Australian Government
National Archives of Australia

Dissecting the Digital Preservation Software Platform

Version 1.0

RKS: 2009/4026

the archives
naa.gov.au

Document Change Record

Version	Changed By	Description of Changes	Change Date
0.1	Allan Cunliffe	Created.	August 2010
0.2	Allan Cunliffe	Edits following initial review.	November 2010
0.3	Allan Cunliffe	Edits following internal review.	November 2010
0.4	Allan Cunliffe	Edits following second internal review.	November 2010
0.5	Allan Cunliffe	Edits following third internal review.	December 2010
1.0	Allan Cunliffe	Minor edits. Marked as final.	February 2011

Related Documentation

Title	Author	Date	URL
An Approach to the Preservation of Digital Records	Helen Heslop Simon Davis Andrew Wilson	December 2002	http://www.naa.gov.au/images/an-approach-green-paper_tcm2-888.pdf
The Benefit of Experience: the first four years of digital archiving at the National Archives of Australia	Michael Carden	August 2010	http://michaelcarden.net/blog/wp-content/uploads/2010/08/michael-carden-conference-paper-final.pdf
Checksum Checker User Manual v1.3	Allan Cunliffe	August 2010	http://checksumchecker.sourceforge.net/documentation.php
Digital Preservation Recorder User Manual v1.2	Allan Cunliffe	August 2010	http://dpr.sourceforge.net/documentation.php
Manifest Maker User Manual v2.2	Allan Cunliffe	September 2010	http://manifestmaker.sourceforge.net/documentation.php
Xena Help	Allan Cunliffe	August 2010	http://xena.sourceforge.net/documentation.php

Table of Contents

1 Introduction

Digital records present many preservation challenges. They are at risk of being lost due to the rapid pace of development in computer hardware, operating systems and application software, coupled with the short effective life of most physical storage media.

The National Archives of Australia (National Archives) has developed a digital preservation methodology, supported by software, to preserve significant digital records as national archives.

This paper details the methods employed by the National Archives to ensure the effective preservation of digital records. It covers our approach to preserving digital records, how we develop our software, what the software does to preserve digital records and how we maintain the integrity of the digital records in our digital archive.

1.1 Background

Over many years, the National Archives considered different ways of preserving significant Australian Government digital records as national archives. In 2000 we commenced research which resulted in the National Archives developing a digital preservation methodology, and building its own digital preservation software and digital archive.

Our suite of software is made up of the following applications:

- Manifest Maker[1] – supports the transfer of digital records from agencies to the National Archives. It produces a list of all digital records being transferred, their checksum, and the relationship between the digital records and their related item numbers in RecordSearch, our archive management system
- Xena[2] (XML Electronic Normalising for Archives) – determines the file format of a digital record and converts it to an appropriate preservation file format based on open standards. Xena can also be used to view and/or export Xena files
- Digital Preservation Recorder[3] (DPR) – manages the digital preservation workflow, recording audit information about each transfer in to the digital archive. DPR uses Xena to perform digital record conversions
- Checksum Checker[4] – monitors the contents of the National Archives' digital archive for data loss or corruption.

All of our software is developed using an open source methodology and is licensed under a free and open source licence (the GNU General Public License, or GPL).

[1] http://manifestmaker.sourceforge.net/

[2] http://xena.sourceforge.net/

[3] http://dpr.sourceforge.net/

[4] http://checksumchecker.sourceforge.net/

1.2 Audience

This document is intended for those interested in digital preservation, including:

- staff of the National Archives
- other archives or collecting institutions.

1.3 Scope

The following sections detail what is treated as being in scope and out of scope of this document.

1.3.1 In scope

The following are considered to be within the scope of this document:

- a description of the National Archives' approach to digital preservation
- an overview of the functions of our digital preservation software
- how we process digital records, including a description of our selected preservation formats.

1.3.2 Out of scope

The following are considered as being out of scope for the purposes of this document:

- all activities relating to the negotiation and processing of transfers prior to the creation of a transfer manifest file
- digital access strategies, including the provision of access to digital records stored in the digital archive
- complex technical information about our software, network or hardware. This includes software architecture and network configuration
- discussion of specific characteristics of the original record that are preserved by our software. This will be covered in a subsequent document.

2 Methodology

The aim of our digital preservation methodology is to ensure that we can read and provide access to digital records in the future.

The work of preserving digital records is done by a small team of professionals. Our solution attempts to maximise the safety and integrity of significant digital records with minimal resources. We achieve a balance between these conflicting goals by:

- converting digital records into openly-specified preservation file formats
- using an open source development methodology and licensing our software under the GPL. This enables us to build upon the efforts of other open source projects, so we can achieve our goals more quickly and with fewer resources
- processing files of the same file format in the same way. For example, all Microsoft Word documents are converted to Open Document Format. This approach is predictable and removes the need to make separate decisions for each new transfer we receive
- processing digital record transfers as soon as a transfer is received to give us the best chance to convert the records to a preservation format. If we encounter any issues, such as file corruption or damaged transfer media, we can address them as soon as possible
- always keeping an exact copy of the original files as they were transferred
- automating the digital preservation process as much as possible.

3 Preservation formats

The basis of the National Archives' digital preservation strategy is the conversion of non-open file formats to open file formats which have a greater potential lifespan.

We select preservation file formats based on open standards, which:

- have full specifications that are publicly documented
- are interoperable with a range of software applications from multiple vendors
- are community developed and not the work of a single entity
- are not affected by changes in commercial property rights over software in the marketplace (that is, no licence or patent restrictions).

One of the challenges of digital preservation is having the software to access the content of digital records in the future. Using preservation file formats that are based on open standards reduces the chance that software required to read the files is not available. However, if the software is unavailable, the file format specification can be used as a basis for recreating the necessary software.

Before we process any digital records, we determine which file formats will be converted to an equivalent preservation file format. The significant characteristics preserved by our process are determined by a combination of the:

- preservation file format selected
- way our software converts the original file format to the preservation format.

