Spreadsheet Procedures

Version 1.118

Created date:	26 January 2012
Last updated:	23 August 2017
Review Due:	23 August 2018
Authors:	Jon Bateman, Jen Mitcham, Gary Nobles, Michael Charno, Kieron Niven, Tim Evans, Ray Moore and Jenny O'Brien
Maintained by:	Digital Archivists
Previous version:	Live

1 Purpose of this document

This page is designed to document current ADS procedures for production of dissemination and preservation copies of spreadsheet data. It contains a list of current dissemination/preservation formats and how to migrate files to required formats. More information on this data type, can be found in the *Guides to Good Practice for Databases and Spreadsheets*.[1]

[1] http://guides.archaeologydataservice.ac.uk/g2gp/DbSht_Toc

2 Formats

Offered format	Accepted	Preservation	Presentation	Notes
Microsoft Excel **.xls**	YES	Comma separated values **.csv (preferred)** or Microsoft Office Open XML **.xlsx**	Comma separated values **.csv (preferred)** or Microsoft Office Open XML **.xlsx**	See footnote.[2]
Microsoft Office Open XML **.xlsx**	YES	Microsoft Office Open XML **.xlsx**	Microsoft Office Open XML **.xlsx**	Can be preserved/disseminated in this format. See above for notes.
OpenOffice Calc 2.0 **.ods**	YES	Comma separated values **.csv (preferred)** or OpenOffice Calc 2.0 **.ods**	Comma separated values **.csv (preferred)** or Microsoft Office Open XML **.xlsx**	See footnote.[3]
OpenOffice Calc 1.0 **.sxc**	NO	N/A	N/A	

[2] Microsoft Excel spreadsheets (XLS and more recently .xlsx) are the most commonly deposited formats and though the software is widely used, Excel spreadsheets are only suitable as a deposit format. For both preservation and presentation they normally need to be migrated to non-proprietary formats CSV). There are exceptions to this rule. When some spreadsheets are transferred to CSV, their significant properties are not preserved. This could be the case if there is complicated formatting (which conveys meaning), use of special characters/symbols which do not translate into CSV, or formula which are considered to be a significant property and can not be preserved in any other way. Where CSV is deemed to be an inadequate preservation format, an XML based file format can be used instead. This may be XLSX or B- it is worth having a go at both of these as one may be better than another for preserving the significant properties in question. In one recent example, an Excel spreadsheet was saved as ODS but this did not preserve the values within the cells, just the formulas. Watch out for special characters too - they may translate differently in XLSX and ODS. Where XLSX or ODS are used as a preservation format, make sure that figures/charts within the spreadsheet are also preserved in a suitable format.

[3] Open Office.org Calc is freely available from http://www.openoffice.org/. It's native spreadsheet format is a compressed XML file, with the actual spreadsheet data being stored in a file called content XML and images in a separate directory. ODS can be used as a format for preservation where csv is not able to adequately preserve all the significant properties of a file. However, It may be worth trying file conversions to XLSX too in order to see which is most effective. Note that Calc provides character set options when exporting to CSV. This is a good option if you need to get from XLS to CSV with UTF-8 encoding.

Delimited text (tab, pipe etc) **.txt**	YES	Comma separated values **.csv**	Comma separated values **.csv**	See footnote.[4]
Lotus 1-2-3 **.123/.wk***	NO	N/A	N/A	
Quattro Pro **.wq***	NO	N/A	N/A	
Portable Document Format **.pdf**	Only if no other version available	Portable Document Format/Archive 1-B **.pdf**	Portable Document Format/Archive 1-B **.pdf**	See footnote.[5]

3 Documentation / Metadata

The following documentation is required for any spreadsheet.

