

Recommendations to ensure the long-term preservation of digital objects stored by institutional repositories

Document details

Project: SHERPA DP
Work Package: 6.5
Author: Gareth Knight
Version: Version 1.0
Document date: 28/02/07
Change history:

<i>Date</i>	<i>Version</i>	<i>Author</i>
20/06/2006	First draft	Gareth Knight
27/06/2006	Second draft	Gareth Knight
28/02/2007	First version	Gareth Knight

Contents

Introduction.....	2
Lifecycle of an e-print and its implications for preservation	2
Monitoring for obsolescence.....	2
Preservation Strategies	3
A migration strategy for the SHERPA DP project.....	3
Preparation of new content.....	8
Recommendations.....	8
References	10

Introduction

Digital preservation may be defined as “all of the actions required to maintain access to digital materials beyond the limits of media failure or technological change” (Beagrie & Jones, 2002). The primary threat to digital objects stored and distributed by institutional repositories is format obsolescence – the encoding method used for certain data types are rendered inaccessible by a number of software applications. Obsolescence may occur for many reasons – a software application ceases to function in a later release of an operating system and an alternative software product must be found; the software developer ceases to support the file format or earlier versions of the format, or the software developer enters liquidation or is bought by another company and the format is abandoned. A key goal of an OAIS-compliant repository is to ensure continued access to the intellectual content stored in a digital object through remedial action, performed on ingest or at a specific stage in the OAIS workflow. This work package outlines actions to be taken by the AHDS to ensure the long-term preservation of digital objects currently stored by institutional repositories. It also makes recommendations for changes at the institutional repository that will simplify the management process for digital objects that are submitted in the future.

Lifecycle of an e-print and its implications for preservation

The preservation of an e-print requires some understanding of its lifecycle prior to deposit with an institutional repository and the likely outcome if action is not taken. The lifecycle model developed by James et al (2004a), investigated in detail in the earlier Sherpa DP document, ‘A lifecycle model for an e-print in the institutional repository’, indicates there are seven distinct stages that an e-print may pass through, from its creation to eventual withdrawal. Although useful, it does not provide practical case studies to identify how the lifecycle may differ for an e-print that contains different types of significant properties. The point of creation is considered to be the most important stage in the e-print lifecycle (James et al, 2004a) that defines the content type that will be contained in the digital object and the encoding method. Two scenarios may be identified:

1. ‘Born Digital’ objects

A significant percentage of e-prints stored in institutional repositories are ‘born digital’ – academic research created and stored in an electronic format, such as PDF or Postscript. These digital objects consist primarily of textual information, often accompanied by images and descriptive metadata that is stored in a manner determined by the creator and/or depositor.

2. Scanned documents

A small, but growing number of objects in institutional repositories originally existed in a printed form and were digitized by IR staff or the author at a later date to allow widespread access and use. These documents may have been created by hand on paper, written on a typewriter, or created in an electronic form on a home computer and subsequently printed by the author. These documents typically consist of a number of pages scanned as raster images and subsequently merged into a single file.

The management process necessary to preserve the digital object will differ as a result of decisions made at the point of creation and subsequent deposit in the institutional repository. It will affect the e-print’s longevity, in terms of the ability to decode the digital object and the problems that may be encountered when migrating an e-print to a different preservation format.

Monitoring for obsolescence

The definition of a file format as active or obsolete is an intellectual exercise that requires some understanding of the digital object and its use in the computing market. Several institutions and projects are currently investigating methods to assess the likelihood of format obsolescence and offer services to identify repository holdings that are at risk (PRONOM &

Global Digital Format Registry). However, monitoring services are in early stages of development and are unlikely to be launched in the timescale allocated for the SHERPA DP project. In the absence of practical alternatives, the AHDS will perform the management processes necessary to monitor digital objects stored in the institutional repositories, assess format obsolescence and ensure that intellectual content remains accessible.