As some characteristics of a record may be lost when it is converted to another file format, we want to make sure we select the best available preservation format we can – one that preserves the most significant characteristics of the original record.

The digital preservation file formats we have selected are listed in the following table. The preferred file formats are the targets for file format conversions. Both the preferred and acceptable preservation file formats are based on open standards and any digital records received in these formats are preserved as they are.

Table 1: Preservation File Formats

Format Category	**Preferred open file format**[5]	**Files in these formats are converted to our preferred open file format**	**Acceptable open file format**[6]
Archive	An index of the contents is created as XML. The content of the archive is converted according to the appropriate preservation file format.	• Compressed archives (gzip, bzip2, war, zip) • Uncompressed archives (jar, tar, zip)	
Audio	Free Lossless Audio Codec (flac).	• Audio Interchange File Format (aiff) • Broadcast Wave File (bwf) • MPEG-2 audio layer 3 (mp3) • Speex (spx) • Vorbis (ogg, oga) • Wave Audio File (wav)	
Computer aided design	Not yet decided.	• Drawing (dws, dwt, dwg) • Design Web Format (dwf)	• Drawing Exchange Format (dxf)
Email	XML and XSL files are created for each email. Any attachments are converted according to the appropriate preservation file format.	• Mailbox (mbx, mbox) • Outlook Mail Message (msg) • Outlook Personal Information Store (pst)	• Email (eml)

[5] Preferred file formats are based on open standards and the targets for file format conversions.

[6] Acceptable preservation file formats are those based on open standards. Any digital records received in these formats are preserved as they are.

Format Category	Preferred open file format	Files in these formats are converted to our preferred open file format	Acceptable open file format
Geospatial data	Not yet decided.	• Spatial Data File (sdf)	• Geography Markup Language (gml)
Image - raster	Portable Network Graphics (png).	• Bitmap (bmp, gif, pcx, pnm, ras, xbm) • Photoshop (psd) • Tagged Image File Format (tiff) • Windows Cursor (cur)	• Open Document Graphics (odg) • Joint Photographic Experts Group (jpeg) • Portable Document Format (pdf)
Image - vector	Not yet implemented.	• Adobe Illustrator (ai) • Encapsulated PostScript (eps)	• Scalable Vector Graphics (svg)
Office documents	Open Document Format (odf).	• Excel (xls, xlsx, xlt) • PowerPoint (pot, pps, ppt, pptx) • Rich Text Format (rtf) • Symbolic Link (slk) • StarOffice (sdd, sdc, sdw, sxc, sxi, sxw) • Word (doc, docx, dot) • Word Perfect (wpd)	• Open Document XML (fodt) • OpenOffice.org XML (stw, stc, std, sti, sxg, sxm)
Project	XML	• Project (mpp)	

Format Category	**Preferred open file format**	**Files in these formats are converted to our preferred open file format**	**Acceptable open file format**
Plain text	Plain text in Unicode or ASCII.		• Style sheets (css, xsl/xslt) • Database tables, such as comma and tab-separated files (csv, tsv) • Scripting files (such as Python, Javascript, Perl, PHP) • Structured Query Language (SQL)
Video – video stream	Not yet decided. Video files are currently preserved as is.	• Flash Movie file (swf) • Motion JPEG/JPEG 2000 • RealVideo (rv, rmvb) • Windows Media Video (wmv)	• Theora (ogg, ogv) • “Raw" video
Video - container	Not yet decided.	• Audio Video Interleave (avi) • Advanced Systems Format (asf) • Flash Video File (flv) • MPEG video (mpeg-2, mpeg-4) • QuickTime Movie (mov) • RealMedia (rm)	• Ogg (ogv) • Matroska (mkv)

Format Category	Preferred open file format	Files in these formats are converted to our preferred open file format	Acceptable open file format
Video – audio stream	Free Lossless Audio Codec (flac).	• Advanced Audio Coding (aac) • MPEG-2 audio layer 3 (mp3) • RealMedia audio (ra, ram)	
Website	XHTML	• HTML (htm, html) • Active Server Page (asp, aspx)	
Website archive	Web ARChive (warc).	• MIME HTML (mht)	• ARC file format (arc)

4 Open source development

To develop the suite of digital preservation software the National Archives adopted an open source development methodology. This provides a level of transparency to our digital preservation processes and also allows others to freely use or modify the software to suit their circumstances.

Wherever possible, we reduce our development effort by making use of software and software libraries developed by others. Software libraries extend the capabilities of our software by providing additional functions, such as file detection or manipulation. Software libraries can also provide a way for our software to access the functions of other software, such as LibreOffice or OpenOffice.org. **Appendix A** lists the software libraries used by version 5 of Xena.

Our open source development methodology covers:

- software licensing
- external contributions
- version control
- programming standards
- change management
- release management
- testing (including unit, integration and regression testing).

4.1 Supporting new file formats

The Xena software architecture is based on the use of plugins. Plugins are a set of software components that add specific capabilities to a larger software application.

A Xena plugin consists of one or more components, each having a specific role in the conversion process. These include file format detection, file conversion, and creation of the Xena XML file that is eventually stored in the digital archive.

Xena uses different plugins to deal with various categories of file types, including:

- audio
- archive
- email
- image.

When we decide to support a new file format, we will either extend an existing plugin or create a new plugin. For example, if the new, unsupported, file format:

- is an audio format, we would extend the existing audio plugin
- does not logically belong to an existing plugin, we will create a new plugin.

5 Processing

The main steps in the National Archives' digital preservation process are:

1. Transfer documentation. This includes a manifest of all digital records to be transferred to the digital archive, including their checksum. The checksum is created using Manifest Maker (or a similar tool). Descriptive information about the records is loaded into RecordSearch[7].
2. DPR processing. Records are checked for viruses and integrity and, where necessary, converted to preservation file formats.
3. Storage. Records are stored in the digital archive and continually monitored for data loss or corruption.

