http://doi.org/10.7207/twgn21-07

# Preserving Documents

## Data Types Series

Artefactual Systems and the Digital Preservation Coalition

## DPC Technology Watch Guidance Note

July 2021

DigitalPreservationCoalition

ISSN: 2048-7916

First published in Great Britain in 2021 by the Digital Preservation Coalition.

## The Data Type Guidance Note Series

Each Guidance Note in the Data Types series is designed to provide a primer on the current state of community knowledge about data types commonly encountered by those seeking to preserve digital holdings. Digital preservation is about keeping information findable, usable, and trustworthy over the long-term. The best approach for any repository will vary according to the scope and content of its holdings, available resources, and the expectations of its funders and users. There are however, broadly applicable good practices that have been established as a result of many years of research, practical implementation, and consensus building. These are presented here as a starting point, along with additional resources for further exploration.

This series of Data Type Guidance Notes has been authored by staff at Artefactual Systems in collaboration with the Digital Preservation Coalition. These notes have been developed in conjunction with the UK Nuclear Decommissioning Authority.

Digital preservation is an evolving field and continues to change and develop in response to external drivers and fresh challenges. New formats, standards, and examples of good practice will emerge over time and the information contained within this report will need to be updated. We welcome comments and feedback to: info@dpconline.org.

# 1 Overview

The term ‘document’ has various broad definitions, and even narrow definitions still encompass many file formats (Buckland, 1998). This Guidance Note will focus on documents as being primarily textual in nature. Documents can be born-digital or digitized from analogue sources, and can range from meeting notes or memos to electronic records, legal documents, and publications. Documents can be related to one another, and indeed often derive much of their meaning from relationships with other documents. Documents may be complete and final versions, drafts, non-draft versions, or copies.

# 2 Preservation Challenges

## 2.1 Lack of records management

Poor or absent records management can have a negative impact on the ability to transfer documents successfully to an archive. If documents are managed in a digital recordkeeping system (such as an EDRMS), records managers need to consider how to extract both documents and related metadata from the system for transfer. Ideally this should be considered when the system is being implemented.

## 2.2 Embedded content

Born-digital documents can have objects embedded in them, often raster images or spreadsheets, less commonly multimedia files and even other documents. Embedded digital objects may not be rendered correctly if opened in software that is different from the document-creating application.

## 2.3 References between files

Documents often contain references to other digital objects such as other documents, multimedia files, spreadsheets or websites. These non-embedded objects may also require preservation in order for the document to retain its full context and meaning.

## 2.4 Missing or incorrectly substituted fonts

A document may depend on fonts that are not embedded within the document file itself. On rendering a file with missing fonts, applications may substitute similar, or sometimes entirely dissimilar, fonts. The result might range from minor aesthetic differences to a complete loss of content where equations or barcodes are present. Research focussing on legacy Word documents indicated that this problem affected around 1 in 5 documents (Brown and Woods, 2009).

## 2.5 Cloud-based formats

Cloud-based document files are widely used, and facilitate collaboration, version control, and document sharing in remote work environments. Cloud-based documents can be exported into a range of formats, including DOCX, PDF, HTML, RTF, and plain text. This process can change the text formatting or other features of the document. For example, pagination may be altered and last modified dates changed or lost (Young, 2021; Mitcham, 2017). The existence of such documents only on the web and ease of alteration by different users makes them complex objects to preserve.

## 2.6 Digital rights management

Digital rights management (DRM) is a set of technical measures designed to constrain the use of digital files, typically to protect intellectual property rights such as copyright (Dingledy and Matamoros, 2016). Encryption and password protection can also be used for other purposes such as restricting access to personal information. Specific restrictions might relate to opening, copying, saving, or printing files, which may hinder their preservation and re-use.

## 2.7 Legacy formats

There have been numerous word processing programs over the years, many of them producing documents in proprietary file formats. Accessing and reading legacy documents can be complicated, and even if a modern word processor or a text editor can render them, there may be a loss of significant properties. Document formatting such as fonts, pagination, headers and footers, tables of contents, bulleted lists, and text underlining or italics, may also be lost or altered in legacy document formats, and embedded objects may fail to open.

