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

Plan Date: 20240322
Template: 202105

Structured Data

Delimited, fixed field, or marked-up structured data files.

NARA Transfer Guidance (Bulletin 2014-04) includes general requirements for structured data transferred to NARA by federal agencies

Significant Properties of Structured Data

Structured data refers to any data that resides in a 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	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 can be embedded. Code lists and data dictionaries will	Structured data should be transferred together with any associated files

	likely be necessary to understand data.	necessary to verify the validity of the data, e.g., DTDs, schemas, and data dictionaries. These associated files should be retained alongside the record.
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 may be automatically embedded in the file header depending on the file format.	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

Bulletin 2014-04

- Preferred:
  - Comma Separated Value (CSV)
  - ASCII Text
  - Extensible Markup Language (XML)
  - JavaScript Object Notation (JSON)
  - OpenDocument Format Spreadsheet (ODS)
  - Software Independent Archiving of Relational Databases (SIARD)

- Acceptable:
  - Microsoft Excel Office Open XML
  - Microsoft Excel 97 Binary Document Format (XLS)

- Acceptable for Imminent Transfer:
  - Extended Binary Coded Decimal Interchange Code (EBCDIC)

Current NARA Format(s) for Public Access and Reference for Structured Data

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: ASCII Text, CSV, EBCDIC, Microsoft Excel Spreadsheet, OpenDocument Format Spreadsheet, XML

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 unrestricted datasets are made searchable at a record level through the Access to Archival Databases (AAD) tool.