Text Procedures

Version 1.130

Created date:	26 January 2012
Last updated:	01 December 2017
Review Due:	24 August 2018
Authors:	Jen Mitcham, Jo Gilham, Kieron Niven, Tim Evans, Gary Nobles, Jon Bateman, Ray Moore and Jenny O'Brien
Maintained by:	Digital Archivists
Previous version:	Live

1 Purpose of this document

This page is designed to document the current ADS procedures for production of dissemination and preservation copies of binary texts (word processed documents) and plain texts. It contains a list of current dissemination/preservation formats and how to migrate files to these formats.

For more information on this data type, please refer to the *Guides to Good Practice for Text and Documents*.[1]

2 Formats

We have adopted a **.docx** / **.pdfa** strategy, although if we receive files in **.odt** or **.txt** they can be left in their original format. Alot of Binary text files contain non-text content, please see the relevant section on how to preserve these elements.

Offered format	Accepted	Preservation	Presentation	Notes
Microsoft Word **.doc**	YES	Microsoft Office Open XML **.docx**	Portable Document Format/Archive 1-A **.pdf**	
Microsoft	YES	Microsoft Office	Portable	

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

Office XML **.docx**		Open XML **.docx**	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 Open XML **.docx**	Portable Document Format/Archive 1-A **.pdf**	
Portable Document Format **.pdf**	YES	Portable Document Format/Archive 1-B **.pdf**	Portable Document Format/Archive 1-B **.pdf**	
Plain text **.txt**	YES	Plain text **.txt**	Plain text **.txt**	Text encoding
XML **.xml**	YES	XML **.xml**	XML **.xml**	Text encoding, DTD or schema
SGML / HTML / XHTML **.sgml** / **.html** / **.xhtml**	YES	SGML / HTML / XHTML **.sgml** / **.html** / **.xhtml**	SGML / HTML / XHTML **.sgml** / **.html** / **.xhtml**	DTD and character encoding

3 Documentation / Metadata

File metadata is based on the *Guides to Good Practice*.[2] Not all fields are applicable to all documents (i.e. Grey Literature).

Element	Description
Title	Title of document.
Abstract	Short description of document.
Date Published	Year of publication.
Published	Full location details of any published versions. Start and end page numbers or number of pages should also be recorded.
Publisher	Name of document publisher and location.
ISBN	ISBN number (where applicable).
DOI	Digital Object Identifier

[2] http://guides.archaeologydataservice.ac.uk/g2gp/TextDocs_Toc

URL	URL (where applicable).
Related files / resources	Details of related files or resources.
Language	The language (English, Spanish, etc.) used within the document. It is recommended that a controlled vocabulary (e.g. RFC 4646) is used and specified.
Author	Name of primary author(s).
Contributor	Names of contributors, should be qualified e.g. Editor, Translator, Contributor.
Email	Email address for author.

4 Accessioning checks

- Do we have necessary documentation (see below)?
- Is document recorded in OASIS?[3]
- 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)
- File validation: Check that files submitted as PDF/A actually verify as PDF/A, if not remove PDF/A information from the file prior to ingest
- 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".
- 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.

Significant properties

The significant properties that we need to be preserving are as follows:

- 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:

[3] http://oasis.ac.uk/pages/wiki/Main

- Track changes functionality (we are primarily interested in the final version of a document)

Significant properties may change depending on the exact nature of the document being preserved. All files should be looked at on a case-by-case basis.

5 How to convert files

Starting format	Procedure	End format
Preservation		
PDF	See the guidance on the PDF/A page[4]	PDF/A 1B
DOC	Open in MS Word DOC, and convert to DOCX (97-2003 compatible)	DOCX
RTF	Open in MS Word DOC, and convert to DOCX (97-2003 compatible)	DOCX
Dissemination		
PDF	See the guidance on the PDF/A page[5]	PDF/A 1B
DOC	Open file in MS Word DOC, convert to PDF/A 1A	PDF/A 1A
RTF	Open file in MS Word DOC, convert to PDF/A 1A	PDF/A 1A

Batch conversion For a lot of doc files, create a macro in MS Word.[6] Further guidance is available here.[7] The macro code on that page can also be edited to do the same for rtf. Batch processing of PDF-PDF/A files is supported in Callas.[8]

Adobe Acrobat has a batch processing function (under Advanced => Document Processing => Batch Processing tool), but I've never got this to work brilliantly. For example, the processing of individual files still requires involvement via Preflight, so can't be left to its own devices.

Batch validation can be done in Adobe.[9]

File naming

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.

[4] Available from ADS internal wiki.
[5] Available from ADS internal wiki.
[6] http://answers.microsoft.com/en-us/office/forum/office_2007-word/is-it-possible-to-convert-a-batch-of-doc-files-to/8f5549be-8143-4f0a-8f82-06c855fcb092
[7] https://www.datanumen.com/blogs/3-quick-ways-to-batch-convert-word-doc-to-docx-files-and-vice-versa/
[8] https://www.callassoftware.com/en
[9] Available from ADS internal wiki.

In these cases a logical naming strategy should be used, and should be accompanied by explanation in the Process table. Then placed in prescribed location (see below).

6 Post-migration checking

We should ensure data consistency by undertaking a check of files post migration. The number of files to be checked is at the discretion of the Digital Archivist. Things to check include:

- 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

At present there are no tools to do this, see JLM's call for help here.[10]

April 2012: The SCAPE project has agreed to take this on as a problem to work on. We have provided them with sample files from the GLL to work with. Watch this space! A tool may be available to us in the future!

Storage

Data should be stored in appropriately named folders, as described in the ADS Repository Operations manual.[11] 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.

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

```
/preservation
        /{original_structure}
                mydocument.docx

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

Storing metadata

File metadata (copyrights, documentation, etc) 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}
                mydocument.docx
                /documentation
                        mydocument_metadata.docx

/dissemination
        /{original_structure}
```

[10] http://wiki.opf-labs.org/display/REQ/Checking+that+significant+properties+are+preserved+after+migration
[11] http://archaeologydataservice.ac.uk/advice/RepositoryOperations.xhtml

mydocument.pdf
/documentation
mydocument_metadata.pdf