Preservation Handbook

[Binary Text / Word Processor Documents]

Author	Rowan Wilson and Martin Wynne
Version	Draft V3
Date	22 / 08 / 05
Change History	Revised by MW 22.8.05; 2.12.05; 7.3.06

Definition

'Binary' is used in contrast with 'plain text' in describing the encoding of textual data. In fact, plain text is actually a type of binary encoding, since any computer file is fundamentally a sequence of bits, but plain text is usually seen as a special case and differentiated from other binary encodings, and this is the usage which is preferred here. A 'plain text' file is a special case of a binary file which contains only text represented by alphabetic, numeric and punctuation characters. For the purposes of this handbook, a 'binary text' file is one in which textual data is encoded in some binary format which includes data other than standard encodings for alphabetic, numeric and punctuation characters.

A binary text document is usually created in a word-processor or as the result of transformation from another format. Microsoft Office is the most popular of these environments, but the Microsoft Word documents which it produces are not suitable for preservation without some form of conversion of the textual data to a format which conforms to open standards, such as Unicode. Rich Text Format is an open specification from Microsoft, but it attempts to amalgamate various Microsoft Word document standards in a tagging structure and should always be converted. Adobe PDF documents are a postscript-derived format, and are not suitable for preservation, and unless the recent addition of structural tagging to the format is exploited (and it rarely is), it has implicit accessibility issues. Wherever possible a conversion to a purely (non-binary) text-based human readable format should be undertaken. Conversions of the documents to plain text, or, where possible structured text formats (such as XHTML or XML) should be undertaken. The guidelines for the preservation of plain text and marked-up text should be consulted for guidance on suitable ways in which the textual data may be encoded in these formats.

Description

Word Processor files are designed to encode the representation of a document, and while the text may appear to have little formatting, the code which underlies the text includes much that is irrelevant to the content of the text, and more to help with the document's presentation and printing. The main problem with the preservation of binary formats is that they are dependent on the specific software that created the original document, and without this software it can be difficult to read the file in the original format intended. The multiplicity of different word processors compounds this problem.

Microsoft Word

MS Word is probably the most common desktop word processor used in the World today. Commonly Word forms part of the Microsoft Office Suite, the versions of which develop with every new release of Windows.

RTF (Rich Text Format)

The RTF Specification provides a format for text and graphics interchange that can be used with different output devices, operating environments, and operating systems. RTF uses the ANSI, PC-8, Macintosh, or IBM PC character set to control the representation and formatting of a document, both on the screen and in print. With the RTF Specification, documents created under different operating systems and with different software applications can be transferred between those operating systems and applications.

PDF

PDF combines the page layout language of PostScript used by many printers with font-embedding to ensure the required fonts are available and compresses them all into a single-file storage system. It is a binary and proprietary format, but Adobe have released the PDF specification as an open standard so that no royalties are incurred for reading or writing PDFs. More modern PDFs are able to embed XML tagging for structure and text alternative to non-textual elements, however while this ability exists, very few people creating PDFs recognise this or exploit its potential. Hence, generally, the text and structural markup in a PDF file is not in an accessible file format and should be converted to a text-based format like XML where possible.

WordPerfect

WordPerfect, a word processor similar to Microsoft Word, is owned by Corel. One of the benefits touted by its loyal supporters is the ability to reveal the formatting codes used by the word processor

in an editable manner and thus enable complete control by the users over the manner in which information is being stored. Nonetheless, its file format is binary and proprietary and should be converted to a text-based format like XML where possible. Where there is a large amount of relevant formatting and other presentational information in the file, RTF is preferable to WordPerfect, and there are few problematic issues in conversion.