Element	Description
Filename	
Description	
Creation Date	Either date created or date of last edit/deposit.
Software used	
Software version	
Supporting documentation filename	
Sheet Documentation	

[4] Tab delimited text (TAB) is another suitable preservation format for spreadsheet data but it is not routinely used by the ADS any more. There may be some data in our archive that has been preserved like this in the past. Spreadsheet data cells are separated by tabs and should not be enclosed by double or single quotes. A new row in the data is indicated by the new line character. This overcomes the problems sometimes found with CSV format as tab characters are not routinely used within spreadsheet data. Indeed, it is impossible to add a tab or a new line character to a cell in Microsoft Excel as these keystrokes are used for navigating around the spreadsheet. Comma separated values (CSV) is the preferred format for dissemination of spreadsheet data. Cells are separated by a comma and enclosed in double quotes. It is a format that most people are familiar with, and has the added advantage that the files will open directly into MS Excel for those users who have this software. This makes it easier for people who are not so comfortable with importing delimited data into a spreadsheet application of their choice. This format is also suitable for preservation.

[5] Portable Document Format (PDF) can occasionally be used to disseminate spreadsheet data. This should only be used where there is information in the spreadsheet which can not adequately be replicated in a CSV file. Where possible, CSV should be used so that users can more effectively re-use the data. It may be necessary to provide a pdf version alongside a CSV version for complicated spreadsheets so that users can see what the spreadsheet looked like but also work with the data. Obviously, PDF(A) should be used rather than standard pdf if we are creating pdfs for dissemination.

Sheet Name	
Sheet Description	
Field Name	
Field Description	

Supporting Documentation

This should include:

- codes used
- units of measurement used in specific fields

4 Accessioning checks

- Where spreadsheets contain multiple sheets, is it clear what each sheet represents? The sheet needs a meaningful name or a heading, or some separate documentation that describes its contents
- Are all the sheets intended for preservation? Sometimes people deposit spreadsheets with sheets that are just random workings out and calculations and are not intended for preservation NB: Check sheet description.
- Embedded objects: many spreadsheet applications allow users to embed other media (especially images) within files. Spreadsheet applications such as Microsoft Excel and OpenOffice Calc allow users to embed graphs and charts generated from data along with other images. Any spreadsheets using charts/graphs should include a caption within the supporting documentation.

Significant properties

- The actual data within the spreadsheet (there may be several sheets of data) - including cell headings and the values themselves. Associated with this is the use of special characters in the dataset, from ampersands to greek characters (common in dating/scientific data). These must be identified and preserved.
- It is advisable that such content is stored and archived separately thereby retaining the original qualities of the content (e.g. image resolution) and allowing it to follow a separate archival strategy to the textual content.
- Formatting/Layout. This is quite tricky, and the Digital Archivist has to decide this on a case by case basis - is the use of colour, formatted text significant? Does the layout need preserving? Are notes/comment fields relevant? Are there manual line breaks in fields (see below)?
- Macros/fomulae: in the past we have (generally) not preserved these properties, being more concerned with the data. However, in some cases it may be decided that these are significant, and need preserving as part of the archive. Hopefully in most cases, this will be highlighted in the documentation (see below). If these are to be preserved then it should be made clear in the supporting documentation.

Comments or Notes

These are preserved in XLSX as a separate .xml file (*/xl/comments.xml*) so this may influence the choice of preservation/dissemination formats.

If the spreadsheet is to be saved as a CSV, then these comments fields will need extracting from the original file and saving as a .txt file. There's no easy way to do this, apart from copying and pasting the text into TextPad, recording the cellid and saving under /documentation/. Sounds boring? Then go with XLSX.

Formulae, Queries, Macros

If the file contains complex formulae or queries that need to be preserved in their own right then these need to be identified as migrated versions of the spreadsheet or database may only preserve the actual values calculated by the functions and not the functions themselves. Complicated formulae (e.g. those linking worksheets) or queries may need to be preserved separately and documented within a text file so the spreadsheet functionality can be recreated at a later date.

5 How to convert files

Problems:

Does the layout and formatting of the text convey extra meaning? For example do headings span multiple rows or columns or is information conveyed through use of colour, borders or font?

> The spreadsheet data will need to be edited by hand before migration to ensure that meaning is not lost. For example, merged cells must be split and the text within them duplicated within each cell.

Does the file contain complex functions/formulae that need to be preserved in their own right?

> Migrated versions of the spreadsheet data will only preserve the actual values calculated by the functions, not the functions themselves. Complicated formulae may need to be preserved separately within a text file so the spreadsheet functionality can be recreated at a later date.

Does the file contain macros or Visual Basic scripts that need to be preserved?

> Macros can be preserved as plain text if necessary.

