http://doi.org/10.7207/twgn21-09

Preserving Spreadsheets

Data Types Series

Artefactual Systems and the Digital Preservation Coalition

DPC Technology Watch Guidance Note

July 2021

DigitalPreservationCoalition

ISSN: 2048-7916

First published in Great Britain in 2021 by the Digital Preservation Coalition.

The Data Type Guidance Note Series

Each Guidance Note in the Data Types series is designed to provide a primer on the current state of community knowledge about data types commonly encountered by those seeking to preserve digital holdings. Digital preservation is about keeping information findable, usable, and trustworthy over the long-term. The best approach for any repository will vary according to the scope and content of its holdings, available resources, and the expectations of its funders and users. There are however, broadly applicable good practices that have been established as a result of many years of research, practical implementation, and consensus building. These are presented here as a starting point, along with additional resources for further exploration.

This series of Data Type Guidance Notes has been authored by staff at Artefactual Systems in collaboration with the Digital Preservation Coalition. These notes have been developed in conjunction with the UK Nuclear Decommissioning Authority.

Digital preservation is an evolving field and continues to change and develop in response to external drivers and fresh challenges. New formats, standards, and examples of good practice will emerge over time and the information contained within this report will need to be updated. We welcome comments and feedback to: info@dpconline.org.

1 Overview

Spreadsheets store tabular data divided into columns and rows of data cells. These spreadsheet cells can possess technical attributes that influence the display of the data within them. There are approximately twenty-one identified groups of technical attributes (Wijsman, 2020), including:

- A format category such as number, currency, accounting, date, time, percentage, or fraction;
- Formulae;
- Comments;
- Pivot tables;
- Styling, including text alignment, font colour and style, cell colour, and border styles. Styling is often used to highlight specific aspects of the data such as totals or negative numbers, but also to convey meaning (Archaeology Data Service (ADS), 2009).

Spreadsheets existed before modern computing, but automated spreadsheets as a software application were invented around 1961 (Dutch National Archives, 2003). Over the years, spreadsheets have evolved to become more than just a single sheet of data delimited by tabs, commas, or other signifiers, and cell-based values. Spreadsheets can hold dynamic information, perform mathematical equations, create visualizations, and store embedded data. Spreadsheets may have multiple sheets or columns that are hidden from view. Because of this, spreadsheets can be more complex than they first appear, which can result in data loss or inability to access content if not handled with appropriate levels of care.

2 Preservation Challenges

Due to the sometimes hidden complexity of spreadsheets, there are some notable preservation challenges. These include handling formulae and macros, managing digital rights, collecting legacy file formats, and cloud-based software systems that do not produce files as we typically interpret them.

2.1 Dynamic content

Preserving the formulae used to create dynamic spreadsheets is a challenge. Dynamic content can include styling, formulae, graphs, charts, or macros. It was reported that 99% of spreadsheets ingested into the Ejournal Collection at the National Archives of the Netherlands were considered dynamic spreadsheets (van Veenendaal, 2019). Specific preservation challenges include the following:

- Formulae. Unlike the literal values shown in spreadsheet cells, the formulae that perform operations or calculations may not be compatible across different spreadsheet software.
- Macros. These are snippets of code that run repetitive tasks on a spreadsheet. They initiate sequences of keystrokes and mouse movements automatically (Technopedia, 2020). Macros are typically written in Visual Basic or Javascript and may not be compatible across different spreadsheet software.

2.2 Embedded content

Spreadsheets can have other data types embedded in them, such as images or other spreadsheets. Data embedded into spreadsheets make preservation more challenging.

- Embedded content may be linked from another location on the computer, a local access network, or the cloud. That connection can be severed when the spreadsheet or linked content is moved.
- Embedded content may be lost if the spreadsheet is migrated to a different format.
- If the creating software is not available in future, rendering of the embedded content in a new software application may not accurately represent the original structure. It may have a different appearance, be in a different location within the spreadsheet, or be lost entirely.

2.3 Cloud-based formats

Internet-only spreadsheets must be exported to a final state format such as Microsoft Excel or Open Document Format for long-term preservation. Current investigations into exporting spreadsheets has revealed some success, however there are issues with capturing Google-specific formulae and Sparklines (line charts embedded in cells) (Young, 2021). PDF and HTML exports are also less complete than Microsoft Excel or Open Document Format. The Google API also provides an opportunity to gather additional metadata not available through direct export from Google Sheets (Young, 2021).