OpenOffice.org

The OpenOffice.org file format has, with the release of OpenOffice version 2 in 2005, adopted the OASIS OpenDocument XML format as its default native file type. It will still read and write files created in earlier versions of the program. The OASIS OpenDocument XML format is very similar to the native XML format of the earlier version of OpenOffice, and both are stored as a compressed zip file which can be uncompressed with any freely available unzip software. The OASIS OpenDocument format is a vendor and implementation independent file format and guarantees freedom and independence from any particular hardware platform or software application. If it is necessary to preserve binary text in its binary form with presentational markup, then this document format is preferable to proprietary formats. (Although, strictly, preservation versions must also include an unzipped version of the file alongside.) There are existing filters for OpenOffice documents which allow exporting directly as TEI XML. If no intellectual content is lost through this conversion (for example the semantics of particular fonts in the context of a specific document) then this can be an acceptable route to provide a suitable preservation version.

Additional Information

Technical Environment

In general, binary file formats are not appropriate for long-term preservation of textual data. Where possible, resource creators should be informed of this and encouraged to prepare textual data in other ways for deposit with the AHDS. In some cases it is possible for the AHDS to convert files deposited in binary formats. OpenOffice is a suitable application for opening and exporting files from most word processors and exporting them as XML. There are few limitations on the type of binary data which can be encoded in a PDF file, so it is not possible to propose a general solution for there conversion. It may be possible to convert a file containing only textual data which has been encoded in a conventional manner in a PDF file using freely available open source tools or Adobe Acrobat Professional, although the extent to which formatting information can be preserved through this migration route is uncertain, due to the large number of ways in which it can be encoded in a PDF.

Common Formats

Format	File Extension	Notes
Adobe Portable Document Format	.pdf	Should be converted to text-based format. Although specification is widely distributed it is proprietary.
Microsoft Word Document	.doc	Proprietary and inherently not suitable for preservation. Should be converted to text-based format.
Rich Text Format	.rtf sometimes .doc	An open specification owned by Microsoft that, although text-based tagging, it has many variants and should be converted to a neutral text-based preservation format.
WordPerfect	.wpd .doc (and others)	Proprietary and dated formats that should be converted to a plain text-based preservation format.
OpenOffice (previously StarOffice)	.odp, .sxw (and others)	Although it can save in a number of proprietary formats, OpenOffice's default text format (.odp) is a zipped XML format. In its unzipped format it is suitable for preservation, but unwieldy. Where possible the existing export filters should be used to convert to a more common XML format (i.e. TEI XML).

Additional Information

Adobe PDF

< http://www.adobe.com/> Last checked 08/03/2006

Open Office

< http://www.openoffice.org/> Last checked 08/03/2006

Technical Information on Microsoft Word

< http://office.microsoft.com/en-us/FX010857991033.aspx> Last checked 08/03/2006

Help with file extensions

< http://filext.com/detaillist.php?extdetail=RTF> Last checked 08/03/2006

Ingest Checklist

Level 1 (Essential)

- Documentation of the specific software package and version used to create the original file.

- Check files for viruses.

- Features within the document which are dependent on the software (e.g. macros, character formatting, tables, footnotes) need to be evaluated. An analysis of the document should be performed before deciding which is the best preservation format.

- The output options provided by the original software should be checked, where possible. Output should be to plain text (possibly with markup) for preservation purposes.

- Once the preservation format has been agreed, the document should be saved as a plain text file and the original software and version used to create the original to be noted in the metadata.

- The plain text file should be given a random check to ensure that the key features of the document have been retained. If the original document contains embedded images, these should be removed and an appropriate marker indicated in the plain text file.

Level 2 (Preferred)

- The document should be saved to a plain text file and the details of the original software used to create the file recorded in the metadata. Non-textual content of the document (tables, images etc) should be evaluated and the appropriate preservation strategy applied.

- Document can be stored in RTF in the meantime while a suitable plain text format is investigated.

Level 3 (Best Practice)

- Advise the depositor on how to prepare the documents in a text format appropriate for long-term preservation and await deposit in such a format.

