SPREADSHEET PROCEDURES (VERSION 1.132)

DIGITAL ARHIVISTS
ARCHAEOLOGY DATA SERVICE
https://archaeologydataservice.ac.uk/

Created date:	26 January 2012
Last updated:	18 December 2019
Review Due:	31 March 2021
Authors:	Jon Bateman, Jen Mitcham, Gary Nobles, Michael Charno, Tim Evans, Kieron Niven, Ray Moore, Jenny O'Brien, Teagan Zoldoske, Digital Archivists
Maintained by:	Digital Archivists
Required Action:	
Status:	Live
Location:	https://archaeologydataservice.ac.uk/advice/PolicyDocuments.xhtml

1. Purpose of this document

1.0.1 This documents current ADS procedures for production of dissemination and preservation copies of spreadsheet data. It contains a list of current dissemination and preservation formats and how to migrate files to required formats. More information on this data type, can be found in the G2GP for Databases and Spreadsheets http://guides.archaeologydataservice.ac.uk/g2gp/DbSht_Toc.

2. Formats[1]

Offered format	Accepted	Preservation	Presentation	Notes
Comma Separated Values **.csv**	YES	Comma Separated Values **.csv**	Comma Separated Values **.csv**	Although probably a rarity, files should be checked to confirm that they are comma separated (with qualifiers) rather than the similarly-named 'Character Separated Values' format. As with all text formats, character encoding should be checked.
Microsoft Excel **.xls**	YES	Comma Separated Values **.csv** (or in certain cases Microsoft	Comma Separated Values **.csv** (or in certain cases Microsoft Office Open XML **.xlsx**)[2] [3]	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-

[1] Generally speaking CSV is our preferred method of preservation and dissemination. XLSX should only be used when CSV is not preserving formatting within the spreadsheet.

		Office Open XML **.xlsx**)[2] [3]		proprietary formats (.csv). For exceptions see Footnote[2]
Microsoft Office Open XML **.xlsx**)	YES	Comma Separated Values **.csv** (or in certain cases Microsoft Office Open XML **.xlsx**)[2] [3]	Comma Separated Values **.csv** (or in certain cases Microsoft Office Open XML **.xlsx**)[2] [3]	Where possible XLSX files should be converted to CSV although rare cases may require preservation/dissemination as XLSX. For exceptions see Footnote[2]
OpenOffice Calc 2.0 **.ods**	YES	Comma Separated Values **.csv** (or in certain cases OpenOffice Calc 2.0 **.ods**)[2]	Comma Separated Values **.csv** (or in certain cases OpenOffice Calc 2.0 **.ods**)[2]	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

[2] 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 (e.g. colour 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 .ods - 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] 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/. Consequently, .xlsx may be more appropriate.

				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.
OpenOffice Calc 1.0 **.sxc**	NO			
Delimited text (tab, pipe etc) **.txt**	YES	Comma Separated Values **.csv**	Comma Separated Values **.csv**	Notes[4]
Lotus 1-2-3 .**123**/ **.wk***	NO			
Quattro Pro **.wq***	NO			
Portable Document Format **.pdf**	YES, but only if no other version available	Portable Document Format **.pdf**	Portable Document Format **.pdf**	Portable Document Format (.pdf) can occassionally be used to disseminate spreadsheet data. This should only be used where there is information in the

[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. 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.

				spreadsheet which can not adequately be replicated in a csv file. [5]

3. Documentation / Metadata

3.0.1 Alongside the standard metadata for files, the following additional documentation is required for any spreadsheet. The current metadata template is available from the Guidelines for Depositors.[6]

3.0.2 The following documentation is required for any data in spreadsheets

Element	Description
Filename	This is the name of the file.
Title	This is the title of the spreadsheet.
Description	A brief description of the contents of the spreadsheet.
Creator	The individual(s) or organisation(s) responsible for the creation of the spreadsheet.
Copyright holder	The individual(s) or organisation(s) who holds the copyright for the spreadsheet.
Period of Creation	The start and end date for the creation of the spreadsheet.
Language	The language(s) used within the spreadsheet.
Software used	The software that was used to create the spreadsheet.
Software version	The version of software used to create the spreadsheet.
Sheet Documentation	
Sheet Name	This should be used to record the name of each individual worksheet used within the spreadsheet.
Sheet Description	Provide a brief description of the function of the worksheet.

[5] 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.

[6] https://archaeologydataservice.ac.uk/advice/guidelinesForDepositors.xhtml.