The following diagram describes how our software is employed at each stage of the digital preservation process.

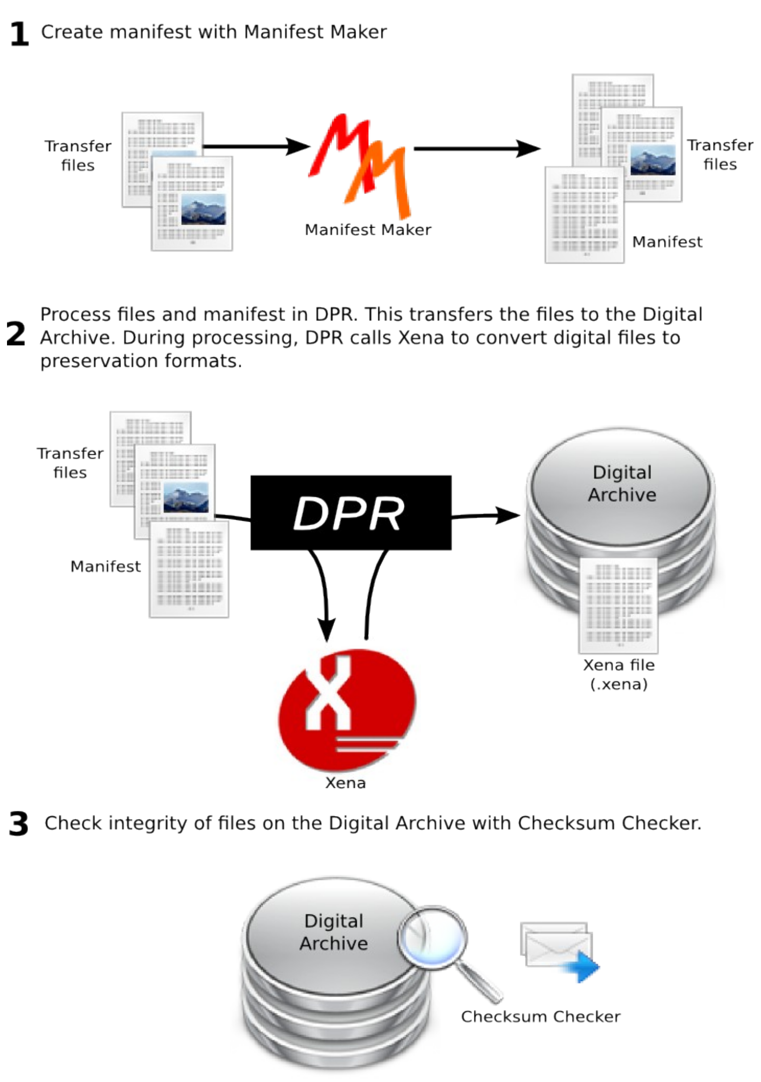
1 Create manifest with Manifest Maker
Transfer files
Manifest Maker
Transfer files
Manifest
2 Process files and manifest in DPR. This transfers the files to the Digital Archive. During processing, DPR calls Xena to convert digital files to preservation formats.
Transfer files
Manifest
DPR
Digital Archive
Xena file (.xena)
Xena
3 Check integrity of files on the Digital Archive with Checksum Checker.
Digital Archive
Checksum Checker

Diagram 1: Digital Preservation Process

[7] All records must meet the conditions of the National Archives' transfer policy and be properly controlled and described – a requirement for both paper and digital records.

There are three stages to DPR processing: Quarantine, Preservation and Storage. During each stage DPR performs a number of functions to ensure that digital records are preserved. All of these functions are fully integrated within the DPR workflow.

Table 2: Stages of DPR Processing

Stage of DPR processing	**Key Functions**[8]
Quarantine	• Checksum checks • Virus checks • Capture of processing metadata
Preservation	• Detection of file formats and conversion to preservation file formats (using Xena) • Automated quality assurance • Manual quality assurance • Checksum creation for newly created Xena files • Checksum checks • Capture of processing metadata
Digital Repository	• Copying digital records to the digital archive • Checksum checks • Capture of processing metadata • Reprocessing

5.1 Checksum checking

From the point of manifest creation, all digital records are assigned a checksum. The checksum uniquely identifies each digital record. Any change in the checksum indicates that the digital record has been corrupted or changed in some way.

Checksum checking is performed throughout all stages of DPR processing.

[8] Unless specified as manual, these functions are fully automated.

5.2 Virus checking

Virus checks are integrated in to the Quarantine stage of DPR processing to ensure that the safety of our digital records is not put at risk by any viruses that may be present on incoming records.

To ensure that virus checks are effective, we check all incoming records for viruses and update our virus signature definitions each day.

The process we follow for checking for viruses is:

1. Check all incoming digital records for viruses:
   - if a virus is found, we cease the transfer and contact the source agency
   - if a virus is not found, the records proceed to the next stage of Quarantine processing.
2. Place the digital records in isolation for 28 days.
3. After 28 days of isolation, check the digital records for viruses again:
   - if a virus is found, we cease the transfer and contact the source agency
   - if a virus is not found, the records proceed to the Preservation stage.

The reason we subject incoming files to two virus checks is the anti-virus software is only able to detect viruses that pre-date the last virus definition update. The first virus check should find any viruses already recorded in the virus definitions of the anti-virus software. If the records contain a virus that is not recorded on our virus definitions, the 28 day quarantine period should provide enough time for us to update our virus definitions and detect the virus.

5.3 File format detection and conversion

During Preservation stage processing, DPR calls the Xena application to identify the file format of digital records and to process them according to system rules defined in DPR and Xena software.

There are three types of file formats that we may receive:

- preservation file format
- supported, non-preservation file format
- non-supported, non-preservation file format.

Preservation processing involves the creation of Xena files from each digital record. Each Xena file is an XML file which contains the content of the source file as a base64 encoded bit stream and file-specific metadata. The Xena files are what we preserve in the digital archive.