Preservation Strategies

A number of approaches may be taken to ensure continued access to an object. When discussing the preservation of digital resources, it is helpful to consider three approaches that may be taken to preserve content (James, 2004b):

1. Preservation of the bit stream (the basic sequence of binary data) that represents the information stored in a digital resource.
2. Preservation of the information content (words, images, etc.) stored as bits and defined by a logical data model, embodied in a media format.
3. Preservation of the experience (speed, layout, characters, etc.) of interacting with the information content.

The three preservation levels are organized in terms of difficulty and are unlikely to be performed sequentially. An institution may choose one or two preservation approaches, according to the type of digital content for which they are responsible. The first stage may be considered the simplest form of preservation, requiring only basic information necessary to describe the purpose of the digital object. The intellectual content is stored in its original form and only limited action is performed. This may be restricted to the creation of descriptive information regarding the purpose of the resource, or may include basic technical information. In itself this stage will not maintain the accessibility and usability of data over long periods of time because of hardware and software obsolescence issues.

To undertake the second preservation level, a detailed understanding of the significant characteristics of the digital object is needed, as well as the identification of the most effective method of migrating these properties to other file formats. For an e-print (or indeed, any type of research paper), the significant properties are likely to include the text (headings, body text, footnotes, end notes) and images present in the original paper that must be preserved.

The third preservation level is likely to be the most complex, requiring an understanding of the significant properties of a digital object, methods of extracting/migrating and validating the data, and some method to replicate the user experience of the original document. This will require replication of the original layout of the document, and may, as defined by the needs of the user, incorporate specific features of the software tool that was used to access the content (e.g. zoom features).

A migration strategy for the SHERPA DP project

The objective of digital preservation in the Sherpa DP project, is to ensure access to the intellectual content of an e-print, using software tools, workflow procedures, and file formats that are sustainable in the long-term. The construction of an Archival Information Package (AIP) – a set of information that has the “all the qualities needed for permanent, or indefinite, Long Term Preservation of a designated Information Object” (OAIS reference model, 4-33) – is considered a priority for the AHDS Preservation Service.

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, e.g. through performing repeat migration. If the intellectual content is stored in a format considered to be ill-suited for preservation, a normalized version may be created. 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.

The migration strategy undertaken by the AHDS for the SHERPA DP project will encapsulate the first and second stages, outlined above. The first level – preservation of the bit stream, as provided by the institutional repository – will be undertaken for *all* digital objects. The digital object will be stored as-is in the digital repository, technical metadata will be created, and appropriate backup procedures will be followed. The second level – preservation of the information– will be applied in circumstances when the AHDS is able to gain full control over the digital object, in order to perform migration.

Although it is preferential for the preservation service to perform migratory action for all digital objects, the complexity of certain file formats and absence of appropriate software tools make them difficult to migrate without some content loss. The file formats stored by institutional repositories participating in the Sherpa DP project are restricted to common formats that may be produced and viewed without significant effort (Knight, 2005; James et-al, 2004). These are represented by a mix of well-documented, proprietary formats (PDF, JPEG) and open formats (HTML, ASCII text). Table 1 indicates the file formats stored in institutional repositories and the level of preservation that the AHDS preservation service will provide.

File Format	Preservation Level
'Born digital' Adobe PDF	Bit preservation only
'Scanned' Adobe PDF	Bit and information content preservation
Microsoft Word	Bit and information content preservation
ASCII text	Bit and information content preservation
Unicode text	Bit and information content preservation
Postscript	Bit and information content preservation
DVI	Bit and information content preservation
TIFF	Bit and information content preservation
JPEG	Bit and information content preservation
GIF	Bit and information content preservation
HTML	Bit and information content preservation
Corel Draw	Bit and information content preservation
Others ¹	Bit preservation only

Table 1: Preservation level for different file formats

The preservation level will vary according to the complexity of the digital object and the availability of appropriate software to perform migratory action. It should also be noted that table 1 indicates the ability to provide a specified level of preservation for specific file formats. However, it does not indicate that preservation action is necessary at the current time. Many of the listed file formats are well documented and there are no noted factors that would cause them to be rendered obsolete in the near future (Library of Congress, 2006). The file formats listed in table 1 may be separated into two types – simple and complex objects.