2.4 Digital rights management

Digital rights management (DRM) is a set of technical measures designed to constrain the use of digital files, typically to protect intellectual property rights such as copyright (Dingledy and Matamoros, 2016). Encryption and password protection can also be used for other purposes such as restricting access to personal information. Specific restrictions might relate to opening, copying, saving, or printing files, which may hinder their preservation and re-use.

2.5 Legacy formats

There have been numerous software applications for creating spreadsheets over the years. Accessing and reading these spreadsheets can be complicated. Even if a modern application can render them, there may be a loss of significant properties. Formulae, macros, graphs, charts, and styling may be lost or altered in legacy spreadsheet formats, and embedded or linked content may not be accessible.

3 Typical spreadsheet formats

There is no single perfect format for the preservation and future use of spreadsheets. Decisions made on file formats should be dependent on the features and functionality to be preserved and the future use cases to be supported. Note that the table below does not provide an exhaustive list of formats suitable for preservation and access. The most suitable format for preserving the important features and functionality of a file may be the original format that it was created in. It is recommended that careful research and analysis is carried out before migrating files to a new format.

File format	**Extensions**	**Brief summary**
Delimiter separated values	.csv, .txt	Comma Separated Value (CSV) is a simple, open format commonly used to store and export spreadsheet data. It uses commas to format columns and rows. Tab Separated Values (TSV) uses tab characters instead of commas (LC, 2021).

		Tab-separated or comma-separated files can be created from plain text files to store structured data. Tabs or commas give the basic rows and columns structure in the text document. It should be noted however that these file formats cannot preserve styling, formulas, graphs, charts, relationships between multiple sheets, or other dynamic spreadsheet functionality (NCDCR, 2012). If none of these features are present, they are the preferred preservation and access formats (LC, 2020-2021; ADS, 2009).
Microsoft Excel	.xlsx	Open Office XML (XLSX) is the default file format of Microsoft Excel software, having replaced Microsoft Excel Binary File format (XLS) in 2007. Standardized as ISO/IEC 29500-1:2016, it is a container format that packages up a set of XML files to provide structure and formatting when rendered by software. Given that the format is XML-based, has an open (although incredibly complex) specification and international standard, and is widely used, XLSX is a preferred preservation format (LC, 2017b; ADS, 2009).
Open Document Format	.ods, .fods	OpenDocument Format, standardized as ISO/IEC 26300-1:2015, is an open XML format for office documents, including spreadsheets. The latest version, ODS version 1.2, has fewer interoperability issues than earlier versions (LC, 2020). It is considered a preferred preservation format (LC, 2020-2021; ADS, 2009).
Google Sheets	N/A	The structure and format of a Google Sheet is opaque to the user who is only able to view a rendered version of the (cloud stored) data, in their web browser. Google Sheets can be exported into a range of formats, including XSLX and ODS. This process may change the formatting and/or result in the loss of certain features. The UK National Archives is currently working on practices for exporting Google Sheets and relevant metadata (Young, 2021).
Portable Document Format	.pdf .pdf/A	PDF was developed by Adobe Systems in 1993 as a proprietary presentation format for documents. In 2008, it was released as the open standard *ISO 32000-1:2008*. PDF 2.0 was released in 2017 and updated in 2020 as *ISO 32000-2:2020*. PDF/A is the preservation format designed by Adobe Systems. For detailed information on PDF/A, please see the Guidance Note on Documents. PDF exports can retain the "look and feel" of a spreadsheet but not the functionality (NCDCR, 2012). Exporting spreadsheets to PDF will remove many of its essential dynamic properties. A PDF can represent different layers and formulae associated with a spreadsheet, but it cannot maintain relationships to other spreadsheets or external data sources (Dutch

		National Archives, 2003). Export to PDF is not advisable as a large-scale solution; however, in some circumstances it can be an acceptable preservation format, depending on the significant properties requiring preservation.

