

The Benefit of Experience: *the first four years of digital archiving at the National Archives of Australia*

Michael Carden - August 2010

Introduction

The National Archives of Australia (NAA) commenced research into the preservation of digital records in the year 2000 with a project that in December 2002 produced our foundational document for digital archiving; *An Approach to the Preservation of Digital Records*¹. A key conclusion of that research was that the preservation of digital records should focus on maintaining the *performance* of a record over time rather than attempting to preserve the specific combination of source data and technology that represent a record at its creation.

The research team looked at the factors affecting the creation of a performance of a record and concluded that a means was required to move digital records away from specific technologies and to represent digital records in openly specified formats based on freely available standards. This gave rise to the development of the *Xml Electronic Normalising for Archives*² (Xena) digital preservation software. The Xena software is designed to determine the file formats of digital records and to perform conversions into appropriate open formats while adding some preservation metadata.

With Xena at the core of the digital preservation process, the NAA then developed a sophisticated tool to manage a digital preservation workflow calling on Xena and other tools while collecting an audit trail of the process. This tool is known as the *Digital Preservation Recorder*³ (DPR) and like Xena, DPR is open source software freely available for download.

With the key software tools in place, the team constructed a computer server and storage facility to host a prototype digital archive and that facility has been in daily use since 2006, preserving the digital records of Australia's Commonwealth government.

The operation of a digital archive has taught the NAA many valuable lessons concerning software development, business processes, scaling, quality assurance and the value of working with each and every digital object in our care. This paper will share some of those lessons so that you might benefit from our experience.

1 http://www.naa.gov.au/images/an-approach-green-paper_tcm2-888.pdf

2 <http://xena.sourceforge.net>

3 <http://dpr.sourceforge.net>

The National Archives of Australia

The National Archives of Australia is an agency of the Australian Government, established under the *Archives Act 1983*. Our head office and exhibition spaces are in Canberra and there is an office and reading room in each state capital.

The National Archives of Australia:

- helps Australian Government agencies create and manage their records;
- selects the most valuable records created by Australian Government agencies to become part of the national archival collection;
- stores, describes and preserves the national archival collection;
- and makes records in the national archival collection that are over 30 years old publicly available.

Our staff ensure that the national archival collection – in formats that include paper, audiovisual and digital material – is controlled and described and that it remains stable and accessible over time.

The National Archives also maintains the administrative history of the Australian Government so that records can be related to their original context. We maintain records repositories around the country and conservation laboratories in Canberra, Sydney and Melbourne. The majority of our audiovisual collection is located in Sydney and our digital collection is located in Canberra.

The Performance Model

Digital records are always mediated by the computing platforms and software applications used to create or render them. Over time, computing hardware has undergone rapid evolution while operating systems and software applications have seen similar high rates of change. These changes, coupled with the relative fragility of digital storage media, pose a threat to the long term access to digital records. The National Archives recognises that unless proactive intervention takes place, digital records are not likely to survive beyond a few years.

The performance of a record is simply the process of rendering it in a meaningful way. For a document or an image or a video, this usually means display via a screen or a projector. For audio recordings a performance means playback via loudspeakers or earphones.

In order to separate the performance of a record from the technologies that created it, the NAA has sought to convert all source data objects into standards based open formats that we believe offer the best opportunity for preservation into the future.

Xena Software

The National Archives' Xena software is written in the Java programming language so that it is not tied to any single computer architecture. Most of our development takes place on Linux machines but we also test on Windows and Mac OS-X systems. The software is developed using an open source methodology where our code is available via the world's largest open source project website, sourceforge.net. The Sourceforge site manages our source code through the GIT⁴ distributed source code management system and hosts bug and feature trackers for us. Most importantly, Sourceforge gives us an easy way to distribute our finished products via a worldwide series of mirror sites from where it is downloaded between five hundred and one thousand times per month. We aim to do a full release of the software approximately every six months, but for those confident in compiling Java source code into a finished application, the code is updated daily and available for copying from our GIT repository⁵.

While Xena can be downloaded and used as a desktop application to perform preservation transformations on collections of files, in practice not many people will use it that way. To easily integrate Xena with other digital archiving software, we have included an Application Programming Interface (API) that allows other software to easily call on Xena's conversion services. Others have taken advantage of this functionality and Xena has been integrated with the DSpace digital archive application and with the Alfresco content management system.