Inform Depositor

Depositor should be informed that binary text formats are not appropriate for long-term preservation, and that other formats are preferred. Depositor should be informed that some information may be lost if the document is heavily reliant on the internal features of a particular piece of software. If the original document is very large, or there are very many files, the depositor should be informed that the process to create a proper preservation format that captures all the original features may take some time. Very old or obsolete word processor files may only be accepted and preserved 'as is', as the AHDS cannot emulate all legacy platforms and applications.

Preservation

Significant Characteristics

One significant characteristic of textual data is the markup scheme employed. An electronic text may be marked up in several different ways, ranging from the so-called "markup-free" ASCII to full-scale SGML encoding. HTML markup is commonly used. Proprietary word-processing software contains its own form of markup, although the XML standard is preferred.

Another characteristic is the extent to which the edition is dependent on specific software. Some electronic texts require specific software to interpret them while others may be handled by a variety of different applications. In some cases, while the text is intrinsically independent, it is nevertheless distributed in a package with specific software.

In some cases the formatting is a significant characteristic but in other cases it is not. This needs to be assessed on a file-by-file basis. For example, where text is held in a complex table structure, that formatting is likely to be necessary to the understanding of the text and therefore will need to be preserved. However, whether the text in a document is aligned to the left or justified in both margins is probably not important. Similarly, the font and font size within a document is probably not significant *unless* it is giving extra meaning to the text.

If images (such as logos or maps) are included within the text it is likely they are also significant characteristics that need to be preserved (probably as separate files to the text itself).

The overall structure or architecture of the edition is significant. Scholarly electronic texts may be presented as simple linear text, but more complex approaches include a segmented linear text, with tables of contents; a collection of images; and a structure which is essentially a database. Combinations of some or all of these are also possible; it might be necessary to assess whether the page breaks are significant or irrelevant to the understanding of the text.

Technique

Establish a system for persistent names for stored digital objects. The naming system must be able to deal with the objects themselves, all the components of the object, and all the components which the objects depend on. The naming system must also be designed to survive changes to, and in, the archiving institutions.

Determine the significant properties of the object to be stored as this will help ensure that the stored object will contain the intellectual content of the original digital object. This can be done by using 'save as' ASCII text and adding references to images/tables in the appropriate place in text.

In HTML by removing the images and coding text and tables using HTML mark-up (preferably by hand to ensure code is clean and minimal). Another method would be to convert using MS Word or other package and then clean up with 'HTML Tidy', inserting links to images using <img> tag.

Determine what needs to be preserved, and preserve it. This may include: the original digital object, as a byte stream; the versions derived from the original object; the context of the object, including the purpose of the object; all the software intermediaries necessary to use the object; and how the object was created.

Validation of Exported Data

The plain text file should be examined to ensure that all the important and relevant features of the original have been captured. Where is has not be possible to capture the original feature, some kind of marker indicating this should be used. Special attention should be paid to documents which have tables, footnotes and embedded images, to ensure that that these are intelligible in the preservation format.

If the word processor document does contain images, these need to dealt with and preserved. The preservation strategy described in the AHDS Bitmap Image Preservation Handbook should be consulted.

Problems and Issues

The most common problem encountered, from an archival point of view, is simply the number of documents that are still prepared in proprietary formats, and the fact that many academics are unaware that these formats pose any preservation problems. Word, for example, is available to most academics on their desk-top set-up, but few realise that the features built-in to aid document creation are also the ones which cause problems when it comes to long term preservation. This situation is becoming more problematic, as recent versions of Word are now unable to read earlier version of files created by the same software. MS Word is also a common carrier of viruses, and thus all documents need to be virus scanned before they are copied onto a system or worked upon.

As a last resort, if the structure and format of a file is so important that out-putting the data to a plain text file will destroy valuable information about the file, and there are not the resources available to convert to structured markup conforming to open standards, then saving the file as a TIFF image should be considered as a means of recording this information.

Additional Information

Common problems with MS Word

< http://www.goldmark.org/netrants/no-word/attach.html> Last checked 08/03/2006