# 3 File formats

There is no single perfect file format for the preservation and future use of documents. Decisions made on file formats should be dependent on the features and functionality to be preserved and the future use cases to be supported. Note that the table below does not provide an exhaustive list of formats suitable for preservation and access. The most suitable format for preserving the important features and functionality of a file may be the original format that it was created in. It is recommended that careful research and analysis is carried out before migrating files to a new format.

| File format | Extensions | Brief summary |
|---|---|---|
| Electronic Publication (EPUB) | .epub | EPUB is an open standard used in publishing since 2007, compatible with e-readers and software for tablets, computers, and mobile phones. It is a container format consisting of text-based files, image files, and supporting metadata files. EPUB files can contain DRM technical protection measures that may be a hindrance to long-term preservation. The Library of Congress (LC) considers EPUB to be an acceptable preservation format 'if the content files are not encrypted, if the file is not subject to technological protection that inhibits long-term preservation and access, and if all content is stored within the EPUB container.' (LC, 2020). |
| OpenDocument Text Document Format (ODT) | .odt<br>.ott | ODT is an open file format specification which, like DOCX, uses XML as its underlying structure. ODT is natively supported by the open-source LibreOffice word processing application. The format is a member of the Open Document Format (ODF) family of XML-based files, and is standardized as ISO/IEC 26300-1:2015. Like DOCX, the fact that it is XML-based and supported by open specifications and standards (although incredibly complex) makes it an acceptable preservation format (LC, 2020-2021; National Archives and Records Administration [NARA], 2020). |
| Plain text | .txt | Plain text files contain characters and whitespaces (such as spaces and line breaks), but have no formatting or embedded objects. They are program-independent and do not rely on specialized software to display as human-readable text. Plain text is ubiquitous, being the underlying format for software code, computer log files, marked-up files such as HTML, XML, JSON, SGML and LaTeX, scalable vector |

| | | |
|---|---|---|
| | | graphics files, and the email backup and export formats MBOX and EML. Any file that can be rendered fully in a text editor is considered a plain text file and is generally suitable for long-term preservation. |
| Portable Document Format/Archive (PDF/A) | .pdf/a | PDF/A is a set of PDF standards designed by Adobe Systems for long-term preservation. Based on PDF 1.4, PDF 1.7 or PDF 2.0, PDF/A prohibits functionality that poses concerns for long-term preservation such as linked fonts and encryption. It also restricts the use of digital signatures, which may make it an unsuitable preservation format for certain types of documents.<br>The different versions and conformance levels (PDF/A-1 to PDF/A-4) allow conformant files to contain different features. The most recent version, PDF/A-4, was released in 2017 and substantially revised in 2020. PDF/A is widely used for both preservation and access. |
| Portable Document Format (PDF) | .pdf | PDF was developed by Adobe Systems in 1993 as a proprietary presentation format for documents. In 2008, it was released as the open standard ISO 32000-1:2008. PDF 2.0 was released in 2017 and updated in 2020 as ISO 32000-2:2020. The PDF Association has numerous resources available about PDF versions and different types. PDF is compatible with a variety of rendering applications and is a common access format. |
| Rich Text | .rtf | Rich Text format is plain text with additional formatting characters and tags, supported natively by Microsoft's Wordpad application and by earlier versions of Microsoft Word. RTF was originally designed to provide interoperability of formats between different versions of Microsoft Word, and between Microsoft Word and other word processing and desktop publishing applications. Although it is a proprietary format, the RTF specification is published, and the Library of Congress considers it to be an acceptable preservation format (LC, 2017). |
| Google Docs | N/A | The structure and format of a Google Doc is opaque to the user who is only able to view a rendered version of the (cloud stored) data, in their web browser. Google Docs can be exported into a range of formats, including DOCX, PDF, HTML, RTF, and plain text. This process may change the formatting and/or result in the loss of certain features. The UK National Archives is currently working on practices for exporting Google Docs and relevant metadata (Young, 2021). |
| TeX | .tex | TeX (or LaTeX) files are marked-up plain text files used to create output formats, such as PDF, for digital distribution. Use of TeX is common in science and mathematics because of the format's |

| | | |
|---|---|---|
| | | accurate rendering of formulae (The LaTeX Project, n.d.). TeX files may not often make their way to a digital repository, since they are an early-state format designed to generate other formats for publication and distribution. If they do, the published versions of the files should also be acquired whenever possible. |
| Word Document Format | .docx<br>.doc | Open Office XML (DOCX) is the default file format of Microsoft Word software, having replaced Microsoft Word Document (DOC) in 2007. Standardized as<br>ISO/IEC 29500-1:2016, it is a container format that packages up a set of XML files to provide structure and formatting when rendered by software.<br>Given that the format is XML-based, has an open specification and international standard, and is widely used, DOCX is an acceptable preservation and access format (LC, 2020-2021). DOC file format versions are listed as acceptable preservation formats by NARA (2020) but not the Library of Congress. |