Base64 encoded bit streams allow us to store the file-specific XML metadata with the content of the digital record. As all the content and metadata are stored in the same file, we do not need to maintain two separate documents, and the link between them, over time. The file-specific XML metadata includes file format information, which is crucial to accessing the content in the future. See Section 5.6.2 for more information on Xena file XML metadata.

Regardless of file format, a Xena file is always created to preserve the content of the original digital record. By preserving the content of the original digital record in the Xena file, we have the ability to reprocess the source records again if we need to. See Section 5.7 for more information on reprocessing.

5.3.1 Preservation file format

A preservation file format has the properties identified in Section 3. When Xena detects digital records in preservation file formats, it converts them to a Xena file containing the original content. The following diagram describes the processing of a preservation file format, using a PNG image file as an example.

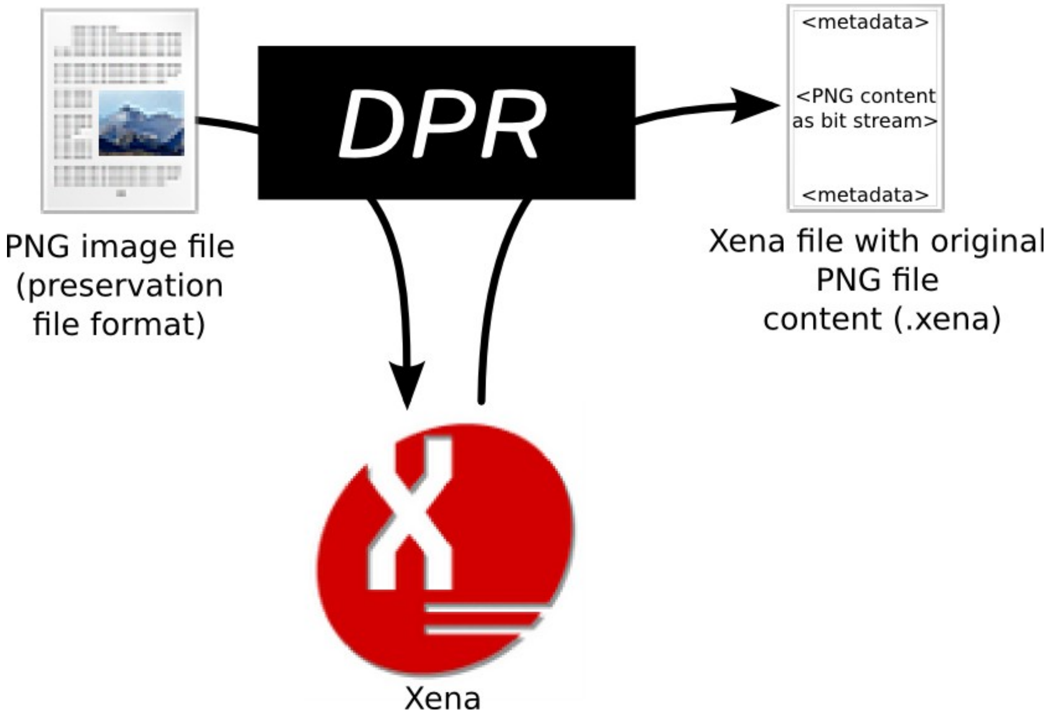
DPR
<metadata>
<PNG content as bit stream>
<metadata>
PNG image file (preservation file format)
Xena file with original PNG file content (.xena)
Xena

Diagram 2: Preservation File Format

5.3.2 Supported, non-preservation file format

A supported, non-preservation file format is one that Xena can convert to an equivalent preservation file format:

- A copy of the source file is converted to the appropriate preservation file format and preserved in a Xena file.
- The original file is preserved within another Xena file.

The following diagram describes the processing of a non-preservation file format, using a Microsoft Word document as an example. Part of the processing involves the conversion of the Microsoft Word file to Open Document Format.

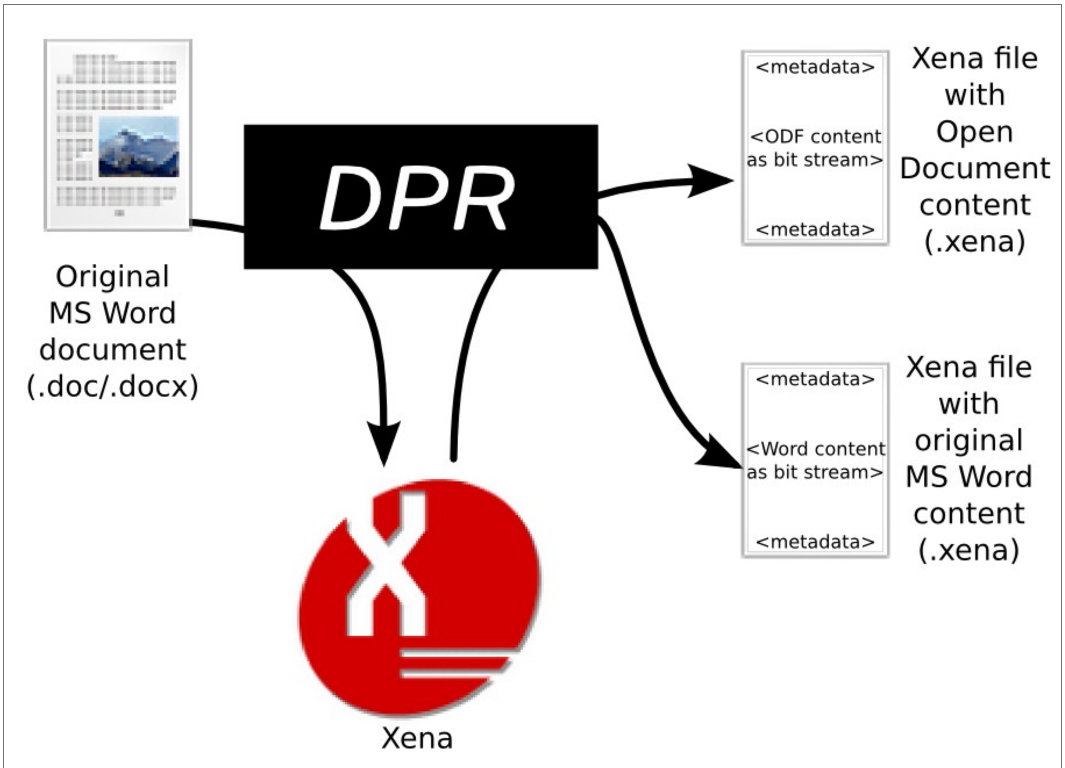
Original MS Word document (.doc/.docx)
DPR
Xena
<metadata>
<ODF content as bit stream>
<metadata>
Xena file with Open Document content (.xena)
<metadata>
<Word content as bit stream>
<metadata>
Xena file with original MS Word content (.xena)

