

An investigation of file formats in use by SHERPA DP repositories

Document details

Project: SHERPA DP
Work Package: 6.1
Author: Gareth Knight
Version: Version 1.0
Document date: 02/03/2007
Change history:

<i>Date</i>	<i>Version</i>	<i>Author</i>
02/03/2007	First version	Gareth Knight

Contents

Introduction	2
File formats held by institutional repositories.....	2
Significant properties of an e-print.....	3
Analysis of repository holdings.....	4
Export of repository holdings to a suitable archival format	5
Export of repository holdings to a suitable dissemination format.....	7
Recommendations	7
References	8

Introduction

The file formats held by institutional repositories represent an easy and convenient method to store research papers in an electronic state. Unlike print resources, however, the hardware and software available to access these electronic formats are likely to change and part, or all, of the research paper may be rendered inaccessible. This paper outlines the preservation strategy to be taken in the Sherpa DP project, indicating the significant properties that must be preserved and the problems that are likely to be encountered when performing migration.

Digital preservation may be considered as the managed activities necessary to ensure the ongoing maintenance of a digital resource and allow continued access to the intellectual content in the long-term (Jantz, 2005). Preservation is a speculative activity that is taken to avoid possible scenarios in which content access is lost, either through the file format being rendered obsolete in later versions of the software or content becoming corrupted. There is no definitive approach to the problem of maintaining digital content across multiple generations of technology. Nevertheless, it is important that organizations with a responsibility to preserve research data should declare to their stakeholders the extent to which they are able to perform this function (OCLC-RLG, 2001). For the SHERPA DP project, the AHDS will be responsible for the preservation of holdings held by 12 institutional repositories. A two-tier preservation approach will be taken:

1. Preservation of the original bitstream that represents the information stored in a digital resource.
2. Preservation of the significant properties of the resource through migration to a file format that will remain accessible over time.

In practical terms, the AHDS will store an original digital object in a file format chosen by the institutional repository (as stated in their deposit guidance) and an archival derivative for preservation purposes. A dissemination copy, stored in another file format, will be generated on request at a later date.

The lifecycle of an e-print is determined at the point of creation through the software application that an author uses to write their paper and the file format in which it is stored. Computer software typically contains a format specification that dictates how information should be organized and stored within the object. By saving their research paper in Microsoft Word, for example, an author is, often unknowingly, specifying how content may be accessed at a later date. In the short-term, the research paper may be interpreted and displayed in the software that was used to create it. However, problems are likely to occur in the long-term when several factors are likely to endanger the accessibility of the resource: the file format may not be supported in later versions of application software; the software developer goes out of business or may cease to support the file format; or the software may be rendered inaccessible when the operating system or hardware is upgraded.

File formats held by institutional repositories

It is considered good practice to limit the number of file formats accepted by a digital repository and to store data in formats that are open or well documented (Knight, 2004). The costs and risks associated with digital preservation tend to grow when a digital collection includes a larger number of diverse file formats (James et-al, 2004; Granger, Russell & Weinberger, 2000) and will increase the risk of loss of access to the intellectual content over time. Institutional repositories participating in Sherpa DP, as well as the Sherpa project as a whole, take a pragmatic approach to the file formats accepted for deposit. In most repositories, several formats are accepted for submission and further work is performed to migrate the content to a standard dissemination format. Table 1 identifies the file formats accepted for deposit and dissemination.

Repository	Deposit Format	Dissemination Format
Birkbeck ePrints Archive	Microsoft Word, PDF, text	PDF

Edinburgh Research Archive	PDF, Microsoft Word, etc.	PDF, DVI, Postscript, JPEG
Glasgow ePrints Service	PDF, Microsoft Word	PDF
Glasgow ERPAePrints Service	PDF	PDF
Glasgow Jelit Service	PDF	PDF
KCL ePrints	PDF, web site (HTML, GIF, Jpegs)	PDF
LSE	PDF	PDF
Nottingham ePrints	PDF	PDF
Royal Holloway	PDF, Postscript	PDF
SOAS	PDF, web site (HTML, GIF, Jpegs), ASCII text	PDF, ASCII text
UCL EPrints	PDF	PDF
White Rose	PDF, HTML, ASCII text, and other non-specified formats	PDF

Table 1: File formats accepted for deposit and dissemination

For convenience, repository staff typically indicate PDF as the preferred deposit format, but accept other file formats that may be convenient for the depositor (ASCII, HTML, Postscript, Rich Text Format, Microsoft Word) to create or are common within a specific subject discipline (TeX, LaTeX). PDF is an open de facto format for electronic documents that is maintained by Adobe Systems Inc. The format is designed to combine the page layout language of Postscript (a scripting languages used by many printers) with font-embedding to ensure the layout of a document is reproduced in a way that is independent of operating system and hardware configuration.

