BINARY AND PLAIN TEXT PROCEDURES (VERSION 1.146)

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

Created date:	26 January 2012
Last updated:	13 January 2020
Review Due:	31 March 2021
Authors:	Jo Gilham, Jen Mitcham, Kieron Niven, Tim Evans, Gary Nobles, Ray Moore, Jenny O'Brien, Leontien Talboom, 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 text files. 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 Text Documents http://guides.archaeologydataservice.ac.uk/g2gp/TextDocs_Toc.

2. Formats

Offered format	Accepted	Preservation	Presentation	Notes
Microsoft Word **.doc**	YES	Microsoft Office Open XML **.docx**	Portable Document Format/Archive 1-A **.pdf**	
Microsoft Office XML **.docx**	YES	Microsoft Office Open XML **.docx**	Portable Document Format/Archive 1-A **.pdf**	
OpenOffice.org 2.0 **.odt**	YES	OpenOffice.org 2.0 **.odt**	Portable Document Format/Archive 1-A **.pdf**	
Rich Text Format **.rtf**	YES	Microsoft Office XML **.docx**	Portable Document Format/Archive 1-A **.pdf**	
Portable Document Format **.pdf**	YES	Portable Document Format **.pdf** (as deposited)	Portable Document Format **.pdf** (as deposited)	PDF files should be preserved as deposited. See PDFGuidance.[1]
Plain text **.txt**	YES	Plain text **.txt**	Plain text **.txt**	Text encoding
XML **.xml**	YES	XML **.xml**	XML **.xml**	Text encoding, DTD or schema

[1] Internal access only.

SGML/HTML /XHTML **.sgml**, **.html**, **.xhtml**	YES	SGML/HTML/XHTML **.sgml**, **.html**, **.xhtml**	SGML/HTML/XHTML **.sgml**, **.html**, **.xhtml**	DTD and character encoding
Apple Pages **.pages**	NO			

3. Documentation / Metadata

3.0.1 Alongside the standard metadata for files, the following additional documentation is required for any text files. The current metadata template is available from the Guidelines for Depositors.[2]

3.0.2 Submissions through OASIS are subject to their own requirements. Metadata for these submissions are stored in two locations - technical metadata (within the OMS) and file-level metadata (within the ADS Library schema).

3.0.3 Submissions as part of a journal or series destined for dissemination through the ADS Library are subject to there own specific metadata requirements. A metadata template is available for the ADS Google Drive.[3]

3.0.4 In instances where text documents are submitted as part of an archive, but are regarded as 'library worthy', file level metadata must be transferred to the ADS Library using either the dedicated uploader, or through the online forms. This essentially duplicates the metadata stored within the OMS.

Element	Description
File name	
Title	
Abstract	
Author/editor (each individual should be added on a new row)	
Page Count	
Date Published	
Publisher	

[2] https://archaeologydataservice.ac.uk/advice/guidelinesForDepositors.xhtml.
[3] Restricted access.

Place Published	
Volume/ Issue/Report Number	
ISBN	
DOI	
URL	
Language	
Software	
Software Version	
Supporting Documentation	
Copyright/GDPR Clearances	

3.1 Associated metadata

3.1.1 It is important that any copyright information/permissions are stored alongside the requisite file in a suitable preservation format.

4. Accessioning checks

4.1 Checks

- Do we have necessary documentation?
- Is document recorded in OASIS? (see What to do with Grey Lit)[4]
- If PDF files are sent but no Word doc (or similar) originals, ask if they have originals so that we can create preservation copies.
- Check that PDF files are not 'secured' i.e. password protected (you can tell this by opening them and seeing if '(secured)' appears in the title bar next to the document file name. Some security can be removed in Acrobat X (Tools > Protection > Encrypt > Remove Security), or that the file is not corrupted or includes content which may cause problems to those viewing the PDF. NB This is particularly important following the change to the preservation workflow for PDF as problems are not likely to be identified further down preservation pathway.
- Look out for compressed word documents as deposits. There's no indication in the file name, but on opening Word tries to reassemble the document using a series of sub-documents, unless these are included in the deposit you get an error message "hyperlink reference is not valid".

[4] Internal access only.