# 4 Tips for creators

Creators working in government, business, or other controlled environments should be aware of their organizations' document handling and records management policies. Adherence to guidelines and requirements for file formats, document formatting, versioning, and metadata creation will help ensure that the documents can be preserved in such a way as to retain their context and meaning over time. Academic settings often provide detailed guidelines for different types of documents (such as theses and publications), which should be followed as closely as possible. Organizations without format policies might find resources such as the Library of Congress' *Recommended Formats Statement* useful substitutes.

Individuals who intend to donate documents to an archive should use well-supported and ubiquitous file formats whenever possible, and be prepared to describe when, how, and for what purposes the documents were created and disseminated.

Other tips for records creators include the following:

- If embedding other digital objects in your files, retain external copies of the embedded objects and transfer them to the archives along with the documents whenever possible.
- If migrating to an alternative document file format, consider retaining the original source file along with the final-state file; for example, retain .tex or .docx files that have been printed to PDF.
- Be aware that using non-standard fonts for presenting content such as equations or barcodes could be problematic for preservation.
- If creating a PDF version, use the application that was used to create the document to generate a PDF version of it; for example, later versions of Microsoft Word come with the ability to export to PDF built in. Otherwise, consider using tools such as Adobe Acrobat, PDF Studio (Qoppa Software, 2021) or pdfaPilot (Callas Software, n.d.).

- If saving to PDF/A, remember to visually inspect the document, and to use the software's PDF/A conformance check functionality if available. Underlying issues with the document could result in the generation of a non-conformant PDF/A file (Oates et al., 2018).
- If creating PDF/A for long-term preservation, consider using conformance level A for born-digital documents and conformance level B for digitized documents (Oates et al, 2018). Both versions have requirements for visual presentation of the files on a computer screen, but Conformance level A is more stringent, requiring content to be structured and tagged in specific ways to facilitate accessibility.

# 5 Tips for archivists

## 5.1 General guidance

The following resources provide guidance on preserving and providing access to documents:

- Library of Congress (2020-2021) Recommended formats statement: ii. Textual Works – Digital.
- The National Archives (2012) Managing digital records without an electronic record management system.
- A number of software tools are available for working with Document data (COPTR, 2021).

## 5.2 Acquisition and appraisal

- Work closely with records managers and/or records creators to help lay the groundwork for successful transfer of documents to archives.
- If records are being extracted from recordkeeping systems, ensure that the metadata being exported with them are sufficient to describe the context of their creation and use during the active stages of their lifecycle.
- If acquiring documents from private individuals, several interviews with the creator(s) may be needed to ensure that the documents can be arranged and described in a meaningful way.
- Archivists should be aware of any DRM or other protection mechanisms, and negotiate with the donor to relax the restrictions, obtain encryption keys or consider rejecting the documents if the restrictions are likely to impede the ability to preserve them (also see Characterisation, below).

## 5.3 Preservation action

- Retain original files. Migration will often result in a loss of formatting and content, but may provide some insurance against format obsolescence. Retaining the original files will facilitate an emulation approach, should this become necessary in the future.
- Be cautious when migrating to PDF/A. Generating reliable, conformant PDF/A documents from word processing files (including PDF files) can be problematic (Klindt, 2017; Oates et. al., 2018).
- For cloud based documents, a dedicated exporting service such as Google Takeout Services may be used, but quality assurance must be undertaken to ensure relevant significant properties are not negatively affected (Young, 2021; Mitcham, 2017). Web archiving is also another option for capturing cloud based documents in their native format (Young, 2021).

## 5.4 Legacy documents

- When acquiring older word processing documents, be aware of any printed versions of the documents already residing in the archives. If printed copies exist it may be more practical to preserve them than to preserve the digital versions, unless the digital versions have metadata and other attributes that add needed context to the documents.
- Adopt an emulation strategy if documents cannot be reliably rendered in modern software. It is now possible to emulate certain early word processing programs using emulation platforms such as WordStar Emulator (Wordstar, 2020), PCjs Machines (PCjs, 2021) and vDosWP (Columbia University, 2020). Projects such as Scaling Emulation as a Service Infrastructure (EaaSI) (Educopia Institute, 2020), led by Yale University, and Emulation as a Service, led by University of Freiburg (2020), aim to provide simplified and scalable access to emulation infrastructure.