4 Tips for creators

Creators working in government, business or other controlled environments should be aware of their organizations' spreadsheet handling and records management policies. Adherence to guidelines and requirements for file formats, formatting and styling, versioning, and metadata creation will help ensure that the spreadsheets can be preserved in such a way as to retain their context and meaning over time. Academic settings often provide detailed guidelines for research data stored in spreadsheets, which should be followed as closely as possible. Organizations without format policies might find resources such as the Library of Congress' *Recommended Formats Statement* useful substitutes.

Good practice techniques suggest considering preservation before creating files. The suggestions for the creation and management of spreadsheets below will help to enable long-term preservation and accessibility:

- Be aware that just because a spreadsheet can be opened does not mean that all of its properties can be accessed.
- Cloud-based platforms (e.g. Google Sheets) create spreadsheets that are not files (Mitcham, 2017). Investigate the takeout or export options for this content and test that the spreadsheet's functionality is still present in the exported files.
- Spreadsheets should be 'self-describing' and able to exist independently. Create meaningful row and column headings, and describe the units used within the spreadsheet. Use controlled vocabularies and established word lists where possible for data entry to ensure consistency and clarity of the data (ADS, 2009).
- The Library of Congress (2020-2021) recommends copying macros or removing them entirely for long-term preservation. Dalgleish (2020) provides advice for storing formulae and macros externally. The advice includes how to add contextual comments within the macro code. The macro code can be stored externally as plain text files.
- The spreadsheet should be 'self-containing', meaning that externally linked content is captured and transferred with the spreadsheet. This will ensure that the relationship and context between linked content and spreadsheet is preserved (LC, 2020-2021).
- Spreadsheet authors should avoid using DRM when possible.
  Textual data in spreadsheets has greater longevity if the text encoding is set to ASCII or UTF-8 (LC, 2020-2021). Other text encodings may create problems with diacritics and other special characters when rendering the spreadsheets in different software and operating system environments.

5 Tips for archivists

5.1 General guidance

The following resources provide guidance on preserving and providing access to spreadsheets:

- The Archaeology Data Service's (2009) *Guides to Good Practice* on spreadsheets has general preservation and file format guidance.
- The Library of Congress' (2017a) *Sustainability of Digital Formats: Planning for Library of Congress Collections* has detailed overviews and information on various spreadsheet file formats.
- The Let's Solve the Format Problem (2019) wiki has a list of spreadsheet file formats that outlines accessible software for legacy formats, such as such as Lotus 1-2-3 and Quattro Pro.
- The Dutch National Archives' (2003) *From digital volatility to digital permanence: Preserving spreadsheets* report has guidance on preserving the authenticity and context of spreadsheets. While an older resource, it is still the most comprehensive advice on how to preserve the authenticity of spreadsheets.

A number of software tools are available for working with Spreadsheet data (COPTR, 2021).

5.2 Acquisition and appraisal

The retention of spreadsheets largely depends on an organization's collecting mandate and the acquisition of other records related to the spreadsheets.

- Often spreadsheets are created to perform calculations which are then incorporated into other records, or to organize information on a temporary basis, and appraisal may show their value as a record is limited and retention unnecessary.
- Some domains (such as the sciences or finance) may place a higher archival value on the content of a spreadsheet and require long-term preservation and access.
- Consider requesting the removal of DRM in spreadsheets before acquisition. Spreadsheets with any DRM features that prohibit access or use should not be collected if possible (LC, 2020-2021).

5.3 Authenticity and context

- Spreadsheets need to be accompanied by relevant descriptive metadata, to provide context. Ensure that spreadsheets and associated records have at least these attributes: the name of the organization, purpose, date, and relationship with other files.
- Write documentation that contains information about the original and current file formats, any associated software used to create or render the spreadsheets, and any preservation actions that have been taken. The minimum requirements for supporting authentic records are worksheets in the correct sequence with correct names, rows and columns; that retain their structure, embedded objects and formulae; and that have correct relationships between cells (Dutch National Archives, 2003).
- Retain aesthetic characteristics like font weight and colours (particularly where it conveys meaning) when possible, but know that exact replication is not always essential.

5.4 Backwards compatibility

- Consider the potential (or lack of) for backwards compatibility. For example, modern versions of Microsoft Excel software such as Office 365 can be used to read files created in the earlier Excel 95 edition of the software and saved in the Excel 95 format (Microsoft, 2021).
- Legacy software documentation can also provide guidance on which software is appropriate for backwards compatibility. The Let's Solve the Format Problem (2019a) wiki discusses backwards compatibility for many legacy spreadsheet formats. Microsoft (2021) has compatibility checkers in modern Microsoft Excel software, and provides a list of compatible formats.
- Relying on backwards compatibility alone may not be sufficient to address continuity of access. Commercial applications may remove support for obsolete formats, or may themselves become obsolete.