1. Simple Objects

Simple objects primarily consist of a single type of intellectual content, in some cases accompanied by metadata. These may be considered the easiest to manage, as the content may be exported to an equivalent format without too many issues. For example, a set of GIF image may be batch exported to TIFF, without significant difficulty. The majority of these formats are also well-documented and supported to some degree by many different software applications. As the basis for other file formats, it is unlikely that ASCII or Unicode text will be rendered unreadable in the near future. The majority of still image formats are also well documented and unencumbered by patent or other

¹ 'Other' unknown file formats that may be accepted by SHERPA DP partners. The AHDS will provide a minimum of bit preservation and may offer preservation of the information content, if the data is stored in a suitable format.

restrictions². Figure 1 lists the simple objects distributed by SHERPA DP partners and the most appropriate preservation format.

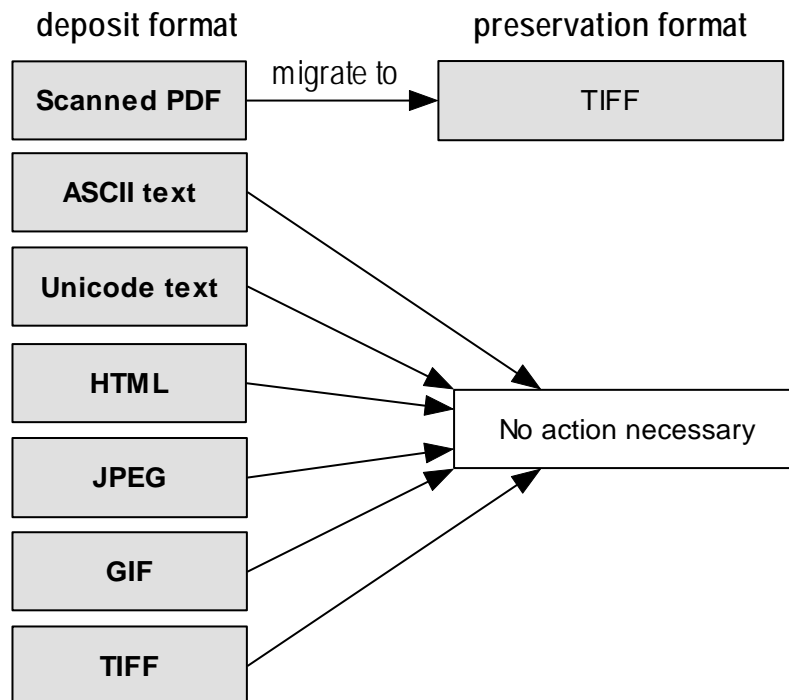


Figure 1: Simple objects received and distributed by institutional repositories and the most appropriate preservation format

No conversion is considered necessary for the majority of simple digital objects, specifically those stored as ASCII text, Unicode text, HTML, JPEG, GIF, or TIFF. An exception is scanned documents. These consist of multiple raster images that reproduce the exact appearance of each page, encapsulated in a PDF container. To avoid unexpected issues at a later date, each page should be exported to a TIFF image and given an identifier that indicates the source object and the page number. For example, the preservation output of <http://www.era.lib.ed.ac.uk/handle/1842/440> may be era_440_001.tiff, era_440_002.tiff, era_440_002.tiff, etc. A replacement DIP (Dissemination Information Package) may be created by converting the TIFF images into an appropriate distribution format³, sequentially ordered and exported to an appropriate packaging format⁴. Although the submission of uncompressed image bitstreams is preferred (e.g. TIFF_UNC_EXIF, TIFF_UNC), lossy compressed images, such as JFIF, JPEG_EXIF (EXchangeable Image File format for digital still cameras) and JPEG 2000 Part 1, Core Coding, Lossy Compression (J2K_C_LSY) are acceptable⁵.

² A possible exception is the JPEG format. Forgent Networks brought patent infringement suits against 31 software and device-manufacturing companies in 2004 (Forgent, 2006). However, it is expected the patent will expire in 2006 (Library of Congress, 2006).