Content migration is often necessary to export the intellectual content and make it available in a file format that is accessible by the majority of target users. For consistency, PDF is widely-used as a dissemination format, however a small number of records are held in a DVI, postscript and JPEG formats (Edinburgh Research Archive). As may be expected from the recent development of e-print archives, the most recent version of PDF (1.6) is also the most common, however a number of e-prints identified possess a creation date in the late 1990s and have been exported using an earlier version (1.1, 1.3, 1.4) of the PDF file format.

Significant properties of an e-print

The significant properties of a resource may be identified as the intellectual content that must be retained in subsequent migrations. The file format in which data is stored will, to an extent, determine the significant properties of the digital object and will contain instructions on how the content should appear, for example, the correct reading order of pages or the placement of diagrams. For research papers, different properties are likely to be important to different stakeholders: the author will primarily be concerned with communicating their ideas and the publisher (if the paper is a publisher version) is likely to be concerned with their layout style (CEDARS, 2002). Four properties may be identified as significant:

1. **Text** - Textual information is the most significant content type held in a research paper. The encoding method is likely to vary according to the method of creation – research papers created prior to the 1980s, such as those held by the Edinburgh Research Archive, are likely to consist of several scanned images. In contrast, recent deposits are likely to be provided in a born-digital ASCII or Unicode original, created in TeX (e.g. <http://www.era.lib.ed.ac.uk/handle/1842/209>), Framemaker 7.0 (<http://eprints.whiterose.ac.uk/archive/00000006/>), or Microsoft Word, and exported to PDF. Certain elements of the text may be corrupted or lost in the conversion process. These include word spaces, hyphens, font emphasis (bold, italics, underline), subscript/superscript, annotations provided by peer-reviewers or other users, and special characters (foreign and mathematical formulas) in unusual fonts (Gross, 2003).
2. **Images** – Many research papers contain images that are used to graphically represent specific types of information. In research papers held by institutional

repositories, images are used to depict cover sheets, graphs, tables, complex mathematical formula and photographs. The majority of these images are black and white, while a small number of papers contains colour. To avoid inadvertent loss of data, images should be exported to a lossless format, such as TIFF or PNG (Gross, 2003).

3. **Layout** – The layout of a research paper can often imply meaning that is significant and which, if not identified correctly, may be rendered incomprehensible. Conversion software to export content to a different file format is likely to handle multiple newspaper-style columns, however text boxes set off to the side of text or commentary text that runs alongside a paragraph may be incorrectly interpreted. Footnotes may also be misplaced if not appropriately tagged (Gross, 2003). Structural information may provide context to the organization of a document that cannot be easily identified without manual validation. PDF Tags are a recent development in the PDF 6.0 specifications that may be applied automatically by the software or, preferably, by the author¹ to identify specific sections of a document. For example, the identification of an abstract or quotations within a paper. A PDF file equipped with well-formed tags may be "reflowed" to fit different page or screen widths and, as a result, will be accessible by users with screen-readers (such as JAWS and Window Eyes), handheld devices, or other audio-visual units (Gross, 2003). Few authors are currently creating tagged PDF files – only one e-print was found in the sample that was identified as a 'tagged PDF'. This may be attributed to a number of reasons: it requires additional effort or it may not be supported in the word processing or desktop publishing software that they currently use (AccessIT, 2003).
4. **Descriptive Metadata** – Resource discovery metadata, created by the author or repository staff, will exist for every e-print within a repository. However, additional metadata may be encoded in the XMP (Extensible Metadata Platform) section of the PDF itself. XMP is a derivative of the Resource Description Framework (RDF) that enables the creation of basic descriptive (Title, Author, Subject, Keywords) and provenance (Creator, Producer, Creation Date, Modification Date) metadata. Descriptive metadata embedded in the e-print often provides details on the creation process and may in the absence of author-created metadata, provide useful information that can be copied to resource discovery metadata. However, XMP fields are often populated by computer software and, as a result, the value of this information varies significantly – the XMP section of many research papers is often incomplete or contains 'junk' values created by the PDF creation tool. For example, one paper held by Nottingham ePrints² list the Title as 'Microsoft Word - sp13076-macdonald1.doc' and Author as 'Administrator'. Additionally, several e-prints held by the White Rose consortium are intentionally encrypted by the author, which makes it difficult to extract XMP tags from the e-print.