Diagram 3: Non-Preservation File Format

5.3.3 Non-supported, non-preservation file format

A non-supported, non-preservation file format either has no equivalent preservation file format, or Xena is unable to identify the file format. In such cases, the original file content is preserved in a Xena file. The following diagram describes the processing of an unsupported file format.

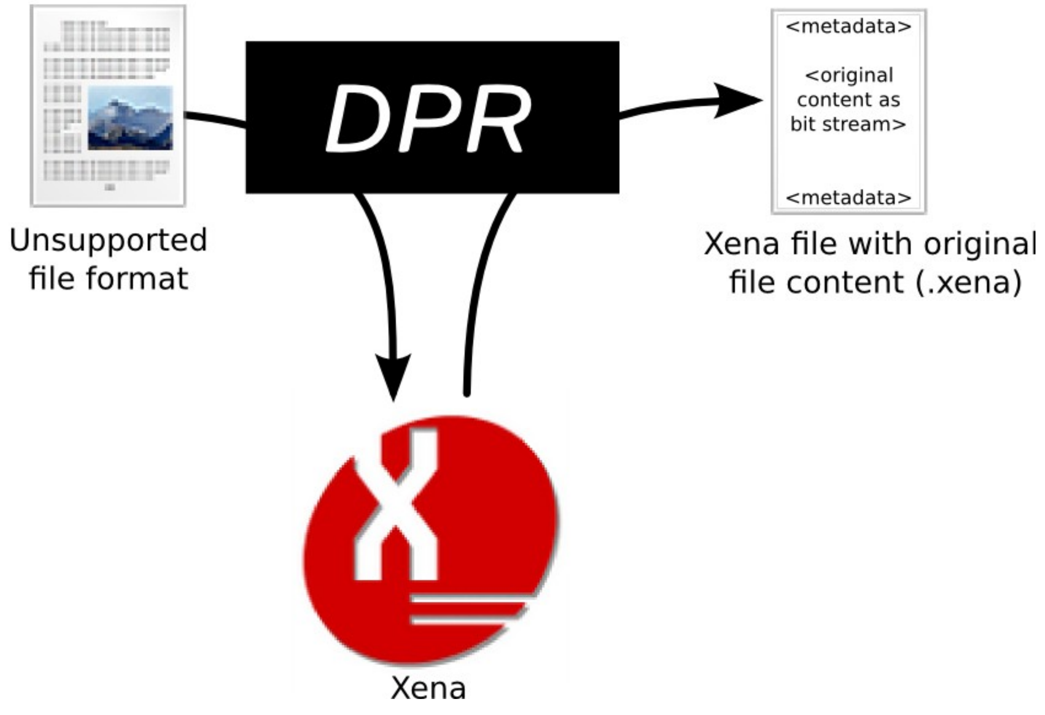
Unsupported file format
DPR
<metadata>
<original content as bit stream>
<metadata>
Xena file with original file content (.xena)
Xena

Diagram 4: Non-Supported File Format

5.4 Automated quality assurance check

During Preservation processing, the DPR performs an automated quality assurance check on all digital records. This check is to ensure that the content of the original record has been encoded accurately in the Xena file.

The automated check involves a comparison of the base64 encoded bit stream of the original content with the original digital record:

1. DPR takes the Xena file and re-creates the record from the bit stream stored in the Xena file.
2. DPR creates a checksum of the re-created record.
3. DPR compares the checksum of the re-created record to the previously stored checksum of the original record. If the checksums are not identical, we know that the original record has not been accurately encoded in the Xena file and we can take the appropriate action.

5.5 Manual quality assurance check

The manual quality assurance check provides a means for an operator to confirm that the conversion process has been performed successfully. It is a simple subjective test, where the operator compares the original to the open format version of the record.

DPR randomly selects a sample of digital records undergoing Preservation processing[9] and presents them to the operator. The operator can make a pass/fail decision and add any relevant comments. The pass/fail decisions and comments form part of the processing metadata stored for the relevant records.

5.6 Capture of metadata

Metadata is captured at two levels: file-specific metadata is captured in the Xena file; DPR processing metadata is captured in DPR database records.

5.6.1 DPR processing metadata

The DPR records metadata relating to each stage of processing. This metadata provides valuable information about the processing of each digital record in the digital archive. Some of the metadata collected includes:

- Xena version used
- normaliser name and version
- virus scanner name and version
- virus definitions version
- files that pass or fail conversion
- date and time of the processing
- original file format
- operator.

[9] The DPR only selects records which have been converted to open formats.

5.6.2 Xena file metadata

All Xena files produced during DPR processing include the Wrapper, Package and the NAA XML schemas. These schemas include AGLS and Dublin Core metadata elements.

A file-specific XML schema is applied according to the type of file format detected (for example, the Audio schema will be applied where the Xena file contains audio file content). If a Xena file contains more than one type of file, multiple XML schemas are applied. For example, an email with attachments.

The following diagram describes the relationship between the various XML schemas[10] applied to a Xena file produced during DPR processing.

