Preservation Action Plan: Structured Data/Plain Text
National Archives and Records Administration (NARA)

DRAFT 20190801
Template: 201907

Electronic Record or Digital Surrogate Types and Associated Formats

Plain-text delimited or marked-up structured data files.

Essential Characteristics of this Record Type

Appearance

Name	Definition	Function Description
Character Encoding	The data used by computers can be: ● ASCII ● Unicode ● EBCDIC ● Plain Text	The sequence of characters (letters, numbers, punctuation, and certain symbols) or coding that translate human readable or natural language characters to a specialized format for efficient transmission or storage. Assumption: Always has to exist and needs to be identified in order to open in a compatible format or to transform to another format, such as ASCII. Must meet Ingest requirements.

Structure

Name	Definition	Function Description
Schema	Record layout is typically embedded, but like databases, code lists and data dictionaries may be necessary to understand data.	

Linkage	Connection between or within records or files. (See also Hyperlinks)	If connections exist, then they are core.
Column Count	Total number of columns with content in the document	Valuable for evaluating the completeness of the content after transformations.
Row Count	Total number of rows in the document	Valuable for evaluating the completeness of the content after transformations.
Technical Metadata	Metadata describing the specific database format, software, software version, etc. This is generally automatically embedded in the file header.	Supports the ability to potentially recreate interactions with the data, such as queries or graphing, can be recreated.

Behavior

Name		
Hyperlinks	Links within the file, to external files, or to external data sources.	Hyperlinks are generally core features. The biggest risk is links to external files that may not be part of the series or to external websites that may not remain active.

Context

Name		
Related Files	A group of related or linked files that are referenced in the spreadsheet.	

Current NARA Transfer Guidance for this Record Type

Preferred:

- Comma Separated Value (CSV)
- ASCII Text
- XML
- JSON

- OpenDocument Format Spreadsheet

Acceptable:
- EBCDIC
- Microsoft Excel Office Open XML
- Microsoft Excel 97 Binary Document Format

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: In general, the records are delivered to researchers in the formats in which they are preserved.

Public Access Format: CSV, ASCII/plain text, and, where available, downloadable in the native format. Some datasets extracted from spreadsheets are made searchable at a row level via Access to Archival Databases (AAD) function of the National Archives Catalog.

Comments and Notes

In general, NARA accessions spreadsheets and databases in formats defined in 36 CFR 1235 and NARA Bulletin 2014-04. The formats defined in those issuances drive the formats we preserve records in and provide access to them.

Comma Delimited ASCII

NARA Format ID: NF00143

Extension(s):

- csv

Documentation

- 7-bit or 8-bit ASCII with structured columns and rows where the delimiters for the values are commas.
- http://fileformats.archiveteam.org/wiki/CSV
- http://fileformats.archiveteam.org/wiki/Tab_delimited

Risk and Prioritization Analysis

Supply the Risk Level and Numeric Rating and the Prioritization Numeric Rating as generated with the Format Risk and Prioritization Matrix for this file format.

✓ **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: The format is considered Preferred as per NARA transfer guidance.

Preferred Processing and Transformation Tool(s)

- The format can be opened and read in any supported text editor.

- The format can be opened in current versions of Excel or OpenOffice.
- Universal Office Converter - Commandline library that converts between any document format supported by LibreOffice/OpenOffice (https://github.com/dagwieers/unoconv)

Preferred Viewer/Access Software

- None - NARA does not currently provide viewer/access software. Some data is provided via Access to Archival Databases (AAD).

Tab Delimited ASCII

NARA Format ID: NF00418

Extension(s):
- tab

Documentation

- 7-bit or 8-bit ASCII with structured columns and rows where the delimiters for the values are tabs.
- http://fileformats.archiveteam.org/wiki/CSV
- http://fileformats.archiveteam.org/wiki/Tab_delimited

Risk and Prioritization Analysis

✓ **Low Risk**
☐ **Moderate Risk**
☐ **High Risk**
20 Numeric Risk Rating
15 Numeric Prioritization Rating

Proposed Preservation Plan

☐ **Retain** file format in its existing format.
✓ **Transform** file to a new format.
Selected Format: Comma Delimited Files are Preferred over Tab Delimited Files
☐ **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: The format is considered Preferred as per NARA transfer guidance. Tab-delimited files should be transformed to comma-delimited files.