The objective of digital preservation is to retain the four properties outlined above in a format that is sustainable in the long-term. However, it is unlikely that different users will require access to this information in its entirety. The researcher, who is accessing the institutional repository, is likely to be concerned that the dissemination version provides an accurate rendition of the text, images and layout of the authors' or published paper; the publisher will be concerned that the layout of their original has been retained; while the preservation service is likely to require information on all five properties, in order to construct a new dissemination version of the e-print.

Analysis of repository holdings

To ensure the preservation of digital holdings provided by institutional repositories, the AHDS performed an assessment of repository holdings to identify potential problems. A sample of 10 e-prints were downloaded from each repository and the file infrastructure analysed using JHOVE – an open source format identification and validation software developed by JSTOR

¹ Manual intervention is necessary to ensure the tagging process was performed correctly. Even seemingly small errors in document structure can render a file completely incomprehensible.

² <http://eprints.nottingham.ac.uk/archive/00000280/>

and Harvard University. The file structure of the PDF document is analysed and validated against Adobe's PDF specifications (Jhove, 2005). Two issues were identified:

1. **Outline Dictionary Missing** –PDF document often contain a document outline that allows the author to specify significant sections in the document (e.g. table of contents, beginning of chapter, index page) and navigate between them. An Outline dictionary defines the different levels of the document hierarchy. This may be compared to a contents page that allows the author to quickly identify the relevant section they require. In one document identified, the outline dictionary was absent and the navigation aids no longer functioned. In an analogy of paper documents it may be compared to the removal of page numbers, which would require the reader to manually search through the document to locate a particular section.
2. **Encrypted XMP metadata** – The PDF specification contains security settings that the author or repository staff may use to configure the allowed use of a document and ensure that it cannot be and has not been changed. These include security passwords necessary to open or modify the document, quality levels for printing, recording of changes to the document, permission setting to copy or extract content, commenting, form field completion, document access and encryption level. A small number of e-prints in the White Rose repository were identified that contain restrictive security settings that may prevent the export of intellectual content. As a result, JHOVE is unable to extract XMP (Title, Author, Creation and Producer) fields from the document, displaying a <may be encrypted> error message.

For authentication purposes, it is useful to identify missing or inaccessible components within a document before migration has taken place, to avoid blame for errors being attributed unnecessarily to the migration process. These issues are unlikely to prevent the long-term maintenance of the research papers, but they do indicate the need to identify potential problems with a document and correct errors before migratory action can be performed. It is preferable for the institutional repository to perform error correction before the papers are harvested by the AHDS. Research papers that have been made secure should be modified to remove encryption. The e-prints security settings should be recorded in preservation metadata and applied to the dissemination version, if the institutional repository requests it. Similarly, the Outline dictionary should be recreated by resaving the file.

Export of repository holdings to a suitable archival format

Format migration is considered to be a cost-effective method of allowing continued access to a resource. The primary goal of this process is to ensure that intellectual content of a resource is held in a file format that is accessible over time and which will limit the long-term costs of preservation. The archival format should retain intellectual content of the e-print, and store any information, such as structural metadata, that is embedded within the resource. Two approaches may be taken to preserving the resource:

1. The PDF may be migrated to another format that can accommodate the various components in a single file, or
2. The resource may be separated into two files – one that contains the intellectual content (text, images, and layout) of the resource and an XML record that contains relevant metadata about the resource.

To identify likely problems that affect the preservation workflow and the chosen archival format, an assessment of possible migration paths should be performed and suitable testing undertaken to identify potential problems that are likely to be encountered. Although the AHDS has not finalized the file format that will be used for preservation³, three formats may be considered – XML, RTF and PDF/A. Several software plug-ins/tools (Adobe Acrobat, Solid Converter PDF, et al) exist that export PDF documents to these and other common formats, however they do not always produce expected results. Possible problems include the removal of text and image data, modification to the page layout and loss of metadata. The likelihood

³ Further investigation of relevant file formats and metadata is necessary and will be the subject of a later work package.

that these problems will occur will vary according to the creation software (e.g. Microsoft Word), conversion software and file format chosen for export. To identify potential conversion issues, 10 e-prints were selected that contained unusual layout and the content was migrated to an alternative format using Adobe Acrobat.

