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

DRAFT 20190801
Template: 201907

Electronic Record or Digital Surrogate Types and Associated Formats

A collection of information, documents, or database that is sent from a server to a browser via Hypertext Transfer Protocol (HTTP) when a URL has been activated and meets the definition of Federal record and is provided via an agency's web site.

Component parts of web content may include other record types where it may be necessary to address the essential characteristics of those record and data types within a separate series of records. However, similar to the record type "email," when encountering another record type within web records, they are not treated as a separate record type.

Essential Characteristics of this Record Type

The rapid evolution of the web, the complexity of the web, and the variety of ways agencies make use of it, limits defining characteristics only in the manner in which NARA can currently accept the transfer of web content.

The presentation of web content, although it may be intended to be fixed is most often changed or mutable, as is the content over time. The website's appearance is variable, as it is altered for the user depending on user preferences, browser characteristics, network limitations, frequency, and topic of use.

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

Web content is either static or dynamic. The static portion of web content does not maintain additional behavior characteristics 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 characteristics. If dynamic, you may encounter another record type, such as a database. Defining characteristics 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. Includes but is not limited to: ● Style ● Format Elements ● Class ● Heading ● List ● Table ● Form ● Canvas ● SVG (Scalable Vector Graphics) ● Client-side Script instructions	Part of the page/source code
Layout/External	External file(s) that identifies layout and design elements, or are required for 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	External to the page, must be captured to accurately capture full content and layout

Structure

Name	Definition	Function Description
Links to External Sites	Significant links should described and documented in external documentation. Linked content in a different	All links are part of the page/source code. Externally referenced content (e.g., accessed via hyperlink) that resides in

	domain should be redirected and considered an essential characteristic only if associated with the agency or externally hosted for the agency though a formal agreement.	a different domain and is not managed for an agency under a formal 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 required. See the essential characteristics for Database Records.	The schema should be present because if there are serious shortcomings in descriptive and technical documentation then according to NARA guidance the web records should not be considered for transfer.

Behavior

Name	Definition	Function Description
Intended Function	The purpose, audience, and functionality of the site	Accompanying documentation that describes the intended audience and functionality.

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. 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.	Part of the page/source code
Crawl/Capture Metadata	Metadata about a web capture (date. mechanism, etc.) needs to be preserved along with the website.	

Current NARA Transfer Guidance for this Record Type

Preferred: None (not identified)

Acceptable:

- Web ARChive Format (WARC)
- Archive File Format (ARC)

Current NARA Public Access/Reference Format(s) for this Record Type

This Plan references existing public access file formats for electronic records at NARA, determined with a survey of the available public access formats in the National Archives Catalog. These references do not represent recommended public access formats under NARA policies. They are intended for informational purposes only.

Reference Format: Native/original format(s).

Public Access Format: Download from the National Archives Catalog if available, or request access through the reference staff in the appropriate custodial unit. For Legislative web content, public access is provided through https://www.webharvest.gov/

Comments and Notes

Access should be provided in the native formats for the site as captured, including HTML and all associated media sites and files required for look and feel. This can be provided in a container format, e.g., WARC or ARC.

Hypertext Markup Language 1.0; 3.0; 3.2; 4.01; 5.0, 5.1

NARA Format ID: NF00208

Extension(s):
- htm
- html

Documentation

- Generic format, but associated with several web page editing applications including Dreamweaver, Microsoft FrontPage, and Homesite, but can be opened in any text editor.
- https://html.spec.whatwg.org/

Risk and Prioritization Analysis

✔**Low Risk**
☐ **Moderate Risk**
☐ **High Risk**
42 Numeric Risk Rating
41 Numeric Prioritization Rating

Proposed Preservation Plan

✔ **Retain** file format in its existing format.
☐ **Transform** file to a new format.
Selected Format:
☐ **Procure/develop tools** to preserve, manage and provide access to records of this type in their existing form.
☐ **Procure/develop tools** to transform the format to the preferred normalized form.
☐ **Provide Additional Information** so that the record type remains understandable/usable over time.
☐ **Explore Additional Options**

Justification: HTML is a plain text format, easily machine and human readable, and a stable and well-documented open format. In addition, HTML and web archives in general should never be migrated, as that would change the fundamental linkages and functionality of web objects.