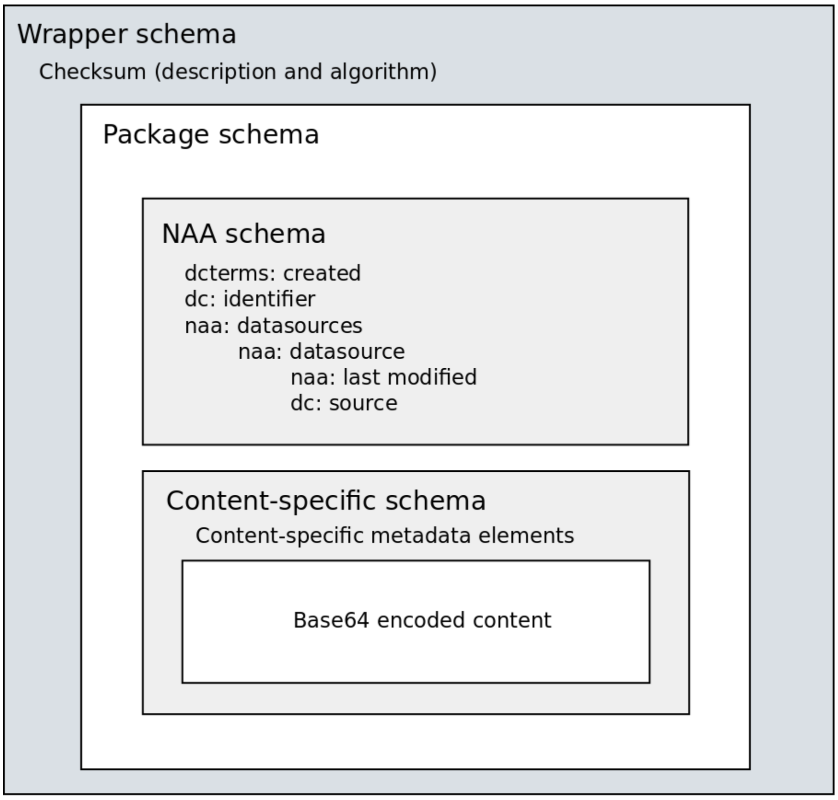

Diagram 5: The Xena Metadata Schema

[10] These XML schemas are available on the Xena wiki at: http://sourceforge.net/apps/mediawiki/xena/index.php?title=Main_Page

5.7 Reprocessing

Reprocessing gives us the ability to select a subset of digital records in the digital archive and put them through the conversion process again. There may be several reasons for wanting to do this:

- a defect in the software is discovered which affects the way the original records were converted. For example, the defect resulted in no conversion taking place or taking place with errors
- we begin to support a file format that was previously processed as an unsupported file format (see Section 5.3.3)
- we enhance our software so that more characteristics of the original record can be retained by the conversion process.

We recognise that any digital preservation conversion is not perfect and, as a result, certain characteristics of the original may be lost. As we preserve the original content of all the digital records we process, we are able to perform any reprocessing on the original digital record rather than the converted version. This helps to prevent any unnecessary loss of the original record's significant characteristics.

6 Maintenance

To ensure that digital records in the digital archive remain safe, we perform a number of regular maintenance activities:

- checksum checking of all digital records in the digital archive
- duplicated storage and routine backups
- software and hardware maintenance
- physical and information technology security.

6.1 Checksum checking

After records are stored in the digital archive, our Checksum Checker software continually checks the checksums of digital records and compares them with the values stored in the digital repository database. If there are any discrepancies detected, administration staff are alerted so they can take appropriate action.

6.2 Duplicated storage

The content of the digital archive is duplicated on two independent storage networks using hardware from different vendors. This eliminates any reliance on a single vendor or technology and helps to ensure that we remain in control of the data stored in the digital archive.

The digital archive is backed up regularly. This currently involves simple backups to tapes which are sent off site. However, the design of the digital archive is such that it allows for a complete mirror to be created in a remote location.

7 Appendix A - Xena third party libraries

The following table lists all third party libraries used by the Xena software.

Library Name	Description	Creator	URL	License
Apache Commons CLI	An application programming interface (API) for parsing command line options passed to programs.	Apache Software Foundation	http://commons.apache.org/cli/	Apache License Version 2.0
Apache POI	A Java API for Microsoft documents. Used for text extraction from emails.	Apache Software Foundation	http://poi.apache.org	Apache License Version 2.0
Batik	Used to manipulate SVG documents.	Apache Software Foundation	http://browserlaunch2.sourceforge.net/	Apache License Version 2.0
BrowserLauncher2	Facilitates opening a browser from a Java application and directing the browser to a supplied URL.	http://browserlaunch2.sourceforge.net/	http://browserlaunch2.sourceforge.net/	LGPL
Crystal icon set	Icon set used in the graphical user interface for Xena.	Everaldo Coelho	http://www.everaldo.com/crystal/	LGPL
GNU JavaBeans Activation Framework	A file type map that specifies the MIME content type for a given file and a way to specify actions for given MIME type content.	GNU	http://ftp.gnu.org/gnu/classpathx/	LGPL
GNU JavaMail	A free implementation of the JavaMail API.	GNU	http://ftp.gnu.org/gnu/classpathx/	LGPL
im4java	A pure-java interface to the ImageMagick commandline.	http://im4java.sourceforge.net/	http://im4java.sourceforge.net/	LGPL 2.1