Rich Text Format

RTF (Rich Text Format) is well-documented, may be read using common software tools (e.g. NotePad), and allows the recreation of the visual aspects of the document layout. However, it is designed primarily as an editing format and the specification does not specify a file structure sufficient to accurately describe the layout or context of a complex document. For e-prints that, in most cases, are created in Microsoft Word it is sufficient.

The Adobe Acrobat software exported text and images correctly and preserved the basic two-panel layout of the research paper. However, images contained within complex tables were not exported correctly to the RTF derivative. For example, certain e-prints held in the Edinburgh Research Archive contain departmental cover sheets (e.g. 'School of GeoSciences' coversheet in <http://hdl.handle.net/1842/823>) that are absent in the migrated version. The appearance was also altered - the font size of certain paragraphs varies between point 8, 10 and 12 in a single paragraph or font alignment was altered from 'left' to 'centre'. Formatting issues were also encountered when migrating research papers created in Microsoft Word and exported to PDF. Although the distributable PDF document provided an accurate rendition of the visual display of the research paper, subsequent conversion identified formatting, such as changes in formatting, font type and size, hidden in the original document that altered the layout of the RTF derivative.

XML

XML (Extensible Markup Language) is a second format that possesses properties that make it suitable for preservation. The format specifies the use of plain text (ASCII) and uses structural markup tags to identify components of a resource. For example, an XML document may identify structural components (e.g. markup that identifies the purpose of the text) or content elements (headings, paragraphs, etc.).

The intellectual components of an e-print are separated and exported in an appropriate format: text is output correctly in an XML file, accompanied by appropriate HTML layout tags and graphics are stored as PNG images. Although the software is able to infer meaning in the layout of a PDF that is not explicitly stated (e.g. identifying paragraphs, headings, etc.), it can be inaccurate if sufficient information is unavailable and will require manual correction. For example, the bullet point • is inferred as the text description for a bulleted list).

PDF/A

A third option is to use PDF/A – a refined form of Adobe PDF 1.4 that has been developed as a suitable format for long-term preservation of electronic page-based documents. Electronic extensions, such as dynamic (Javascript and executable files) or audiovisual content are disallowed and the specification does not contain any security settings. The standard (ISO/CD 19005-1) is currently in draft status.

The format offers a degree of certainty that the content will not be rendered inaccessible or obsolete in a later revision of the software. However, the PDF/A format should only be considered a holding format and cannot be described as a preservation solution. It does not resolve the underlying problem that the significant properties of an e-print cannot be exported to another file format in their entirety without manual intervention.

It is evident that no single file format is suitable to describe the significant properties found in e-prints. RDF metadata for example cannot be exported to Rich Text Format and must be held elsewhere. XML, by design, may hold any type of information, but a suitable application profile – a customized schema – must be developed to contain it. It is also evident that

migration tools cannot export intellectual content held in the PDF specification without manual intervention to correct errors and perform 'cleanup' tasks to modify vague or inaccurate descriptions. The formatting hidden in the MS Word and PDF versions of the research paper, in particular present complications when exporting intellectual content to a suitable preservation format.

Export of repository holdings to a suitable dissemination format

It is envisaged the preservation service, to be offered by the AHDS to Sherpa repositories, will be responsible for the creation of archival and dissemination versions of the PDF. Unlike the archive format, it is likely a replacement dissemination version will be created on several occasions during the lifetime of the repository. This may be the result of two situations:

1. The dissemination file format in use at the institutional repository (PDF1.1-1.6) is rendered obsolete.
2. The digital holdings of the institutional repository are lost in an unspecified event (e.g. fire)

The replacement dissemination version should be held in a file format that is usable by the user community of the institutional repository and is optimized for the format. E-prints held in a PDF-based format, for example, should be linearized⁴ to allow the first page to be displayed in the user's web browser before the entire file has been downloaded. Linearization reduces the amount of time between the user choosing to open a file and it being displayed on screen, reducing the likelihood that users will become frustrated or impatient. The file format of the replacement version will be decided at a later date and should be the subject of bi-annual review and revision during the lifetime of the project.