Preferred Processing and Transformation Tool(s)

- Any supported Text Editor
- Any supported Web Browser

Preferred Viewer/Access Software

- Any supported Web Browser

MIME Hypertext Markup Language, AKA Microsoft Hypertext Markup Language

NARA Format ID: NF00334

Extension(s):

- mht
- mhtml

Documentation

https://tools.ietf.org/html/rfc2557

Risk and Prioritization Analysis

✓ **Low Risk**
☐ **Moderate Risk**
☐ **High Risk**
38 Numeric Risk Rating
38 Numeric Prioritization Rating

Proposed Preservation Plan

✓ **Retain** file format in its existing format.
☐ **Transform** file to a new format.
Selected Format:
☐ **Procure/develop tools** to preserve, manage and provide access to records of this type in their existing form.
☐ **Procure/develop tools** to transform the format to the preferred normalized form.
☐ **Provide Additional Information** so that the record type remains understandable/usable over time.
☐ **Explore Additional Options**

Justification: MHTML is a plain text format, easily machine and human readable, and a stable and well-documented open format. In addition, HTML formats and web archives in general should never be migrated, as that would change the fundamental linkages and functionality of web objects.

Preferred Processing and Transformation Tool(s)

- Any supported Text Editor
- Any supported Web Browser

Preferred Viewer/Access Software

- Any supported Web Browser

eXtensible Hypertext Markup Language 1.0

NARA Format ID: NF00185

Extension(s):

- xhtm
- xhtml

Documentation

- https://www.w3.org/TR/xhtml1/

Risk and Prioritization Analysis

✓ **Low Risk**
☐ **Moderate Risk**
☐ **High Risk**
33 Numeric Risk Rating
33 Numeric Prioritization Rating

Proposed Preservation Plan

✓ **Retain** file format in its existing format.
☐ **Transform** file to a new format.
Selected Format:
☐ **Procure/develop tools** to preserve, manage and provide access to records of this type in their existing form.
☐ **Procure/develop tools** to transform the format to the preferred normalized form.
☐ **Provide Additional Information** so that the record type remains understandable/usable over time.
☐ **Explore Additional Options**

Justification: xHTML is a plain text format, easily machine and human readable, and a stable and well-documented open format. In addition, xHTML and web archives in general should never be migrated, as that would change the fundamental linkages and functionality of web objects.

Preferred Processing and Transformation Tool(s)

- Any supported Text Editor
- Any supported Web Browser

Preferred Viewer/Access Software

- Any supported Web Browser

eXtensible Hypertext Markup Language 1.1

NARA Format ID: NF00186

Extension(s):

- xhtm
- xhtml

Documentation

- https://www.w3.org/TR/xhtml11/

Risk and Prioritization Analysis

✓ **Low Risk**
☐ **Moderate Risk**
☐ **High Risk**
35 Numeric Risk Rating
35 Numeric Prioritization Rating

Proposed Preservation Plan

✓ **Retain** file format in its existing format.
☐ **Transform** file to a new format.
Selected Format:
☐ **Procure/develop tools** to preserve, manage and provide access to records of this type in their existing form.
☐ **Procure/develop tools** to transform the format to the preferred normalized form.
☐ **Provide Additional Information** so that the record type remains understandable/usable over time.
☐ **Explore Additional Options**

Justification: xHTML is a plain text format, easily machine and human readable, and a stable and well-documented open format. In addition, xHTML and web archives in general should never be migrated, as that would change the fundamental linkages and functionality of web objects.

Preferred Processing and Transformation Tool(s)

- Any supported Text Editor
- Any supported Web Browser

Preferred Viewer/Access Software

Any supported Web Browser

Cascading Style Sheets

NARA Format ID: NF00141

Extension(s):

- css

Documentation

- https://www.w3.org/Style/CSS/specs.en.html

Risk and Prioritization Analysis

✓ **Low Risk**
☐ **Moderate Risk**
☐ **High Risk**
31 Numeric Risk Rating
31 Numeric Prioritization Rating

Proposed Preservation Plan

✓ **Retain** file format in its existing format.
☐ **Transform** file to a new format.
Selected Format:
☐ **Procure/develop tools** to preserve, manage and provide access to records of this type in their existing form.
☐ **Procure/develop tools** to transform the format to the preferred normalized form.
☐ **Provide Additional Information** so that the record type remains understandable/usable over time.
☐ **Explore Additional Options**

Justification: CSS is used to control and look and feel/rendering of HTML or xHTML. It is a plain text format, easily machine and human readable, and a stable and well-documented open format. In addition, CSS and web archives in general should never be migrated, as that would change the fundamental linkages and functionality of web objects.