5.5 Preservation action

- Retain original files to keep options open over time. Quality checking preservation actions is typically costly, challenging and lacking in tool support for automation.
- An emulation approach (Morrissey, 2020) can be applied to preserve and provide access to spreadsheets and the software with which they were created.
- Migration from one format to another may provide a viable approach in some cases but could result in the loss of some data and/or functionality, particularly dynamic content.
- Migration to a lower fidelity format (such as CSV, Text, or PDF/A) may simplify preservation challenges of the resulting files, whilst accepting the loss of some functionality and/or data. This strategy may be useful in tandem with the more complex emulation or full migration approaches.
- Be aware of macros and any other scripting functionality. Create quality assurance workflows that help ensure that data are not being lost during any preservation processing. It is recommended to store macros separately from spreadsheets (Dalgleish, 2020; ADS, 2009).

5.6 Characterization

- Identify file formats with a tool such as DROID (The National Archives, n.d.), FIDO (Open Preservation Foundation, 2020), or Siegfried (Lehane, 2020) that uses the PRONOM file format registry (The National Archives, 2020).
- Use PDF validation tools such as JHOVE (Open Preservation Foundation, 2020) and VeraPDF (Open Preservation Foundation, 2020) to check conformance to published PDF specifications and metadata standards. Validation tools for other spreadsheet formats are not available currently.

6 References

Archaeology Data Service [ADS] (2009) *Databases and spreadsheets: A guide to good practice.* Available at: https://web.archive.org/web/20201215095638/https://guides.archaeologydataservice.ac.uk/g2gp/DbSht_Toc

COPTR (2021) *Spreadsheet.* Available at: https://web.archive.org/web/20210705083538/https://coptr.digipres.org/index.php/Spreadsheet

Dalgleish, D. (2020) Contextures Blog. Available at: https://web.archive.org/web/20200623025727/https://contexturesblog.com/archives/2020/01/30/keep-notes-on-excel-formulas-and-macros/

Dingledy, F. W. and Matamoros, A. B. (2016) *What is Digital Rights Management?* Available at: https://web.archive.org/web/20200821144227/https://scholarship.law.wm.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1121&context=libpubs

Dutch National Archives (2003) *From Digital volatility to digital permanence: Preserving spreadsheets.* Available at: https://web.archive.org/web/20130903083642/http://en.nationaalarchief.nl/sites/default/files/docs/kennisbank/volatility-permanence-spreadsh-en.pdf

ISO (2020a) *ISO/IEC 26300-1:2015: Information technology — Open Document Format for Office Applications (OpenDocument) v1.2 — Part 1: OpenDocument Schema.* Available at: https://web.archive.org/web/20201223100915/https://www.iso.org/standard/66363.html

ISO (2020d) *ISO/IEC 32000-2:2020: Document management — Portable document format — Part 2: PDF 2.0.* Available at: https://web.archive.org/web/20201218045916/https://www.iso.org/standard/75839.html

ISO (2018) *ISO/IEC 32000-1:2008: Document management — Portable document format — Part 1: PDF 1.7.* Available at: https://web.archive.org/web/20201125031501/https://www.iso.org/standard/51502.html

Klindt, M. (2017) *PDF/A considered harmful for digital preservation.* Available at: https://web.archive.org/web/20200917210028/https://ipres2017.jp/wp-content/uploads/15.pdf

Lehane, R (2020) *Siegfried.* Available at: https://web.archive.org/web/20201028192837/https://github.com/richardlehane/siegfried

Let's solve the format problem (2020) *Quattro Pro.* Available at: https://web.archive.org/web/20200714113418/http://fileformats.archiveteam.org/wiki/Quattro_Pro

Let's solve the format problem (2019a) *Documents.* Available at: https://web.archive.org/web/20200701000610/http://fileformats.archiveteam.org/wiki/Document

Let's solve the format problem (2019b) *Lotus 1-2-3.* Available at: https://web.archive.org/web/20200702010518/http://fileformats.archiveteam.org/wiki/Lotus_1-2-3