Do any of the cells contain textual notes or comments?

> Migration process described below will not save comments. Before migration, comments will need to be stored in a separate text file with a clear indication of which cell the comment relates to.

Does the spreadsheet contain multiple sheets where the sheets are linked by formulae?

> Migration strategies for complex spreadsheets such as these should be decided on a case by case basis.

Does the spreadsheet contain hidden cells?

> Information within hidden cells probably needs to be preserved too (check with depositor) so ensure your migration strategy takes account of this.

Does the spreadsheet contain symbols that can not be converted to UTF-8?

Odf seems to do a good job of preserving special characters, but it is also worth trying xlsx.

Unicode data not displaying properly

Create your Unicode CSV file either with MS Access or Open Office as described. Then open the CSV file in Notepad (make sure its Notepad - Textpad won't do it but will mess up all the funny characters). Then save as in Notepad and specify the character encoding as UTF-8. You will then have a CSV file that displays the characters correctly and opens in MS Excel

Line breaks in cells (only first line will export to CSV)

Search and replace for line breaks. Open the Replace dialogue box and, in the 'find what' box, type 010 while holding alt (a dot will appear). Ensure that the search is for 'values'. Add appropriate character to the replace box.

Images within files

Ideally these should have been deposited as separate raster files, but if no, please follow these procedures: Save file as an XLSX. Using a file manager such as 7-Zip or WinZip, 'extract files'. This should create an unzipped folder with the same name as the file. Within this structure should be the folder /xl/media/ with all images in their original format (the format they were put in the file in). Convert these images to uncompressed tiff (v6). Store in the /preservation/tif/ folder. Ideally, replicate the archive structure and put the files in a folder with the same name as the original file, so for example: /preservation/tif/finds_spreadsheet/my_finds_spreadsheet/. Create a CMS process record detailing this event, making sure to record the location of the preservation images.

File naming

Where possible files should retain the same name as the original sheet.

Where multiple worksheets of a spreadsheet are being converted, folder name should reflect the name of the original spreadsheet and the name of the sheet the data came from, for example:

myspreadsheet-[sheet name].csv

Where extra files are created to hold images, notes and formulae, these should also be named in logical way that makes it easy to trace exactly where they came from, for example

myspreadsheet-[sheet name]-[chart name].tif

Then placed in prescribed location (see below).

6 Post-migration checking

- Check row counts after export
- Check text field lengths
- Check all sheets have been exported
- Check any special characters have been preserved.
- Check whether cells contain line breaks (only 1st line will export to **.csv**, solution above)

Storage

Data should be stored in appropriately named folders, as described in the ADS Repository Operations document.[6] Any directory structure from the SIP should try to retained in the AIP. In some cases editing/restructuring may be required, any restructuring must be recorded in the Process table in the CMS.

Spreadsheets will often be used to record file-level and project metadata and archive documentation. Where this is the case, files should be stored in a relevant folder, for example:

Otherwise, store data in one of the following directory structure:

```
/preservation
        /{original_structure}
                myspreadsheet-sheet1.csv
                myspreadsheet-sheet2.csv
                myspreadsheet-sheet3.csv
                myspreadsheet-sheet3-chart1.tif
                myspreadsheet-sheet3-chart2.tif

/dissemination
        /{original_structure}
                myspreadsheet-sheet1.csv
                myspreadsheet-sheet2.csv
                myspreadsheet-sheet3.csv
                myspreadsheet-sheet3-chart1.jpg
                myspreadsheet-sheet3-chart2.jpg
```

Storing metadata

File metadata should be stored in an appropriate archival format with the preservation/dissemination files in a "documentation" folder within the requisite file type, for example:

```
/preservation
        /{original_structure}
                myspreadsheet-sheet1.csv
                myspreadsheet-sheet2.csv
                myspreadsheet-sheet3.csv
        /documentation
                myspreasheet_metadata.csv

/dissemination
        /{original_structure}
                myspreadsheet-sheet1.csv
                myspreadsheet-sheet2.csv
                myspreadsheet-sheet3.csv
        /documentation
                myspreasheet_metadata.csv
```

[6] http://archaeologydataservice.ac.uk/advice/RepositoryOperations.xhtml