- If a deposited MS Word doc uses the 'track changes' feature, need to inform data producer that we will only archive the final version of the document and not maintain the track changes functionality.
- Check for any OLE objects in Word doc? Check we are able to preserve these too.
- Check for Sensitive Data[5]

4.2 Significant properties

- The words and order of the words in the document
- The hierarchical structure of the document (ie: different levels of headings)
- Formatting within the document (ie: where some text might be highlighted in bold or italics)
- The page numbering of a document. This is particularly important where a document is a published or unpublished report or thesis. If people want to use the resource they may wish to cite and reference exact page numbers. We therefore need to ensure that the same page numbering is maintained in each migration of the document.
- Any other non-text content (images, data tables etc). This may in some situations need to be preserved separately
- The properties that we generally do not see as significant are as follows:
- Track changes functionality (we are primarily interested in the final version of a document)

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

[5] http://archaeologydataservice.ac.uk/advice/sensitiveDataPolicy

5 How to convert files

Starting Format	Procedure	End Format	Checks
Microsoft Word **.doc**	**Using Microsoft Word.** 1) Open the required file 2) File > Save As. 3) Select location where file should be saved. Enter filename and the appropriate 'Save as type' (i.e. Word Document (*.docx))	Microsoft Office Open XML **.docx**	• Number of pages is the same as original • Word count matches original • Images are in their correct position on the page, and on the right page. • Formatting has been retained • If a PDF: OCR is present in file • The document has edits/track changes disabled • check formatting, pagination is not lost
Microsoft Word **.doc**	**Using Microsoft Word.** 1) Open the required file 2) File > Save As. 3) Select location where file should be saved. Enter filename and the appropriate 'Save as type' (i.e. PDF (*.pdf)). From the 'Options' ensure that 'ISO 19005-1 compliant (PDF/A)' is ticked	Portable Document Format/Archive 1-A **.pdf**	As above.
Microsoft Word **.doc**	**Using the batch conversion software Kutools - https://www.extendoffice.com/product/kutools-for-word.html** This add-in has a set of pre-written macros that can be used. When it has been downloaded two tabs will have been added to Word. Within the Enterprise tab there's a button on the left hand side called 'DOC/DOCX' to convert the files. This will open a pop-up window. Here you can select the directory with the doc files you want	Microsoft Office Open XML **.docx** or Portable Document	As above.

	to convert. It is also possible to convert the files from DOC to PDF.	Format/Archive 1-A **.pdf**	

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.[6] 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}/
                mydocument.docx
                other_document.pdf

/dissemination
        /{original_structure}/
                mydocument.pdf
                other_document.pdf
```

6.2 Storing metadata

6.2.1 File metadata should be stored in an appropriate archival format with the preservation/dissemination files in a "documentation" folder within the requisite folder.

```
/preservation
        /{original_structure}
                mydocument.docx
                other_document.pdf
                /documentation
                        document_metadata.csv
```

6.2.2 For dissemination, in many circumstances dissemination of separate/downloadable versions of the metadata may not be necessary as much/all important metadata can be displayed within the archive interface. However, in some circumstances it may be necessary to disseminate the metadata separately.

```
/dissemination
        /{original_structure}
                mydocument.pdf
                other_document.pdf
                /documentation
                        document_metadata.csv
```

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

6.2.3 Text documents 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. In such circumstances care should be taken to ensure that the appropriate data type has been assigned to the file (typically Documentation or Admin). This can be carried out within the CMS,

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

Item	Person	Date
See PDFGuidance.[7] Following discussions at the 'Technical and Preservation Strategy' meeting (17th May 2019) it was agreed that the extant preservation strategy involving the normalisation of PDF to a form of PDF/A would cease. Further discussion at a 'CATs Meeting' (25th June 2019) identified some concerns with the use of 'password protection' and corrupt files. This also highlighted that some form of normalisation would still be required to preserve PDF in stable versions of the format (e.g. PDF 1.4 or PDF 1.7), but that this would be managed through an archive wide program of normalisation on a regular basis, rather than being enacted at ingest.	DAs	17/05/2019

10. References

- ADS Guide to Good Practice http://guides.archaeologydataservice.ac.uk/g2gp/TextDocs_Toc
- PDF to PDF/A conversion. Page based on previous policy of PDF/A normalisation which is retained for future reference.
- JISC Technology Watch Report, XML-based Office Document Standards (TSW0702) (August 2007) http://archive.alt.ac.uk/alt.newsweaver.co.uk/www.jisc.ac.uk/media/documents/techwatch/tsw0702pdf.pdf
- OpenOffice Migration Guide (May 2006): https://www.openoffice.org/documentation/manuals/oooauthors/MigrationGuide.pdf
- Recommendations for the creation of PDF files for long-term preservation and access (Koninklijke Bibliotheek, 2007): https://docplayer.net/24113254-Pdf-guidelines-recommendations-for-the-creation-of-pdf-files-for-long-term-preservation-and-access.html

[7] Internal access only.