³ TIFF images may be used as a distribution format. However, it is likely the institutional repository will require data to be stored in a compressed file format, such as lossy JPEG2000, that produce a smaller file size and, as a result, take less time to download.

⁴ The Sherpa DP OAIS model, section 3.4 provides further information on the requirements for a Dissemination Information Package (DIP).

⁵ In the context of institutional repositories, it is impractical to insist on lossless formats that are particularly large in file size. In the proposed infrastructure for the SHERPA DP project, the AHDS preservation service will harvest the same data that is made available to the institutional repository's user community. Although certain lossless formats are appropriate for preservation, researchers, who are the primary audience for the IR, are likely to consider small file size as their primary concern, rather than the high quality.

2. Complex digital objects

Complex digital objects may contain many different types of embedded information. Common formats, such as the Portable Document Format (PDF) and Microsoft Word formats may encapsulate text, raster images, vector images, sound, video and metadata. Complex digital objects are likely to be 'born digital' and, as a result, may be considered to be at most risk - an equivalent printed copy is less likely to exist, in comparison to scanned objects, and the digital object is likely to contain different types of content information that may have different lifecycles. It is possible to envisage a time when a single component, e.g. embedded vector objects, are obsolete and cannot be rendered, but the container format itself remains accessible.

Two methods are available to migrate complex objects: 1) the export of each e-print component to an appropriate format (e.g. text may be exported to XML markup, images exported to TIFF, metadata saved as Dublin Core, etc.) or 2) migrate all content to a single format that is capable of storing the significant properties (e.g. a Microsoft Word document may be migrated to Rich Text Format or Open Document Format). The first method is preferred, to allow the AHDS to manage the intellectual content in its simplest form. However, such an approach has more complex issues of metadata management and object linking that must be catered for. Born digital objects also contain properties, such as fonts, vector designs, and text encoded in different character sets that may be difficult to export without corruption⁶. Therefore, the second approach is the pragmatic approach at the current time⁷.

The preservation strategy for complex objects should vary, according to whether and what other content types are encapsulated in the digital object. Figure 2 summarises the formats that institutional repositories are likely to provide and the conversions that may be necessary.

Figure 2: Complex digital objects received and distributed by institutional repositories and the chosen preservation format

Complex digital objects should, in most circumstances, be converted to an appropriate preservation format on ingest. If appropriate software tools are made available, the

⁶ An investigation of tools, performed by the author of this work package, found that only a small number of software tools are able to open PDF documents and export the content to another file format, and currently none of these are able to do so without some data loss or corruption. The Linux-based Kword, for example, successfully displays PDF documents, but corrupts vector images when exporting to KWord format and altered layout when exporting to the XML-based OpenDocument format. Adobe Acrobat 7.0 Professional also caused problems when exporting to the Adobe XML format, missing bitmap images that were in the original document. Innate problems of exporting PDFs to other common file formats are also apparent, such as inability to export font information, which would result in scientific formulas being rendered unreadable to many researchers.

⁷ The preservation policy will be reviewed on a bi-annual basis to identify software conversion tools and suitable preservation formats.

preservation service should perform actions necessary to ensure the complex object formats are complete and reduce the possible preservation risks through migration to a suitable file format. An XML-based format is considered to be the most effective method of preserving digital objects, due to its human and machine readable design (Potter, 2002). OpenDocument is considered to be the most appropriate format for preservation of documents and conversion tools are widely available.

The majority of digital objects stored and distributed by SHERPA DP project partners are encoded in the PDF format. PDF is a pragmatic format that fulfills certain user requirements. However, as noted in the investigation of file formats in use by SHERPA DP repositories, it is ill-suited to preservation (OASIS Open, 2006). If a file is in a format considered to be less than optimal for digital preservation a normalized version of the file may be created. The normalized file contains the same content in a more preservation-worthy format. Normalized versions may not be equivalent to originals in appearance or functionality. For example, a PDF file can be normalized into a set of page-image TIFFs. In this case the appearance of the content is retained but functionality such as actionable hyperlinks is lost. Normalization is performed on Ingest and normalized versions are stored in the AIP.