Recommendations

1. For authentication purposes, the AHDS should ensure that any metadata held about the digital object includes details of the file structure of the submitted e-print, including details of linearization and tagging.
2. The AHDS is responsible for the creation of new dissemination files in the event that existing versions of PDF or other formats are rendered obsolete. The AHDS should take into account the needs of the institutional repository's user community and the existence of suitable migration tools. Preservation staff should ensure that special features, identified in recommendation 1, are translated correctly for the replacement dissemination copy.
3. The AHDS should adopt appropriate automated software tools, such as Jhove or the NLNZ Metadata Extractor, to identify file formats submitted for preservation.
4. The AHDS should perform a detailed assessment of the risks associated with each file, in terms of the file format and the individual structure of each e-print. Appropriate action should be taken to convert e-prints into file formats that are suitable for long-term access.
5. The Institutional Repository should correct any file structure errors identified by the AHDS prior to the AHDS taking responsibility for the items. This may require the creation of a data dictionary, removal of security settings, or other remedial action.
6. To reduce the possible preservation risks, the Institutional Repository should encourage authors to deposit e-prints in file formats other than PDF.

⁴ Linearization is enabled by default in Adobe Acrobat, but may be disabled or absent in other software applications.

7. Proprietary file formats present the greater risk to the preservation of e-prints over the long-term. The AHDS should adopt an open-standards-based file format, taking into account the issues outlined in this document. To identify if XML is a suitable format in which to store the intellectual content of the e-print, the AHDS should generate XSL transformations to correct common errors (i.e. bullet points) and convert it into a suitable dissemination format.
8. The AHDS should consider develop detailed guidelines that indicate the level of preservation that they can provide for different file formats. The AHDS Deposit Formats list (<http://www.ahds.ac.uk/depositing/deposit-formats.htm>) may be used as a guideline.
9. The AHDS should review their preservation format policy on a regular basis (e.g. bi-annual) to identify new file formats that may be suitable.

References

- AccessIT. (2003). Is PDF accessible?. Retrieved on December 1, 2005 from: <http://www.washington.edu/accessit/articles?2>
- Adobe (2002) How To: Write an XML packet that conforms to XMP into a PDF file without using Acrobat or PDF Library APIs. Retrieved on December 1, 2005 from: <http://support.adobe.com/devsup/devsup.nsf/docs/52016.htm>
- Adobe (2005). Adobe PDF Security— Understanding and Using Security Features with Adobe Reader and Adobe Acrobat. Retrieved on December 1, 2005 from: <http://www.adobe.com/products/acrobat/pdfs/AdobePDFSecurityGuide.pdf>
- Cedars (2002). Cedars Guide to: Digital Preservation Strategies. Retrieved on December 13, 2005 from: <http://www.leeds.ac.uk/cedars/guideto/dpstrategies/dpstrategies.html>
- Gross (2003). Converting From PDF To XML & MS Word: Avoiding The Pitfalls. Retrieved on December 1, 2005 from: http://www.dclab.com/converting_from_pdf.asp
- Jantz, R & Giarlo, M.J. (2005). Digital Preservation: Architecture and Technology for Trusted Digital Repositories. D-Lib Magazine. Retrieved on December 1, 2005 from: <http://www.dlib.org/dlib/june05/jantz/06jantz.html>
- James, H. et al, (2003). Feasibility and Requirements Study on Preservation of E-Prints. Retrieved on December 1, 2005 from: http://www.jisc.ac.uk/uploaded_documents/e-prints_report_final.pdf
- Jhove (2005). PDF-hul Module. Retrieved on December 1, 2005 from: <http://hul.harvard.edu/jhove/pdf-hul.html>
- Knight, G. (2004). Selection Criteria for the Preservation of E-prints. Retrieved on December 13, 2005 from: http://www.sherpa.ac.uk/documents/D4-4_Preservation_Selection_Criteria.pdf
- Library of Congress (2005) PDF/A, PDF for Long-term Preservation. Retrieved on December 1, 2005 from: <http://www.digitalpreservation.gov/formats/fdd/fdd000125.shtml>
- OCLC-RLG. (2001). Attributes of a Trusted Digital Repository: Meeting the Needs of Research Resources. Retrieved on December 1, 2005 from: <http://www.rlg.org/en/pdfs/attributes01.pdf>
- W3C (2001)/ PDF Techniques for Web Content Accessibility Guidelines 1.0 and 2.0. Retrieved on December 1, 2005 from: <http://www.w3.org/WAI/GL/WCAG-PDF-TECHS-20010913/>