Plugins

Our software has been designed with a 'plug-in' architecture in which we create an expert plug-in for each genre of file format. We currently have plug-ins for audio, csv, email, html, image, office, pdf, project and zipped files. Each of these plug-ins caters for a range of file formats of its type. For example, the image plug-in can recognise some fifteen different image file types including BMP, GIF, JPG, TIFF, PSD and others. The full list of supported file types is available via the Help⁶ pages on the Xena website.

Our small team cannot hope to be experts on every file format in existence. For this reason we always try to leverage the expertise embodied in open source software written by others to work with different file formats. Where necessary we write the code for our plug-ins, but where possible we will seek pre-existing code that we can use. A great example of this is our plug-in for Office document formats. A large part of our Office plug-in is the OpenOffice.org free and open source

4 <http://git-scm.com/>

5 <http://sourceforge.net/projects/xena/develop>

6 <http://xena.sourceforge.net/help.php?page=normformats.html>

office suite. The OpenOffice developers have done an excellent job of reverse engineering proprietary document formats and we are able to build on their work. Parts of our image plug-in rely on the open source ImageMagick set of tools while we have had to write some parts ourselves to deal with some obscure image formats. One of the great benefits to us of developing open source software is this ability to mix our code with pre-existing code released under compatible open source licenses.

Format Determination

The first thing that Xena must do when presented with a data object is to determine what its file format is. Doing so accurately is not as simple as merely looking at the three letter extension on a file name. Xena does this by introducing the data object to each of the plug-ins in turn. The plug-ins evaluate the data object and return to Xena a score based on how likely it is that the object is a file of a particular type.

For example, if a TIFF image file is passed to the Plaintext plug-in, the plug-in will look at the file name and at the file header to see if there are any clues to indicate that it may be of the Plaintext type. The plug-in may notice some familiar characteristics because a TIFF header is mostly text, so it might return a score of 100. Then the TIFF is passed to the Office plug-in where it is again examined. This time the plug-in really doesn't recognise anything about the file, so it returns a score of minus 500. Next, the TIFF is handed to the image plug-in where again the file name and internal structure are checked. This time the plug-in sees a lot that it likes and makes some further checks. Finally the plug-in attempts to render the TIFF image internally and this succeeds. As a result, the image plug-in has a very high confidence that the file is a TIFF and it returns a score of 5000. Once Xena has exposed all of the plug-ins to the data object, it evaluates the response scores and selects the most likely format based on the highest score.

This method of calling on the expertise of a range of file format plug-ins has proved to be surprisingly accurate in determining file types. Careful adjustment of the confidence scores returned by each plug-in has resulted in a high degree of accuracy and as more plug-ins are developed, the overall accuracy improves. We are now in a situation where most cases of format identification failure are caused either by corrupt data or by encrypted files.

Conversion

Once Xena has determined a data object's format, it makes use of the individual plug-ins to perform format conversions.

Each plug-in has a target preservation format for its genre of formats. The image plug-in converts to the Portable Network Graphics (PNG) file type, the Office plug-in converts to OpenDocument

Format (ODF), the audio plug-in converts to Free Lossless Audio Codec (FLAC) and so on. If a data object is already in an appropriate format, Xena makes no changes to its format.

The preservation formats that we have selected for Xena are based on a small set of criteria that we believe are key to the longevity of digital records:

- Standards based - unrestricted access to the format specification.
- Community developed - not the work of a single entity.
- Multiple implementations - a wide choice of software implementing the format.
- Maintains significant properties of source formats – essential characteristics.
- No patent or license restrictions - no risk of paying to use the format.

The underlying principle for these criteria is that if a researcher encounters a digital object in the future and that digital object is encoded in a format which is openly described and widely understood, it should be possible to locate or write software to interpret the format.

Once Xena has converted a data object into an open format, it wraps the resulting file with a small set of XML preservation metadata relevant to that data object. The metadata wrapper used by the NAA contains elements that suit our business needs but may not suit others, so we have designed the wrapper to be easily modified to fit any digital preservation workflow.

Text Extraction

Access to all records in the care of the National Archives, whether paper, audiovisual or digital, is controlled by our *RecordSearch*⁷ collection management application. The system provides a search interface that allows users to find records based on titles and a range of other intellectual control metadata, but does not provide a facility to search within individual records. We have recently enhanced our Xena software to make possible a full text search of digital records. Any record that has a document representation, such as a word processor document or a PDF or an email, can have text extracted during Xena processing to create an index for searching. Extracting text from these types of data objects is not difficult, but if the source document is a scanned image of a paper document stored as an image file, it is necessary to perform Optical Character Recognition (OCR) to recover text from the image.