Preservation projects are taking different approaches to the long-term maintenance of PDF objects: the UK Data Archive stores the objects as-is and do not guarantee any form of migration; the Florida DAITSS project (Dark Archive in the Sunshine State) suggests a possible approach is to normalise PDF files into a set of page-image TIFFs. By performing such an action, they note the content appearance will be retained, but specific functionality, such as actionable hyperlinks will be lost (Florida Center for Library Automation, 2006).

In the absence of appropriate XML-based alternatives, the AHDS may take action necessary to ensure that PDF objects are complete and reduce the possible preservation risks. The most viable alternative is to convert PDF-based, 'born digital' objects to a stable version of the format, PDF Archive. For the majority of e-prints, stored in the PDF v1.1-1.4 format, compliance checks will be performed to validate the object in comparison to the PDF/A-1b specification⁸. Possible compliance issues for PDF-based objects may be caused by reference to external components, most notably fonts, and the absence of XMP metadata⁹. Remedial action should be taken to correct any errors, when appropriate software tools exist.

A small number of items stored by SHERPA DP project partners are stored in other format types that are simpler to manage. OpenDocument is considered to be the most appropriate format for preservation of documents and conversion tools are widely available. Several software applications, notably Open Office, are able to import Microsoft Word and Rich Text Formats and export the intellectual content to the OpenDocument format without significant loss. Similarly, DocVert is a web service that may be used to convert multiple Microsoft Word files to the OpenDocument specification. However, a side

⁸ The PDF/A standard is a format specification intended for "*the archival storage of electronic documents*" (Aiiim, 2006). The standard is a derivative of the PDF Reference v1.4 specification⁸ that imposes limitations on the content type that may be encapsulated in an object, restricting it to text, images and metadata. Embedded files that may have dependencies outside the Adobe software, such as interactive or audiovisual content are not allowed. Fonts must also be embedded in the document itself, to avoid the possibility that the character encoding will be lost, rendering the text unreadable. The PDF/A reference specification identifies two levels of compliance: PDF/A-1a a strict level of compliance, and the less stringent requirements defined by PDF/A-1b. Notably, PDF/A-1b compliance does not require the existence of Unicode character mapping or appropriate text descriptions for non-text components. New versions of PDF/A will be developed to conform to PDF reference versions later than 1.4.

⁹ A compliance check of a small sample of the digital holdings of several Sherpa DP partners performed using the Adobe Acrobat 7 software indicated that the majority of PDFs reference external fonts located on the authors' computer, most notably Times New Roman, and a smaller number did not contain any metadata.

effect of this process is that certain embedded objects, such as Smart Draw vector images, are no longer editable. Complex vector objects, such as Corel Draw, should be converted to the XML-based SVG format. Postscript is a well-documented format intended for printing. As a format intended for layout and printing, the Postscript standard is unlikely to change and may be interpreted, if not easily read, by text viewers. As a result, no format conversion is necessary.

Preparation of new content

The options available to preserve existing content are limited. However, there is an opportunity for repository implementers to ensure that new content is simpler to manage, for a third party or in-house preservation service. This will require some changes to the activities performed by IR staff and authors.

To understand how e-prints may be preserved, it is useful to have an understanding of the lifecycle of an e-print and the decisions that take place at each stage. The e-print lifecycle model, developed by James et al (2004) and subsequently adapted for the Sherpa DP disaggregated model in Work Package 2.13, identifies key events, from creation of a first draft by an author, to technical obsolescence and possible withdrawal from the institutional repository. In particular, the model notes that decisions made during the creation and submission of an e-print into an e-print archive are likely to have long-term consequences for the preservation of a digital object, in terms of costs and preservation strategy. At creation, the author makes choices about which software package to use to write the e-print, the type of content (text, images, etc.) it will contain, and the file format in which to store it. Subsequent decisions are made at the submission stage, such as the file format in which the e-print must be deposited, the extent and type of metadata provided with the digital object, and the allowed use (as defined in the deposit licence). Although the institutional repository cannot dictate the creation of an e-print, it may establish specific requirements with which an author should comply when depositing their e-print.

