels and crawl frequency based on collection scope.

Test crawls are conducted upon initial acquisition of web archives to determine and potentially correct crawling imperfections, avoid an over-abundance of captures beyond scope of web content, determine size of captured data, and help identify crawl frequency and capture levels.

Northwestern University Libraries makes an effort to produce exact captures of targeted websites, but acknowledges the imperfections of web crawling technologies. In-scope embedded web content, such as images, audiovisual materials, documents, etc., will be included in crawls as much as possible. Although captured, some features of websites may not be rendered within the searchable database. As web crawling technologies are updated to accommodate various web dynamics and features, acquisition practices may evolve to accommodate these changes.

The frequency of web captures for live web content will vary with each case and will be identified by the subject specialist(s) within a Distinctive Collections unit in tandem with the Digital Archivist. Inactive, static websites will be crawled upon initial acquisition/identification by a Distinctive Collections unit. Web sites containing comparatively static content or otherwise not changing very often will be crawled and evaluated once a year until noticeable changes are observed. Highly active websites which undergo regular to frequent changes may be crawled periodically throughout the year.
Arrangement

Due to legacy collecting practices of web archives and limitations of the Archive-It system, intellectual arrangement and virtual arrangement may not necessarily be in alignment. Descriptive practices within Archive-It and other discovery tools are intended to help clarify the arrangement to users. As web preservation and storage capabilities improve, virtual re-arrangement will occur, as needed.

At its top level, a web archives are grouped into a “collection” in Archive-It, which includes web documents curated around a common theme, topic, or domain. The curation of an Archive-It “collection” is created or amended in accordance with the Scope categories outlined in this document.

In scenarios in which seeds (i.e., URLs) have already been captured, saved, and arranged in Archive-It prior to 2018, web archives will continue to be intellectually arranged (i.e., described) as a collection; due to Archive-It system limitations, content may not be able to be arranged as an Archive-It “collection” in the discovery system. As a result, seeds will also be categorized within Archive-It as “groups.” Seed arrangement as Archive-It “groups” are followed in accordance with a collection’s scope:

1. Archival collection

Seeds from an archival collection’s associated web archives will be categorized together based on the *fonds* or record group of the archival collection. The Archive-It “group” will be named after the archival collection’s title.

2. Curated cultural heritage

Seeds saved within Archive-It which are not part of the *fonds* of an archival collection will be categorized within a new/existing cultural heritage collection of its respective Distinctive Collections’ unit. The Archive-It “group” will be named by the Curator and/or Distinctive Collections’ subject specialists.
Description

Descriptive metadata will be added to collections and/or seeds of web archives to enhance context and retrieval of content.

Descriptions will be encoded within Archive-It at collection and item level using Dublin Core’s standard fifteen fields. Descriptive practices will follow the OCLC’s recommendations for Descriptive Metadata for Web Archiving.

The level of description will vary with each case and may be dependent upon the amount of information available upon acquisition, staff time and resources, and/or amount of content saved by the web crawler.

For collections within Archive-It, the following fields are required at minimum:

- Title
- Creator
- Description
- Date (if known)
- Identifier

For items within Archive-It, such as seeds (URLs) or embedded documents, Title is the only minimum field requirement. The need for added descriptions beyond minimum requirement will be identified by the subject specialist(s) within a Distinctive Collections unit in tandem with the Digital Archivist.

Further descriptive practices are followed in accordance with a collection’s scope:

1. Archival collection

   Descriptions of published web archives for an archival collection will be included in the collection’s published finding aid as part of the notes section and/or the collection inventory. If possible, embedded links to the Archive-It database for the collection will also be included in the finding aid.

   Web content which is embargoed will not be described in finding aids until its embargo period has lapsed. However, information about dark web crawls associated with the collection will be documented within the processing notes.