Preferred Processing and Transformation Tool(s)

- The format can be opened and read in any supported text editor.
- The format can be opened in current versions of Excel or OpenOffice.

- Universal Office Converter - Commandline library that converts between any document format supported by LibreOffice/OpenOffice (https://github.com/dagwieers/unoconv)

Preferred Viewer/Access Software

- None - NARA does not currently provide viewer/access software. Some data is provided via Access to Archival Databases (AAD).

Extended Binary Coded Decimal Interchange Code - EBCDIC

NARA Format ID: NF00183

Extension(s):
- ebcidic

Documentation

- EBCDIC (Extended Binary Coded Decimal Interchange Code) is a family of character encodings used in a number of IBM mainframe systems in the pre-PC era. It is not strictly speaking a structured data format, but is used as the character encoding for structured data exported from IBM mainframe systems in use in the U.S. government.
- http://fileformats.archiveteam.org/wiki/EBCDIC
- https://en.wikipedia.org/wiki/EBCDIC

Risk and Prioritization Analysis

☐ **Low Risk**
✓ **Moderate Risk**
☐ **High Risk**
12 Numeric Risk Rating
12 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: The format is considered Acceptable as per NARA transfer guidance.

Preferred Processing and Transformation Tool(s)

- The format can be opened and read in any supported text editor.
- The format can be opened in current versions of Excel or OpenOffice.

- Universal Office Converter - Commandline library that converts between any document format supported by LibreOffice/OpenOffice (https://github.com/dagwieers/unoconv)

Preferred Viewer/Access Software

- None - NARA does not currently provide viewer/access software. Some data is provided via Access to Archival Databases (AAD).

Structured Data eXchange Format

NARA Format ID: NF00415

Extension(s):
- sdxf

Documentation

- A data serialization format that allows arbitrary structured data of different types to be assembled in one file for exchanging between arbitrary computers.
- https://www.ietf.org/rfc/rfc3072.txt
- http://www.pinpi.com/SDXF.htm
- https://en.wikipedia.org/wiki/SDXF

Risk and Prioritization Analysis

☐ **Low Risk**
✓ Moderate Risk
☐ **High Risk**
-2 Numeric Risk Rating
-12 Numeric Prioritization Rating

Proposed Preservation Plan

☐ **Retain** file format in its existing format.
✓ **Transform** file to a new format.
Selected Format: TBD, preferably CSV
☐ **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: The format is very uncommon though well-documented. It is not an ASCII or XML-based format so cannot be opened with text editors. Research is needed to identify an appropriate tool to transform the format into CSV.

Preferred Processing and Transformation Tool(s)

- Unknown at this time.

Preferred Viewer/Access Software

- None - NARA does not currently provide viewer/access software. Some data is provided via Access to Archival Databases (AAD).

eXtensible Markup Language

NARA Format ID: NF00187

Extension(s):
- xml

Documentation

- Marked-up plain text Unicode files which can separately or in combination be used to represent, parse, format, and display structured data.
- May be accompanied by: Document Type Definition (dtd), eXtensible Markup Language Schema (xsd), and eXtensible Style Language (xsl, xslt)
- 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.

Standard Generalized Markup Language

NARA Format ID: NF00410

Extension(s):

- sgm
- sgml

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: XML
☐ **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: SGML is no longer actively used, having been replaced by XML and JSON.

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.

JavaScript Object Notation

NARA Format ID: NF00218

Extension(s):

- json
- txt

Documentation

- https://tools.ietf.org/html/rfc8259

Risk and Prioritization Analysis

✓ **Low Risk**
☐ **Moderate Risk**
☐ **High Risk**
49 Numeric Risk Rating
49 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: JSON 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.

eXtensible Metadata Platform

NARA Format ID: NF00189

Extension(s):

- xmp

Documentation

- Marked-up plain text Unicode/XML files which can separately or in combination be used to represent, parse, format, and display structured data.
- 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**
37 Numeric Risk Rating
37 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)

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

Preferred Viewer/Access Software

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