The requirements established for submission vary between Sherpa DP partners – certain repositories establish the right to distribute the e-print, others do not; certain repositories indicate that e-prints must be deposited as a PDF format only, others establish a less restrictive list of 5+ common file formats that may be deposited and repository staff perform migratory action. When contemplating the long-term management and preservation of e-prints, it is useful to reconsider these policies and modify them if necessary. In regards to the subject of this report, it is useful to consider two issues that may affect format obsolescence and the ease with which migration processes may be performed.

Recommendations

1. Deposit Formats

The investigation of file formats in Work Package 6.2 identified several file formats accepted by Sherpa DP partner repositories at the submission stage¹⁰. For the purpose of consistency and to limit the amount of work necessary, deposit guidance typically identifies PDF as the preferred deposit file format, but accepts other formats that may be convenient for the depositor. These formats may be migrated to PDF by repository staff or made available as-is¹¹. Although the PDF format is widely accepted as a convenient

¹⁰ PDF was identified as the preferred file format for deposit followed, in no particular order or preference, by Microsoft Word, postscript, ASCII text, web site formats (HTML, GIF, JPEG, etc.) and other non-specified formats.

¹¹ The decision to migrate or make the deposited object available in its original format differs between institutional repositories and, according to discussions with Sherpa DP partners is based upon a range of factors, such as the availability of migration software, the number of objects that require conversion, and the likelihood that the conversion will be successful.

distribution format, it remains difficult to export intellectual content to other formats without some form of content loss.

Recommendation 1: repository management should revise policy on deposit formats, to encourage the deposit of particular file formats that can be more easily managed over time.

This may require the IR to diversify the number of formats it accepts, if it currently supports the Portable Document Format (PDF) only (e.g. Nottingham ePrints), or to refine its list of accepted formats, if it accepts several formats. Recommended formats for deposit include OpenDocument Text, Postscript, Rich Text Format (RTF) and Microsoft Word. To educate potential depositors, it may be useful to organize deposit formats into 'preferred', 'acceptable' and 'problematic' categories¹².

2. Document styles

The layout and formatting of e-prints deposited with institutional repositories is likely to be defined by the creator prior to submission. The author may choose to use the default design template provided by the editing software or may refine the template according to personal preference, or to accommodate particular content (e.g. images). In many cases, the creator will also use document styles to provide aesthetic consistency, indicating how key sections of the document should appear. For example, a title heading may be formatted in the Arial font, size 20, a section heading may be formatted to Arial, size 18, and a smaller heading may be formatted to Arial, size 16. Although the primary purpose of such formatting is aesthetic, these styles often indicate the content type. For example, <heading1> is often used to indicate the paper title in Microsoft Word.

Recommendation 2: institutional repositories should develop a standard template and educate authors on the benefits of its use.

The consistent use of styles and basic character formatting will improve the likelihood that contextual information will be maintained when migrating to an XML-based format¹³ (Dawson, 2006), and may be useful for value added features, such as the automatic extraction of resource discovery metadata. Dawson (2006) has published a document template for electronic documents that may serve as a useful starting point on which to base a standardized repository template.

3. Distribution formats

Institutional repositories establish a select number of file formats intended for distribution. Typically, PDF is the chosen format for distribution, possibly accompanied by other formats, such as Corel Draw vector designs, that cannot be converted without specialist software or loss of specific functionality. The harvesting method established for the Sherpa DP project uses these object as source data to create replacement Dissemination Information Packages (DIP) as necessary. However, as noted above, distribution formats are often ill suited to the requirements of preservation.

Recommendation 3: institutional repositories should make e-prints available in submission AND dissemination formats, to simplify the preservation process and allow the option for different migration paths.

The influence of the Sherpa consortium may be used to encourage repository developers to implement necessary functionality, if this is not possible for technical reasons (e.g. repository software is capable of distributing one object per item-level record only). As

¹² A possible template for a list of deposit formats may be found on the AHDS web site <http://ahds.ac.uk/depositing/deposit-formats.htm>.

