

1. TEXT DOCUMENTS

Definition / examples

Documents conveying written information primarily intended for reading. Examples include plain text files (txt), pdfs, documents produced by word processing applications (e.g. Microsoft Word, docx) and desktop publishing software (e.g. Adobe InDesign, idn).

Preferred formats	<u>Preservation format</u>	<u>Access format</u>	<u>Level of support</u>
Acrobat PDF/A (pdf)	Original	Portable Document Format (pdf)	Full
Portable Document Format (pdf)	Acrobat PDF/A (pdf)	Original	Full

Acceptable formats	<u>Preservation format</u>	<u>Access format</u>	<u>Level of support</u>
Microsoft Word (doc, docx)	Original	Portable Document Format (pdf)	Watch
OpenDocument Text (odt)	Acrobat PDF/A (pdf)	Portable Document Format (pdf)	Normalize
Plain Text File (txt)	Original	Portable Document Format (pdf)	Full
Rich Text Format (rtf)	Acrobat PDF/A (pdf)	Portable Document Format (pdf)	Normalize
WordPerfect (wpd)	Original	Original	Bit-level

Do not transfer	<u>Preservation format</u>	<u>Access format</u>	<u>Level of support</u>
Google Document Link File (gdoc)	n/a	n/a	Cannot be preserved

Description / guidance

Preservation format(s) – PDF Archival (PDF/A) is the Archives' designated preservation format. It is an open specification, and it preserves both the informational content and the visual appearance of text documents.

Plain text files are left in their native formats, but most other formats will be normalized where possible to pdf/a as the preservation copy. Proprietary formats (e.g. Microsoft Word) cannot be reliably converted to PDF/A in the Archives' preservation system (Archivematica). This is because Archivematica is an open-source software and as such cannot incorporate proprietary tools that can reliably automate bulk conversion.

Where feasible, the Archives may manually normalize files in proprietary formats; typically however they are currently left as is in their original formats.

Access format(s) – PDF is the Archives' main access format. Software for viewing pdfs is widely available, and pdf can be readily displayed online in the Archives' Access-to-Memory (AtoM) application.

The Archives' preservation system (Archivematica) can automate the conversion of some proprietary formats (e.g. Microsoft Word) to PDF via OpenOffice. The resulting pdfs do not always reliably preserve visual formatting; they are adequate for access (but not preservation) purposes.

Transfer guidance – Ideally producers should convert files in proprietary formats (e.g. Microsoft Word) to pdf before transfer to Archives.

Do not transfer Google link files with the GDOC extension. These are pointer files that reference documents that remain on Google Drive. The Archives cannot typically access these drives. If you wish to transfer material from Google Drive, export the documents as PDF and transfer the pdfs to the Archives.

Current gaps / strategy – The Archives currently has no ability to ensure the accessibility of older, obsolete proprietary desktop publishing or word-processing formats (e.g. WordPerfect). These are not common in the Archives' holdings, so addressing this gap is not a high priority at present.

A more pressing issue is to develop a workflow for manually normalizing current proprietary formats like Microsoft Word to PDF/A preservation copies as part of regular archival processing.