## 5.5 Characterization

Characterization can be useful to identify file formats, extract metadata, identify broken or encrypted content, or check conformance to profiles or standards. Tool support and effectiveness can vary considerably for different file formats.

- Identify file formats with a tool such as DROID (The National Archives, n.d.), FIDO (Open Preservation Foundation, 2020), or Siegfried (Lehane, 2020) that uses the PRONOM file format registry (The National Archives, 2020).
- Use PDF validation tools such as JHOVE (Open Preservation Foundation, 2020) and VeraPDF (Open Preservation Foundation, 2020) to check conformance to published PDF specifications and metadata standards. Use LibreOffice's ODF Validator (LibreOffice, n.d.) for validating Open Document Format (ODF) standard conformance.
- Characterization tools can be employed to identify files with DRM, encryption or other protection methods

## 5.6 Metadata

- Documents can contain descriptive and technical metadata generated automatically during creation and modification. Some document creation tools support the ability to add additional descriptive metadata manually.
- Embedded metadata can be viewed and modified using applications such as the open-source ExifTool (Harvey, 2021) and commercial tools such as GroupDocs Metadata (2020) and MetaWiper (2021). The ability to view embedded metadata may be diminished over time as the file formats approach obsolescence.

# 6 References

Brown and Woods (2009) *Born Broken: Fonts and Information Loss in Legacy Digital Documents.* Available at: https://web.archive.org/web/20201030203247/https://escholarship.org/uc/item/53z897zb

Buckland, M (1998) *What is a "digital document"?*. Available at: https://web.archive.org/web/20201124085206/https://people.ischool.berkeley.edu/~buckland/digdoc.html

Callas Software (n.d.) *pdfaPilot.* Available at: https://web.archive.org/web/20210107171004/https://www.callassoftware.com/en/products/pdfapilot

Columbia University (2020) *vDosWP System for WPDOS under 64-bit Windows.* Available at: https://web.archive.org/web/20201130042537/http://www.columbia.edu/~em36/wpdos/vdoswp.html

COPTR (2021) *Documents.* Available at: https://web.archive.org/web/20210707081618/https://coptr.digipres.org/index.php/Document

Digital Preservation Coalition (2020) *Handbook Glossary: Electronic Records*. Available at: https://web.archive.org/web/20201028205756/https://www.dpconline.org/handbook/glossary#E

Dingledy F. W., and Matamoros, A. B. (2016) *What is Digital Rights Management?* Available at: https://web.archive.org/web/20200821144227/https://scholarship.law.wm.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1121&context=libpubs

Educopia Institute (2020) *Scaling Emulation as a Service Infrastructure (EaaSI) (subcontract)*. Available at: https://web.archive.org/web/20201129194437/https://educopia.org/emulation-as-a-service-eaasi/

Google Developers (2019) *Structure of a Google Docs document*. Available at: https://web.archive.org/web/20201101023029/https://developers.google.com/docs/api/concepts/structure

GroupDocs (2020) *GroupDocs.Metadata*. Available at: https://web.archive.org/web/20200809000226/https://products.groupdocs.app/metadata/docx

Harvey, P. (2021) *Exiftool.* Available at: https://web.archive.org/web/20210122225751/https://exiftool.org/

ISO (2020) *ISO/IEC 26300-1:2015 Information technology — Open Document Format for Office Applications (OpenDocument) v1.2 — Part 1: OpenDocument Schema*. Available at: https://web.archive.org/web/20201112002833/https://www.iso.org/standard/71691.html

ISO (2016) *ISO/IEC 29500-1:2016: Information technology — Document description and processing languages — Office Open XML File Formats — Part 1: Fundamentals and Markup Language Reference.* Available at: https://web.archive.org/web/20201112002833/https://www.iso.org/standard/71691.html

ISO(2018) *ISO/IEC 32000-1:2008: Document management — Portable document format — Part 1*: PDF 1.7. Available at: https://web.archive.org/web/20201125031501/https://www.iso.org/standard/51502.html

ISO (2020) *ISO/IEC 32000-2:2020: Document management — Portable document format — Part 2: PDF 2.0*. Available at: https://web.archive.org/web/20201218045916/https://www.iso.org/standard/75839.html

InterPARES Project (1999-2018) *Glossary*. Available at: http://interpares.org/display_file.cfm?doc=ip1_glossary.pdf [accessed 25 January 2021]