Number of Rows	Provide a figure for the number of rows used within the worksheet. This provides an easy mechanism to audit gaps in the data.
Field Documentation	
Field Name	Please provide a list of fields used within the worksheet.
Field Description	Here you should provide a brief textual description of the data contained within each field.
Supporting Documentation	
Supporting documentation	Enclose any supporting documentation associated with the spreadsheet, this would typically include any codes, abbreviations, measurements or terminology utilised in the spreadsheet.

3.1 Associated metadata

3.1.1 It is important that the metadata should include documentation of any extra features the spreadsheet may contain, i.e. formulae, macros, charts, comments and any significant characteristics to be preserved.

4. Accessioning checks

4.1 Checks

- Do 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.
- Formulae, Queries, Macros 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.

4.2 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)?
- 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 manger 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.

- The workbook contains a large amount of worksheets DAs should Count worksheets (in order to check that any batch export has worked on all sheets) csv files should then be named accordingly as described below.
- 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.

4.3 File-naming

4.3.1 Where possible files should retain the same name as the original. On occasion (and normally for dissemination), it may be necessary to create different versions of the same file. In these cases a logical naming strategy should be used, and should be accompanied by explanation in the Processes section of the CMS.

4.3.2 Where a spreadsheet contains multiple sheets these will, typically, be preserved individually. Consequently, each file should prepend the original spreadsheet name to the sheet name.

myspreadsheet-sheet1.csv
myspreadsheet-sheet2.csv
myspreadsheet-sheet3.csv

4.3.3 In instances where the spreadsheet includes embedded content, e.g. graphs, illustrations etc., each object should be extracted and stored separately. Each file should prepend the original spreadsheet name to the object name.

myspreadsheet-sheet3-chart1.tif
myspreadsheet-sheet3-chart2.tif

4.3.4 If the spreadsheet uses formula or queries that are regarded as a significant property these should be extracted and stored in an appropriate format for preservation and named accordingly

myspreadsheet-forumlas.txt
myspreadsheet-queries.pdf

4.3.5 All files and metadata should be placed in the appropriate location as outlined below.

5 How to convert files

Starting Format	Procedure	End Format	Checks
Microsoft Excel **.xls** or Microsoft Office Open XML **.xlsx**	**Microsoft Excel (Batch exporting sheets using VBA)** 1) Press Alt + F11 keys simultaneously to open the 'Microsoft Visual Basic Application' window. 2) In the 'Microsoft Visual Basic Application' window, click Insert > Module. 3) Copy and paste the following code into the Module window. Code in footnotes below [1] 4) Press the F5 key to run the code. You will see all exported csv files are located on the 'Documents' folder.	Comma Separated Values **.csv**	• 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)

6 Storage

6.1 Storing data

6.1.1 Data should be stored in appropriately named folders, as described in the ADS Repository Operations manual.[7] Any directory structure from the SIP should be retained in the AIP. In some cases editing/restructuring may be necessary, but such restructuring should be recorded in the Processes section of the CMS.

/preservation
    /{original_structure}
        mytable1.csv
        mytable2.csv
        mytable3.csv

/dissemination
    /{original_structure}
        mytable1.csv
        mytable2.csv
        mytable3.csv

6.1.2 In instances where the spreadsheet includes multiple tables and/or embedded content then these should be stored accordingly

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

6.1.3 In those cases where large numbers of sheets/tables are included it may be appropriate to group sheets together, for example

/preservation
    /{original_structure}
        /myspreadsheet1
            myspreadsheet1-sheet1.csv
            myspreadsheet1-sheet2.csv
            myspreadsheet1-sheet3.csv
            myspreadsheet1-sheet3-chart1.tif
            myspreadsheet1-sheet3-chart2.tif

[7] https://archaeologydataservice.ac.uk/advice/PolicyDocuments.xhtml#RepOp.

/myspreadsheet2
myspreadsheet2-sheet1.csv
myspreadsheet2-sheet2.csv

6.2 Storing metadata

6.2.1 File and associated metadata (formula, queries, etc) should be stored in an appropriate archival format with the preservation/dissemination files in a "documentation" folder within the requisite folder. Any metadata extraction should be recorded in the Processes section of the CMS.

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

6.2.2 For dissemination

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

6.2.3 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.

7. Creating and linking objects in the OMS tables

7.0.1 See Match Objects Overview for general overview {internal access only}
see also CMS-OMS TableStructure for MOS data requirements {internal access only}

8. Tech watch / things to note

9. Archival notes

10. References