Library Name	Description	Creator	URL	License
inetLIB	Internet address manipulation routines.	GNU	*http://ftp.gnu.org/gnu/classpathx/*	LGPL
International Components for Unicode (icu4j)	ICU is a set of C/C++ and Java libraries providing Unicode and globalization support for software applications.	*http://site.icu-project.org/*	*http://site.icu-project.org/*	Basic BSD Style License
JavaHelp System	An online help system used by Xena.	Sun Microsystems	*https://javahelp.dev.java.net/*	GPL2 only with Classpath exception see: *http://www.gnu.org/software/classpath/license.html*
javalayer	A Java library that decodes MP3 files in real-time.	*http://www.javazoom.net/javalayer/javalayer.html*	*http://www.javazoom.net/javalayer/javalayer.html*	LGPL 2.1
jflac	jFLAC is a port of the Free Lossless Audio Codec (FLAC) library to Java.	Josh Coalson	*http://jflac.sourceforge.net*	GPL 2+ or Modified BSD
JGoodies Looks	JGoodies Looks is a library used to help improve the user interface for the Xena application.	Jgoodies, Karsten Lentzsch	*http://jgoodies.dev.java.net* and *http://www.jgoodies.com/downloads/libraries.html*	Modified BSD License (No advertising clause)
jorbis	A Java Ogg Vorbis decoder.	*http://www.jcraft.com/jorbis/*	*http://www.jcraft.com/jorbis/*	LGPL 2.1

Library Name	Description	Creator	URL	License
JPedal	Java PDF library, providing a Java PDF viewer, PDF to image conversion, PDF printing or adding PDF search and PDF extraction features.	http://www.jpedal.org/	http://www.jpedal.org/	LGPL 3
jreleaseinfo	Creates a Java source file during the build process.	http://jreleaseinfo.sourceforge.net/	http://jreleaseinfo.sourceforge.net/	Apache License Version 2.0
jspeex	An encoder and decoder for the Speex audio codec.	Jean-Marc Valin	http://jspeex.sourceforge.net/	Modified BSD
juh.jar	Provides interoperability between Java and the Open Office API	OpenOffice.org / Uno Module	http://udk.openoffice.org/	LGPL 3
jurt.jar	Provides interoperability between Java and the Open Office API	OpenOffice.org / Uno Module	http://udk.openoffice.org/	LGPL 3
mp3spi	A Java Service Provider Interface that adds MP3 (MPEG 1/2/2.5 Layer 1/2/3) audio format support for Java Platform.	http://www.javazoom.net/mp3spi/mp3spi.html	http://www.javazoom.net/mp3spi/mp3spi.html	LGPL 2.1
MPXJ	Provides a set of facilities to allow project information to be manipulated in Java and .Net.	http://mpxj.sourceforge.net/	http://mpxj.sourceforge.net/	LGPL
ridl.jar	Provides interoperability between Java and the Open Office API	OpenOffice.org / Uno Module	http://udk.openoffice.org/	LGPL 3
Sanselan	Reads and writes a variety of image formats, including fast parsing of image info (such as size, colour space) and metadata.	Apache Software Foundation	http://incubator.apache.org/sanselan/	Apache License Version 2.0

Library Name	Description	Creator	URL	License
TAR	Allows for creating and extracting of TAR archives.	com.ice	http://www.trustice.com/java/tar/	Public Domain
Toastscript	A Java implementation of the PostScript language.	http://sourceforge.net/projects/toastscript/	http://sourceforge.net/projects/toastscript/	GPL 2+
tritonus	An independent implementation of the Java Sound API	http://www.tritonus.org	http://www.tritonus.org/plugins.html and https://sourceforge.net/scm/?type=cvs&group_id=1390	LGPL 2
unoil.jar	Provides interoperability between Java and the Open Office API	OpenOffice.org / Uno Module	http://udk.openoffice.org/	LGPL 3
w3c SVG and SMIL	The Java language binding for the SVG and SMIL object models.	w3c	http://www.w3.org/TR/SVG11/java.html and http://www.w3.org/TR/smil-boston-dom/java-binding.html	Modified BSD
w3c SAC	An API for Cascading Style Sheets (CSS)	w3c	http://www.w3.org/Style/CSS/SAC/	Modified BSD
Xalan-Java	Xalan-Java is an XSLT processor for transforming XML documents into HTML, text, or other XML document types.	Apache Software Foundation	http://xml.apache.org/xalan-j/	Apache License Version 2.0
Xerces	A collection of software libraries for parsing, validating, serializing and manipulating XML.	Apache Software Foundation	http://xml.apache.org/dist/xerces-j/	Apache License Version 2.0

Library Name	Description	Creator	URL	License
xt.jar	An implementation of XSLT in java.	*http://www.blnz.com/xt/index.html*	*http://www.blnz.com/xt/index.html*	BLNZ License

8 Appendix B - Glossary

Term	Definition
AGLS	AGLS Metadata Standard. The AGLS Metadata Standard (Australian Standard 5044) is a set of descriptive properties to improve visibility and availability of online resources. Developed by the National Archives of Australia.
AI	Adobe Illustrator vector image.
AIFF	Audio Interchange File Format. An audio file format, most commonly used on Apple Macintosh computers.
ASCII	American Standard Code for Information Interchange. A character encoding scheme for representing the English alphabet and punctuation (limited to 128 characters). Common character encoding format for plain text files (see also **Unicode**).
Automated Quality Assurance	During Preservation processing, the DPR automatically compares the binary normalised records with the original digital records to check that they have been encoded accurately.
Base64	Representation of binary data in an ASCII string format.
Binary	A system of counting using 1s and 0s. A binary file is a computer file which may contain any type of data for computer processing and storage.
Binary Normalisation	Base64 encoding of the content of a digital object, which is then wrapped in XML metadata.
Bit	A binary digit. In computing, a bit can either be a 1 or a 0.
BMP	Bitmap image file format.
BWF	Broadcast Wave Format. An extension of the Microsoft WAVE audio format.
Character Encoding	A system for representing individual characters with a code, such as a sequence of numbers. ASCII, ISO 8859 and Unicode are some popular character encoding schemes.
Checksum	An alpha-numeric value calculated from the contents of a digital object. For the purposes of digital preservation, the checksum is used to verify the integrity of the digital object – by comparing a recently determined checksum with a stored one, you can tell if the digital object has changed.
Checksum Checker	Software that monitors the digital archive for data loss or corruption.