Preferred Processing and Transformation Tool(s)

- Macromedia Director
- Any supported Text Editor
- Any supported Web Browser

Preferred Viewer/Access Software

- Any supported Web Browser

eXtensible Style Language

NARA Format ID: NF00190

Extension(s):

- xsl
- xslt

Documentation

- https://www.w3.org/Style/XSL/

Risk and Prioritization Analysis

✓ **Low Risk**
☐ **Moderate Risk**
☐ **High Risk**
31 Numeric Risk Rating
26 Numeric Prioritization Rating

Proposed Preservation Plan

✓ **Retain** file format in its existing format.
☐ **Transform** file to a new format.
Selected Format:
☐ **Procure/develop tools** to preserve, manage and provide access to records of this type in their existing form.
☐ **Procure/develop tools** to transform the format to the preferred normalized form.
☐ **Provide Additional Information** so that the record type remains understandable/usable over time.
☐ **Explore Additional Options**

Justification: XSL (also called XSLT) is used to control and look and feel/rendering of HTML or xHTML. It is a plain text format, easily machine and human readable, and a stable and well-documented open format. In addition, XSL and web archives in general should never be migrated, as that would change the fundamental linkages and functionality of web objects.

Preferred Processing and Transformation Tool(s)

- oXygen
- Any supported Text Editor
- Any supported Web Browser

Preferred Viewer/Access Software

- Any supported Web Browser

Document Type Definition

NARA Format ID: NF00162

Extension(s):

- dtd

Documentation

- https://www.w3.org/XML/1998/06/xmlspec-report-19980910.htm

Risk and Prioritization Analysis

✓ **Low Risk**
☐ **Moderate Risk**
☐ **High Risk**
44 Numeric Risk Rating
44 Numeric Prioritization Rating

Proposed Preservation Plan

✓ **Retain** file format in its existing format.
☐ **Transform** file to a new format.
Selected Format:
☐ **Procure/develop tools** to preserve, manage and provide access to records of this type in their existing form.
☐ **Procure/develop tools** to transform the format to the preferred normalized form.
☐ **Provide Additional Information** so that the record type remains understandable/usable over time.
☐ **Explore Additional Options**

Justification: A DTD is a specialized XML file that provides structure and instructions for the display/rendering of content marked up as XML on the web. It is a plain text format, easily machine and human readable, and a stable and well-documented open format. In addition, DTDs and web archives in general should never be migrated, as that would change the fundamental linkages and functionality of web objects.

Preferred Processing and Transformation Tool(s)

- oXygen
- Any supported Text Editor
- Any supported Web Browser

Preferred Viewer/Access Software

- Any supported Web Browser

eXtensinble Markup Language Schema

NARA Format ID: NF00188

Extension(s):

- xsd

Documentation

- https://www.w3.org/standards/xml/schema

Risk and Prioritization Analysis

✓ **Low Risk**
☐ **Moderate Risk**
☐ **High Risk**
44 Numeric Risk Rating
44 Numeric Prioritization Rating

Proposed Preservation Plan

✓ **Retain** file format in its existing format.
☐ **Transform** file to a new format.
Selected Format:
☐ **Procure/develop tools** to preserve, manage and provide access to records of this type in their existing form.
☐ **Procure/develop tools** to transform the format to the preferred normalized form.
☐ **Provide Additional Information** so that the record type remains understandable/usable over time.
☐ **Explore Additional Options**

Justification: A Schema is a specialized XML file that provides structure and instructions for the display/rendering of content marked up as XML on the web. It is a plain text format, easily machine and human readable, and a stable and well-documented open format. In addition, Schema and web archives in general should never be migrated, as that would change the fundamental linkages and functionality of web objects.

Preferred Processing and Transformation Tool(s)

- oXygen
- Any supported Text Editor
- Any supported Web Browser

Preferred Viewer/Access Software

- Any supported Web Browser

Uniform Resource Locator

NARA Format ID: NF00430

Extension(s):

- url

Documentation

- https://url.spec.whatwg.org/

Risk and Prioritization Analysis

✓ **Low Risk**
☐ **Moderate Risk**
☐ **High Risk**
40 Numeric Risk Rating
40 Numeric Prioritization Rating

Proposed Preservation Plan

✓ **Retain** file format in its existing format.
☐ **Transform** file to a new format.
Selected Format:
☐ **Procure/develop tools** to preserve, manage and provide access to records of this type in their existing form.
☐ **Procure/develop tools** to transform the format to the preferred normalized form.
☐ **Provide Additional Information** so that the record type remains understandable/usable over time.
☐ **Explore Additional Options**

Justification: A URL is a plain text format, easily machine and human readable, and a stable and well-documented open format. In addition, URLs and web archives in general should never be migrated, as that would change the fundamental linkages and functionality of web objects.

