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

Plan Date: 20200629
Template: 201907

## Electronic Record or Digital Surrogate Types and Associated Formats

A spreadsheet is an electronic document in which data is arranged in grid-like rows and columns can be manipulated and be acted upon by formulae. Spreadsheet software may allow for multiple interacting sheets, AKA a workbook, and can display data as text, numerals, symbols, or in graphical form. So spreadsheet files may consist of not only the data, but also contain charts or visualizations based on the data and formulae. Each cell may contain either raw data or the results of formulas that automatically calculate and display a value based on the contents of other cells from the same or other pages/sheets, as well as external data sources.

## Essential Characteristics of Structured Data/Spreadsheets

Spreadsheets pose a challenge due to the various ways they can be used. They may contain only data and/or formulas, but they may be used to present information in a table format, combining text, numeric data, and possibly other visual cues such as color. They can also contain visual presentations created by the spreadsheet software itself such as charts, graphs and tables.

Due to the variability in use of spreadsheet applications, some of the questions to consider when preserving textual formats are relevant for spreadsheets as well:

- Would a change in the record's appearance alter its meaning?
- Does changing the record's appearance diminish its value? For example, if the records have been appraised as permanent for their informational value, and not evidential, then appearance characteristics may not need to be preserved.
- Would a change in the record's technical structure alter its appearance?
- Would a change in the record's technical structure affect its possible behaviors?
- Does deletion of a sheet/page always materially affect the content?

Built in tools such as macros with built in reports or external links are related to other record or data types which have their own essential characteristics.

## Appearance

| Name | Definition | Function Description |
|---|---|---|
| Fonts | Includes characteristics of type used in the document such as:<br>● Typeface (Arial, Times New Roman, etc.)<br>● Size (10 pt, 18 pt, etc.)<br>● Pitch<br>● Spacing<br>● Emphasis (bold, italic, strikethrough, underline, etc.) | There will always be a font in order to have text, however, font is only a core characteristic if it conveys meaning. |
| Color | Identification of the use of color in text and layout elements, e.g. borders, boxes.<br>● Hue: color family or name<br>● Saturation: purity or sharpness<br>● Brightness: shade or tint<br>● Contrast: range of optical density or tone | Color is essential if it bears meaning and/or value. |
| Formatting | Includes features of the document that determine how information is presented, such as:<br>● conditional formatting that provides visualization of data<br>● color coding<br>● formatting data in a table layout<br>● themes or templates | Not all formatting is essential to understanding the records. If the formatting conveys information, such as visualization of data, it may be core. |
| Annotations (comments) | Annotations included in the document. | |
| Graphics | Images embedded in the document, or graphics created using a spreadsheet graphics feature. | Not all graphics features are supported by all programs, but graphics often provide visual information that must be maintained as part of the record. |

| | | |
|---|---|---|
| | • WordArt<br>• SmartArt diagrams<br>• 3D shapes<br>• Pictures<br>• Shapes | |
| Pivot Tables | A pivot table is a table of statistics that summarize the data of a more extensive table. This summary might include sums, averages, or other statistics, which the pivot table groups together in a meaningful way. | Pivot tables can be used to draw attention to important information and therefore can be core. |
| Freeze panes | The ability to lock rows and/or columns into place so that when the page is scrolled, those rows/columns remain in place. | Indicates how the spreadsheet was viewed while in use, but not likely a core characteristic. |

**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 worksheets. (See also Hyperlinks) | If connections exist, then they are core. |
| Character Encoding | Encoding schema, e.g., US-ASCII, EBCDIC, UTF-8. | Required for the proper parsing and rendering of the record content. |
| 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. |

### Behavior

| Name | Definition | Function Description |
|---|---|---|
| Macros | A set of automated tasks that apply to content in the spreadsheet. | Generally not a core feature of a spreadsheet as macros are used to create content and appearance, but the instructions in and of themselves are ultimately not significant. |
| Formulas | An instruction in a cell or column that allows for automating calculations. | The final data in the cell itself is the core feature. Since information should not be changed once a spreadsheet is accessioned, maintaining formulas is likely not a core feature. |
| 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/Spreadsheets**
Bulletin 2014-04

- Preferred:
    - Comma Separated Value (CSV)
    - OpenDocument Format Spreadsheet (ODS)
    - ASCII Text

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

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

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 are delivered to researchers in the formats in which they are preserved.

**Comments and Notes**

Microsoft has documented some of the differences between ODS and XLSX (https://support.office.com/en-us/article/Differences-between-the-OpenDocument-Spreadsheet-ods-format-and-the-Excel-xlsx-format-4311c54f-ee86-4197-bd2d-5ecc35deb138) as related to opening and saving ODS files in various versions of Excel.

Formats that can be opened in OpenOffice are identified here: https://wiki.openoffice.org/wiki/Documentation/OOo3_User_Guides/Getting_Started/File_formats