Term	Definition
CSS	Cascading Style Sheets. A style sheet language used to describe the formatting of a document written in a markup language, such as **XML** or **HTML**.
Digital Archive	Permanent storage for significant digital records. The digital archive is accessed through the Digital Repository (DR) stage of DPR.
Digital Object	An object composed of a set of bit sequences. A single record (file or document) to be archived.
Digital Preservation Recorder	Software that manages the workflow for the National Archives' digital archiving process. It consists of three distinct stages: • Quarantine Facility • Preservation Facility • Digital Repository.
Digital Repository	Third stage in the DPR workflow. Copies digital objects to the digital archive. Also includes functions for managing and displaying DPR process metadata.
DPR	See **Digital Preservation Recorder**.
Dublin Core	A set of metadata elements for describing and cataloguing resources, such as books or digital materials. Dublin Core is defined by the International Standards Organisation (ISO) Standard 15836.
DOC	Microsoft Word Document.
DOCX	Microsoft Office Open XML Document.
DPR	See **Digital Preservation Recorder**.
DR	See **Digital Repository**.
EPS	Encapsulated PostScript. A file format containing vector and sometimes bitmap data.
FLAC	Free Lossless Audio Codec. A free and open source software tool and file format for lossless audio data compression.
GIF	Graphics Interchange Format.
GZIP	A software application used for file compression.
HTML	HyperText Markup Language. The main markup language for web pages.
Item Number	An item is a discrete unit within a series which can contain one or more digital objects. An item is uniquely identified by the item number. The item number is used to identify, locate and request items on RecordSearch.

Term	Definition
JAR	Java Archive. A JAR file combines many other files into one.
JPEG	Joint Photographic Experts Group file. A file format which employs a lossy compression for digital images.
Library	Libraries contain code and data that provide services to independent computer programs.
Magic Number	A numeric or text value used to identify a file format.
Manifest File	A machine-readable list of data objects in a Transfer Job, along with their **item number**. The Manifest File is created outside of the DPR workflow.
Manual Quality Assurance	A subjective test, where the operator compares the original file to the open format version. Performed during DPR Preservation processing on a randomly selected sample of digital records.
Metadata	Data about other data. DPR collects and makes available information about the following: • contents of items and digital objects • events in the processing of transfer jobs (actions taken, dates, results) • user activity.
MBX	Mailbox message file. A mailbox or mail folder that contains Microsoft Outlook Express e-mail messages.
MIME type	Multi-purpose Internet Mail Extensions. RFC2045 Internet standard allowing email to support attachments and non ASCII text. Defines the kind of data formatting used by a particular digital object.
MP3	MPEG-2 audio layer 3 (mp3) audio file. A lossy compressed audio format developed by the Moving Picture Experts Group.
MPP	Microsoft Project file.
Normalisation	Conversion of a digital object into an open standards based format which is then base64 encoded and wrapped in XML metadata.
Normaliser	The normaliser is a component of a Xena **plugin** responsible for taking an input file and transforming it into a Xena file.
ODF	Open Document Format. An XML-based file format for representing spreadsheet, text or presentation data.
ODG	Open Document Graphics file.
PDF	Portable Document Format.
PF	See **Preservation Facility**.

Term	Definition
Plugin	Plugins are a set of software components that add specific capabilities to a larger software application. To process digital records, **Xena** utilises different plugins for various categories of file types. For example, audio, email and image.
PNG	Portable Network Graphics file. An image format that employs lossless data compression.
PSD	Adobe Photoshop document. An image file created by Adobe Photoshop.
PPT	Microsoft Powerpoint Presentation.
PPTX	Microsoft Powerpoint Office Open XML Presentation.
Preservation Facility	The second stage of the DPR workflow. Involves the use of **Xena** to perform conversion of digital records to open standards based file formats.
PST	Personal Storage Table. A Microsoft Outlook file format used to store email messages, contacts other data.
QA	See **Quality Assurance**.
QF	See **Quarantine Facility**.
Quality Assurance	A step within the **Preservation Facility** stage of processing. Consists of **Automated Quality Assurance** and **Manual Quality Assurance**.
Quarantine	Before data can be normalised and archived, it must be scanned for viruses and stored on an isolated carrier for 28 days before a second virus and checksum check.
Quarantine Facility	Initial stage of DPR workflow includes manifest file processing, pre-quarantine processing and post-quarantine processing.
RecordSearch	The National Archives' archive management system. Items described in RecordSearch (including digital objects) are identified by an **Item Number**.
Reprocessing Job	If the policy or technology for normalising a particular file type changes, digital objects of that type can be retrieved from the digital archive and processed again according to the new circumstances.
RTF	Rich Text Format file. A method for encoding formatted text and graphics for transfer between applications.
Significant characteristics	Characteristics that must be preserved for the digital object to maintain its meaning over time.
SQL	Structured Query Language. A database computing language for managing the contents of a relational database. Includes insertion, query, update and deletion of data.

Term	Definition
SVG	Scalable Vector Graphics. An XML-based file format for describing two-dimensional vector graphics.
TAR	Consolidated Unix File Archive. A file archive in an uncompressed format created by the Unix Tar utility.
TIFF	Tagged Image File Format. Graphics container that can store both raster and vector images.
Unicode	Industry standard for encoding text characters from most of the world's languages. It is a common character encoding format for plain text files.
UTF-8	An 8-bit character encoding for **Unicode**, which is backwards compatible with the **ASCII** standard.
WAV	Waveform Audio file.
Xena	File normalisation software. Xena functionality is integrated in to the Preservation Facility stage of DPR. Xena can also be used as a stand-alone product.
Xena File	An XML file containing base64 encoded source file content, wrapped in metadata.
XHTML	eXtensible HyperText Markup Language.
XLS	Microsoft Excel Spreadsheet format.
XLSX	Microsoft Office Open XML Workbook.
XML	eXtensible Markup Language.
XSL	eXtensible Stylesheet Language.
XSLT	XSL Transformations. An XML language for transforming XML documents.
ZIP	A lossless compressed file archive format.