¹³ A macro may be used to process styles created in Microsoft Word and convert them to the equivalent XHTML markup. For example, Heading1 will become <h1>, heading2 becomes <h2>.

part of the process, IR staff are encouraged to perform migration of deposited e-prints into a PDF object, rather than request the author or depositor perform the process.

Conclusion

The standardization of digital content to an open, well-documented format, or limited range of formats is the preferred migration approach for the project. However, the complexity of the various file formats makes it difficult to perform such migration without some loss. The migration methods available to preserve existing content are, unfortunately, limited by the availability of software tools. As a result, a pragmatic approach is necessary – the AHDS will perform actions necessary to allow continued access to digital objects, as defined by preservation level 1 (table 1) for difficult to handle formats. Other formats that, at the current time, do not require any conversion will be stored in their delivery formats. In the event that these formats are rendered obsolete or are restricted in some manner (e.g. by patent restrictions, similar to the recent GIF LZW algorithm issue), the AHDS preservation service will take action necessary to export content. It is evident, however, that changes will be necessary to manage the migration process in a systematic manner. To improve the likelihood that the digital objects in an institutional repository will be preserved, fundamental changes must be made to repository workflow and policies, as well as to the repository software itself.

References

- Beagrie, N. & Jones, M. (2002). Digital Preservation Coalition Handbook. Retrieved on July 11, 2006 from:
<http://www.dpconline.org/graphics/intro/definitions.html>
- Cedars (2002). Cedars Guide to: Digital Preservation Strategies. Retrieved on June 26, 2006 from: <http://www.leeds.ac.uk/cedars/guideto/dpstrategies/dpstrategies.html>
- Dawson (2006). A modular methodology for converting large, complex books into usable, accessible and standards-compliant ebooks. Available shortly from <http://ahds.ac.uk>
- DocVert (2006). DocVert Home Page. Retrieved on August 2, 2006 from:
<http://holloway.co.nz/docvert/>
- Florida Center for Library Automation (2006). DAITSS Overview. Retrieved on June 26, 2006 from:
<http://www.fcla.edu/digitalArchive/pdfs/DAITSS.pdf>
- Forgent (2006). Intellectual Property. '672 patent. Retrieved on June 26, 2006 from:
<http://www.forgent.com/ip/672patent.shtml>
- Gross (2003). Converting From PDF To XML & MS Word: Avoiding The Pitfalls. Retrieved on June 26, 2006 from:
http://www.dclab.com/converting_from_pdf.asp
- Library of Congress (2006): JPEG Lossy (DCT) Compression Encoding. Retrieved on June 26, 2006 from: <http://www.digitalpreservation.gov/formats/fdd/fdd000017.shtml>
- James, H. et al, (2004a). Feasibility and Requirements Study on Preservation of E-Prints. Retrieved on June 26, 2006 from:
http://www.jisc.ac.uk/uploaded_documents/e-prints_report_final.pdf
- James, H. (2004b). Collections Preservation Policy. Retrieved on June 26, 2006 from:
<http://ahds.ac.uk/documents/colls-policy-preservation-v1.pdf>
- Knight, G. (2005). An investigation of file formats in use by SHERPA DP repositories. Internal project report.

Library of Congress (2005) PDF/A, PDF for Long-term Preservation. Retrieved on June 26, 2006 from: <http://www.digitalpreservation.gov/formats/fdd/fdd000125.shtml>

National Library of Australia (2002). Digital Preservation Policy. Retrieved on June 28, 2006 from: <http://www.nla.gov.au/policy/digpres.html>

OASIS Open (2006). OASIS Open Document Format for Office Applications (OpenDocument). Retrieved on June 26, 2006 from:
http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=office

President and Fellows of Harvard College (2006). GDFR – Global Digital Format Registry (2006). Retrieved on August 2, 2006 from:
<http://hul.harvard.edu/gdfr/>

The National Archives (2006). PRONOM – The Technical Registry. Retrieved on August 2, 2006 from: <http://www.nationalarchives.gov.uk/pronom/>