2. Curated cultural heritage

   When possible, curated web archives collections will be linked to related resources within its respective Distinctive Collections unit. If a published finding aid exists for a related collection and the web archives is not dark, a Related Resource note will be included in the finding aid(s).
**Maintenance and Re-Appraisal**

Web archives cannot be deleted in Archive-It once their crawls have been saved. Consistency during the identification and testing crawl periods are necessary to maintain responsible custody of the web archives. Yearly, web archives for both archival collections and curated cultural heritage are assessed overall to determine whether the captures and discoverability in the database can still be identified within the collecting scope of NUL web archives. If not, seeds and/or collections will no longer be made visible to the public. Similarly, active web crawls are assessed yearly to determine whether the current frequency of scheduled web crawls is necessary and/or sustainable. If not, crawl frequencies may be modified, lessened, or halted.

If archived versions of a website exhibit severe technical issues to the extent that patrons will not be able to use them for research, the captures may be unpublished and seeds’ crawl frequencies and/or capture level(s) may be altered, lessened, or halted. Curators may continue to monitor such seeds of live web content if they are associated with an archival collection or have been identified within the curatorial cultural heritage scope.

Further maintenance and re-appraisal procedures are followed in accordance with a collection’s scope:

1. Archival collection

   For new acquisitions of archival collections, web content to be crawled will be noted in the accession record. For existing, processed collections, web content to be crawled will be forwarded to the Digital Archivist, as soon as it is identified. Levels of capture, identification of descriptive metadata, crawl frequency, and access restrictions are determined. As crawls are tested and saved, the collection’s web archives remain unpublished until processing for the entire archival collection is complete and a finding aid is published. Redaction and/or restriction embargoes are determined during processing.

2. Curated cultural heritage

   Yearly, Curators research the live web, analyze existing web archives collections, and propose the creation of new curated web archives collections, as needed. Curators research external web archiving initiatives to verify that no other parties are saving identified seeds and making them available. Curators identify the level(s) of capture, descriptive metadata, crawl frequency level(s), and access restrictions. Test crawls and quality assurance are conducted.

   Quarterly, Curators identify additional websites (i.e., seeds) to add to existing collections, as needed. Upon identification, Curators determine levels of capture, identification of descriptive metadata, crawl frequencies, and access restrictions for new seeds.
Presentation and Access

Wherever possible, all feasible efforts will be made to render crawled web archives as closely as possible to its original web content. In instances when test crawls cannot be modified in a way to fully emulate the original web experience, capture deficiencies will be documented in the collection’s processing notes.

The designation within Archive-It to make a seed or collection “Visible to the public” or not will be evaluated and selected for every seed and collection.

Publicly accessible web archives will remain available in the Archive-It database, regardless of crawl frequency designation.

Researchers’ access points to the web archives include NUL’s Archive-It instance, links to collections/groups from finding aids, and links from the NUL website.

Newly added seeds to a collection are not available for access immediately to account for completion of crawl testing, quality assurance, and scoping tasks. Once test crawls are completed and seeds are scheduled for crawl(s) in Archive-It, their respective web archives will become available approximately 24 hours after the crawl has been completed.

Embargoed, or dark, web archives will not be made discoverable on the public-side of the Archive-It database until its embargo period has lapsed.
Appendix A: Glossary

The following terms' definitions are clarified here to account for web archiving jargon and provide context for proprietary terms used by the Archive-It system.

**Capture level**
The amount of web content to be crawled of a seed; Archive-It has the ability to capture the front page of a website without saving any links to other pages or it can be crawled thoroughly to include external links.

**Collection**
Archive-It uses this term to signify a group of archived web documents curated around a common theme, topic, or domain.

**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 seeds and scope. *Crawl* can also reference the archived content associated with the action.

**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).

**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.

**Descriptive metadata**
Information that refers to the intellectual content of materials and aids discovery of such materials; Archie-It uses the Dublin Core standard.

**Group**
A proprietary term used by Archive-It to document a function which intellectually groups seeds from the same collection together.

**Seed**
The starting point URL for a crawler and access point to archived collections.
Appendix B: References

The following policies and professional literature were referenced during preparation of Northwestern University Libraries’ Web Archives Collection Development Policy:


