Preservation Action Plan: Web Records
National Archives and Records Administration (NARA)

Plan Date: 202503
Template: 202105

Web Records

Web records are a collection of information, documents, or database records that are sent from a server to a browser via Hypertext Transfer Protocol (HTTP) when a Uniform Resource Locator (URL) has been activated.

Though web records may include components that seemingly belong in other content categories (documents, databases, images, etc.), they are treated as a whole instead of separately. See the Preservation Action Plan for Email for similar treatment.

Where possible, access should be provided in the native formats for the site as captured, which can include HTML and all associated media sites and files required for look and feel. Web records may be provided in a container format, e.g., WARC or WACZ, for replay.

This plan doesn't currently explicitly include social media records.

Significant Properties of Web Records

The presentation of web content is often changeable or mutable, based on variables such as a user's browser preferences/permissions, browser characteristics (support for plug-ins, desktop vs. mobile presentation), or network limitations. The "replay" presentation of archived websites may also vary due to "leaks" from the live web, cookies, and lack of modern browser support for older web technologies such as Flash. When replaying archived websites, completeness may vary due to scoping parameters such as the depth of the crawl, inclusion or exclusion of multimedia, external links, etc.

The Structure and Appearance/Layout are the most important properties. The manner in which elements are organized, interrelated, and displayed can be found in one or more of the following: source code, record layout, table and frame structure, linkage, site map, and hypertext.

Web content is either static or dynamic. The static portion of web content does not maintain additional behavior properties beyond hypertext/internal links. Dynamic "deep" web content is usually managed through the use of databases and style sheets which are component parts of the Web. If static, capturing the source code of web content generally will encompass content and aspects of the appearance, and structure properties. If dynamic, you may encounter

another record category such as a database. Not all records stored on the web are best preserved as a web record. Defining properties for these component parts should be addressed within a separate series for that component and its record type.

Appearance

Name	Definition	Function Description
Layout/Inline	Inline or embedded layout and look and feel of page content.	Part of the page/source code. Includes but is not limited to: ● Style ● Format Elements ● Class ● Heading ● List ● Table ● Form ● Canvas ● SVG (Scalable Vector Graphics) ● Client-side Script instructions
Layout/ External	External file(s) that identifies layout and design elements, or are required for layout.	External to the page, must be captured to accurately capture full content and layout. Includes but is not limited to: ● Linked objects/files to be embedded in the layout, such as images, video, audio, dynamic or static output from external applications/APIs/web services ● Server Scripts ● Cascading Style Sheets

Structure

Name	Definition	Function Description
Links to External Sites	Significant links should be described and documented in external documentation. Linked content in a different domain should be redirected and considered a significant property	All links are part of the page/source code. Externally referenced content (e.g., accessed via hyperlink) that resides in a different domain and is not managed for an agency under a formal

	only if associated with the agency or externally hosted for the agency through a formal agreement.	agreement will not be accepted for transfer.
Site Organization	Logical organization of web content, including navigation, embedded objects or actions.	Part of the page/source code and documented through an external Site Map file.
Schema for Dynamic Content	If Database-driven, the schema for the database is necessary for understanding the structure. See the significant properties for Database Records.	Databases and datasets may not be properly captured through web crawling and instead should be transferred and preserved using other methods.

Behavior

Name	Definition	Function Description
Replay	Replay, or playback, refers to the presentation of archived web records as they existed on the live web at the time they were crawled, or otherwise captured.	Replay is done through a web browser using specialized web archives replay software.

Context

Name	Definition	Function Description
Descriptive Metadata	Information contained within the record (intrinsic) that refers to the intellectual content of material and aids discovery of such materials.	Part of the page/source code. Includes but is not limited to: Title Meta/Author Meta/Description Meta/Keywords Caption/Subject/Date/Event/Transaction can all add value to the record.
Crawl/Capture Metadata	Metadata about a web capture (date, mechanism, scope, etc.) needs to be preserved with the website.	

Current NARA Transfer Guidance for Web Records

Bulletin 2014-04

- Preferred:
    - Web ARChive Format (WARC) 1.0
    - Web ARChive Format (WARC) 1.1
    - Web Archive Collection Zipped (WACZ)

- Acceptable:
    - Archive File Format (ARC)

Current NARA Format(s) for Public Access and Reference for Web Records
Formats for Public Access are those made available online through the National Archives Catalog and on the Congressional Web Harvest website. Formats for Reference are defined as those made available to researchers upon direct requests for digital copies.

Format(s) Available for Public Access: Content created or delivered for public access in the Catalog is delivered primarily in the following file formats: PDF (Textual and Image), JPEG (Textual and Image), MP3 (Audio), and MP4 (Audio/Video) and ASCII (Datasets). Other file formats may be present depending on when they were added to the Catalog.

The 2004 Federal Term Harvest and the Congressional Web Harvest (from 2006 to present), are made publicly accessible through https://www.webharvest.gov/.

Format(s) Available for Reference: When available, records may be delivered to researchers in the formats in which they are preserved.