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

Plan Date: 20200629
Template: 201907

Electronic Record or Digital Surrogate Types and Associated Formats

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

Essential Characteristics of Structured Data/Plain Text Records

Structured data refers to any data that resides in a fixed field within a record or file. This can include data contained in relational databases, spreadsheets, or marked up text. It requires a data model describing what categories of data will be stored in which fields, columns or tags; data types (numeric, currency, alphabetic, name, date, address); and controlled vocabulary.

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

Current NARA Transfer Guidance for Structured Data/Plain Text Records
Bulletin 2014-04

- 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 Format(s) for Public Access and Reference for Structured Data/ Plain Text Records

Formats for Public Access are those made available online through the National Archives Catalog. Formats for Reference are defined as those made available to researchers upon direct requests for digital copies.

Formats 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.

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

Comments and Notes

Some datasets extracted from spreadsheets are made searchable at a row level through the Access to Archival Databases (AAD) tool.