Library of Congress (2021) *TSV, Tab-Separated Values.* Available at: https://web.archive.org/web/20210211153542/https://www.loc.gov/preservation/digital/formats/fdd/fdd000533.shtml

Library of Congress (2020-2021) *Recommended formats statement: Datasets.* Available at: https://web.archive.org/web/20201116041423/http://www.loc.gov/preservation/resources/rfs/data.html

Library of Congress (2020) *OpenDocument Spreadsheet Document Format (ODS), Version 1.2, ISO 26300:2015.* Available at:

https://web.archive.org/web/20201031173456/https://www.loc.gov/preservation/digital/formats/fdd/fdd000439.shtml

Library of Congress (2017a) *Sustainability of digital formats: Planning for Library of Congress collections.* Available at: https://web.archive.org/web/20201113043001/https://www.loc.gov/preservation/digital/formats/intro/intro.shtml

Library of Congress (2017b) *XLSX Transitional (Office Open XML), ISO 29500:2008-2016, ECMA-376, Editions 1-5.* Available at: https://web.archive.org/web/20201120235444/https://www.loc.gov/preservation/digital/formats/fdd/fdd000398.shtml

Microsoft (2021) *File formats that are supported in Excel.* Available at: https://web.archive.org/web/20201111210012/https://support.microsoft.com/en-us/office/file-formats-that-are-supported-in-excel-0943ff2c-6014-4e8d-aaea-b83d51d46247

Mitcham, J. (2017) *How can we preserve Google Documents?* Available at: https://web.archive.org/web/20201029193019/http://digital-archiving.blogspot.com/2017/04/how-can-we-preserve-google-documents_35.html

Morrissey, S. (2020) *Preserving Software: Motivations, Challenges and Approaches*. Available at: http://doi.org/10.7207/twgn20-02

North Carolina Department of Cultural Resources [NCDCR] (2012) *File Format Guidelines for Management and Long-Term Retention of Electronic records.* Available at: https://web.archive.org/web/20201031113655/http://digitalpreservation.ncdcr.gov/file_formats_in-house_preservation.pdf

Open Preservation Foundation (2020) *Format Identification for Digital Objects (FIDO).* Available at: https://web.archive.org/web/20200916134739/https://github.com/openpreserve/fido

Open Preservation Foundation (2020) *JHOVE.* Available at: https://web.archive.org/web/20201031215050/https://openpreservation.org/products/jhove/

OpenPreservation Foundation (2020) *veraPDF.* Available at: https://web.archive.org/web/20201031220541/https://openpreservation.org/products/verapdf/

Technopedia (2020) *Macro.* Available at: https://web.archive.org/web/20210123172223/https://www.techopedia.com/definition/3833/macro

The National Archives (n.d.) *Digital Object Record Identification (DROID).* Available at: https://web.archive.org/web/20201015033155/https://github.com/digital-preservation/droid

The National Archives (2020) *The Technical Registry: PRONOM.* Available at: https://web.archive.org/web/20201111032324/http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

The National Archives (2013) B*est practice guide to appraising and selecting records for The National Archives.* Available at: https://web.archive.org/web/20200310033845/https://www.nationalarchives.gov.uk/documents/information-management/best-practice-guide-appraising-and-selecting.pdf

van Veenendaal, R., Kjærskov, F. H., Sein, K., O'Sullivan, J., Nielsen, A. B., Tømmerholt P. M., and Tømmerholt, J. (2019) *Significant properties of Spreadsheets.* Available at: https://web.archive.org/web/20191202082644/https://ipres2019.org/static/pdf/iPres2019_paper_48.pdf

W3C (2006) *Extensible Markup Language (XML) 1.1 (Second Edition).* Available at: https://web.archive.org/web/20201023210151/https://www.w3.org/TR/2006/REC-xml11-20060816/

Wijsman, L. (2020) *The Significant Properties of Spreadsheets: Stakeholder Analysis.* Available at: https://doi.org/10.5281/zenodo.3971833

Young, P. (2021) *What's Up, (with Google) Docs? – The Challenge of Native Cloud Formats.* Available at: https://web.archive.org/web/20210304124326/https://www.dpconline.org/blog/whats-up-with-google-docs