Klindt, M. (2017) *PDF/A considered harmful for digital preservation*. Available at: https://web.archive.org/web/20200917210028/https://ipres2017.jp/wp-content/uploads/15.pdf

Lehane, R (2020) *Siegfried.* Available at: https://web.archive.org/web/20201028192837/https://github.com/richardlehane/siegfried

Library of Congress (2020-2021) *Recommended formats statement: ii. Textual Works - Digital*. Available at: https://web.archive.org/web/20201107175543/https://www.loc.gov/preservation/resources/rfs/text.html

Library of Congress (2020) *Sustainability of Digital Formats: Planning for Library of Congress Collections: EPUB (Electronic Publication) File Format Family*. Available at: https://web.archive.org/web/20201023234157/https://www.loc.gov/preservation/digital/formats/fdd/fdd000310.shtml

Library of Congress (2017) *Sustainability of Digital Formats: Planning for Library of Congress Collections: Rich Text Format (RTF) Family*. Available at: https://web.archive.org/web/20200917211110/https://www.loc.gov/preservation/digital/formats/fdd/fdd000473.shtml

LibreOffice (n.d.) *ODF Validator.* Available at: https://web.archive.org/web/20210708081233/https://odftoolkit.org/conformance/ODFValidator.html

MetaWiper (2021) *MetaWiper.* Available at: https://www.metawiper.com/word-metadata-changer [accessed 25 January 2021]

Mitcham, J. (2017) *How can we preserve Google Documents?*. Available at: https://web.archive.org/web/20201029193019/http://digital-archiving.blogspot.com/2017/04/how-can-we-preserve-google-documents_35.html

Oates, A. I., Downie, S. J., Halvarsson, E., and Popham, M. (2018) *Navigating the PDF/A Standard: A Case Study of Theses in Oxford's Institutional Repository*. Available at: https://www.ideals.illinois.edu/handle/2142/100236

Open Preservation Foundation (2020) *Format Identification for Digital Objects (FIDO).* Available at: https://web.archive.org/web/20200916134739/https://github.com/openpreserve/fido

Open Preservation Foundation (2020) *JHOVE.* Available at: https://web.archive.org/web/20201031215050/https://openpreservation.org/products/jhove/

OpenPreservation Foundation (2020) *veraPDF.* Available at: https://web.archive.org/web/20201031220541/https://openpreservation.org/products/verapdf/

PCjs (2021) *PCjs machines.* Available at: https://web.archive.org/web/20210108064056/https://www.pcjs.org/software/pcx86/app/microsoft/word/5.00/

Qoppa Software (2021) *PDF Studio.* Available at: https://web.archive.org/web/20210102222204/https://www.qoppa.com/pdfstudio/.

Society of American Archivists (2005-2020) *Dictionary: document*. Available at: https://web.archive.org/web/20210107121155/https://dictionary.archivists.org/entry/document.html (Wayback Machine capture from 7 January 2021)

Tech Terms Computer Dictionary (2020) *Rich Text*. Available at: https://web.archive.org/web/20201022230151/https://techterms.com/definition/richtext

The LaTeX Project (n.d.) *LaTeX – A document preparation system*. Available at: https://web.archive.org/web/20210105071245/https://www.latex-project.org/

The National Archives (n.d.) *Digital Object Record Identification (DROID).* Available at: https://web.archive.org/web/20201015033155/https://github.com/digital-preservation/droid

The National Archives (2021) *The Technical Registry: PRONOM.* Available at: https://web.archive.org/web/20210108231304/https://www.nationalarchives.gov.uk/PRONOM/Default.aspx

The National Archives (2012) *Managing digital records without an electronic record management system*. Available at: https://web.archive.org/web/20200809232536/https://www.nationalarchives.gov.uk/documents/information-management/managing-electronic-records-without-an-erms-publication-edition.pdf

University of Freiburg (2020) *bwFLA — Emulation as a Service*. Available at: https://web.archive.org/web/20201103031720/http://eaas.uni-freiburg.de/

Wordstar (2020) *WordStar Emulator.* Available at: https://web.archive.org/web/20200601225007/http://www.wordstar.org/index.php/wordstar-emulator

Young, P. (2021) *What's Up, (with Google) Docs? – The Challenge of Native Cloud Formats.* Available at: https://web.archive.org/web/20210304124326/https://www.dpconline.org/blog/whats-up-with-google-docs