Preferred Processing and Transformation Tool(s)

- Any supported Text Editor
- Any supported Web Browser

Preferred Viewer/Access Software

- Any supported Web Browser

Archive File Format

NARA Format ID: NF00112

Extension(s):

- arc

Documentation

- ARC Container/Wrapper: https://archive.org/web/researcher/ArcFileFormat.php
- DAT within an ARC: https://archive.org/web/researcher/dat_file_format.php
- CDX within an ARC: https://iipc.github.io/warc-specifications/specifications/cdx-format/cdx-2015/

Risk and Prioritization Analysis

✓ **Low Risk**
☐ **Moderate Risk**
☐ **High Risk**
30 Numeric Risk Rating
30 Numeric Prioritization Rating

Proposed Preservation Plan

✓ **Retain** file format in its existing format.
☐ **Transform** file to a new format.
Selected Format:
☐ **Procure/develop tools** to preserve, manage and provide access to records of this type in their existing form.
☐ **Procure/develop tools** to transform the format to the preferred normalized form.
☐ **Provide Additional Information** so that the record type remains understandable/usable over time.
☐ **Explore Additional Options**

Justification: An ARC is a container format created when harvesting and creating web archives. It is a container wrapped around plain text files metadata (DAT) and indexing (CDX) files, as well as the component files that comprise the archived web sites. The container wrapper and its DAT and CDX files are machine and human readable, and it is a stable and well-documented open format. In addition, ARCs and web archives in general should never be migrated, as that would change the fundamental linkages and functionality of web objects.

Preferred Processing and Transformation Tool(s)

- Wayback software

- Any supported Text Editor
- Any supported Web Browser

Preferred Viewer/Access Software

- Wayback software
- Any supported Web Browser

Web Archive File Format (WARC)

NARA Format ID: NF00439

Extension(s):

- warc

Documentation

- WARC Container/Wrapper: https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.0/
- CDX within a WARC: https://iipc.github.io/warc-specifications/specifications/cdx-format/cdx-2015/

Risk and Prioritization Analysis

✓ **Low Risk**
☐ **Moderate Risk**
☐ **High Risk**
36 Numeric Risk Rating
36 Numeric Prioritization Rating

Proposed Preservation Plan

✓ **Retain** file format in its existing format.
☐ **Transform** file to a new format.
Selected Format:
☐ **Procure/develop tools** to preserve, manage and provide access to records of this type in their existing form.
☐ **Procure/develop tools** to transform the format to the preferred normalized form.
☐ **Provide Additional Information** so that the record type remains understandable/usable over time.
☐ **Explore Additional Options**

Justification: A WARC is a container format created when harvesting and creating web archives. It is a container wrapped around a plain text indexing (CDX) file, as well as the component files that comprise the archived web sites. The container wrapper and its CDX file is machine and human readable, and it is a stable and well-documented open format. In addition, WARCs and web archives in general should never be migrated, as that would change the fundamental linkages and functionality of web objects.

Preferred Processing and Transformation Tool(s)

- Wayback software

- Any supported Text Editor
- Any supported Web Browser

Preferred Viewer/Access Software

- Wayback software
- Any supported Web Browser

eXtensible Markup Language 1.0, 1.1

NARA Format ID: NF00187

Extension(s):

- xml

Documentation

- https://www.w3.org/TR/xml/
- https://www.w3.org/TR/2006/REC-xml11-20060816/

Risk and Prioritization Analysis

✓ **Low Risk**
☐ **Moderate Risk**
☐ **High Risk**
45 Numeric Risk Rating
45 Numeric Prioritization Rating

Proposed Preservation Plan

✓ **Retain** file format in its existing format.
☐ **Transform** file to a new format.
Selected Format:
☐ **Procure/develop tools** to preserve, manage and provide access to records of this type in their existing form.
☐ **Procure/develop tools** to transform the format to the preferred normalized form.
☐ **Provide Additional Information** so that the record type remains understandable/usable over time.
☐ **Explore Additional Options**

Justification: XML is a plain text format, easily machine and human readable, and a stable and well-documented open format. It is a preferred format under NARA Transfer guidance.

Preferred Processing and Transformation Tool(s)

- Any supported Text Editor
- Any supported Web Browser

Preferred Viewer/Access Software

- Any supported Web Browser or XML parsing/display tool.