The most common file format created during the scanning of paper documents is the Tagged Image File Format (TIFF). To give Xena the ability to extract text from TIFF files, we have integrated it with Google's open source OCR software, *Tesseract*⁸. The Tesseract software is very accurate, is

7 <http://naa.gov.au/collection/recordsearch/index.aspx>

8 <http://code.google.com/p/tesseract-ocr/>

well documented, is freely available and has easily integrated with Xena.

Aside from the benefits that a search index can bring to researchers looking for another way to access our records, the index has the potential to greatly assist the National Archives staff responsible for examining records to provide access clearance.

Digital Preservation Recorder

Our process of digital preservation involves moving data between computer systems, checking data to guarantee its integrity, file format transformations, quality checking and the collection of preservation metadata. Once we had developed our Xena software to manage format transformations, a tool was needed to collect an audit trail of the processes completed during the ingest of digital records to the digital archive. This need has been met by the development of the Digital Preservation Recorder (DPR) – a desktop application that manages antivirus software, checksum verification and the Xena software. In addition, the DPR collects an audit trail of preservation metadata detailing the preservation process for each and every data object stored in the digital archive.

Though our digital preservation process requires a number of different pieces of software to perform a range of tasks, staff who operate the systems do not need to interact with separate pieces of software. Everything is managed as a guided workflow through the Digital Preservation Recorder. The software manages a three stage work flow, conducted on three separate computer systems which are not connected to one another nor to any other networks. This strategy was chosen in order to protect the integrity of the records held in the digital archive in the event of computer viruses or malicious software accidentally arriving via a records transfer.

Quarantine

The first stage of DPR processing is the quarantine stage. Digital records arrive at the National Archives on physical media such as tape or disk and are accompanied by a manifest file that records the name and checksum of each data object.

The quarantine process begins with an automated check of the manifest to make certain that the transfer contains all of the data objects that it should and that no data has been corrupted in transit. This is followed by an automated antivirus check and the data is copied to one of our carrying devices. Four years ago our carriers were 200 gigabyte external hard disks with USB connections. Currently we're using 1 or 2 terabyte external hard disks with E-SATA connections.

The carrier is then disconnected from the quarantine network and placed into secure storage for 28

days. During the 28 day quarantine period, the antivirus definitions on our quarantine network are updated daily. It is our hope that any computer virus hiding in the data and which we fail to detect on an initial scan, may be detected by our updated software 28 days later.

At the end of the quarantine period, the carrier is reconnected to the quarantine network for a second scan. If the data passes the second scan, the quarantine process is completed and the carrier is removed from the quarantine network to be connected to the preservation network.

Preservation

The preservation system is where the DPR calls on the Xena software to determine file formats and perform conversions where needed. In addition, the DPR uses Xena to wrap each data object in some preservation metadata, so at the end of this process we have two things to store in our digital archive; the original data object wrapped in metadata and the open format conversion, also wrapped in metadata.

For quality control purposes, the DPR selects a sample of the converted data objects and provides the operator with an opportunity to make a 'before and after' evaluation of the conversion process. Ideally we would like to completely automate this part of the workflow, but technology does not yet offer the means to replicate the ability of a person to make a subjective judgement.

Once format conversions have been completed and checked, the DPR creates a new checksum for each of the newly created data objects and stores those on the carrier with the data.

Digital Archive

The final stage of our process is the ingest into the digital archive. The carrier device is removed from the preservation network and connected to the digital archive network where all of the data objects and their checksums are copied to high volume storage.

To achieve additional security and redundancy, our digital archive is actually two completely separate systems. We use two different operating systems, two different types of disk storage and two different file systems. At present we employ Windows, Viper RAID and Microsoft's NTFS on one system while the other system has Linux, Apple X-Serve RAID and EXT3. The same data is copied to both systems so that in the event of a failure on either system, we have a full copy the data on the other system. Ideally we would locate one of the two systems in another city, but since both systems are currently in the same building we also create off-site backup tapes.

Once the data is secured within the digital archive, we commence a continuous integrity check by

reading each data object and checking its checksum against the value we store in our database. This provides us with an early warning of hardware or systems failures that may damage the records.

Lessons Learned

Open Source

The National Archives decided very early that any software developed for digital preservation should be released under an open source license. In order to protect the organisation's intellectual property, the GPL2 license was initially selected. Open source software offers the prospect of collaborating with other interested parties without barriers to engagement and helps in getting the software in front of the largest possible audience.

Open source development allows our small team to create large software projects by building on top of existing open source code libraries. We could not have achieved our current level of development without access to open source code.

Finally, open source software for digital preservation opens our processes to external scrutiny. Doing digital preservation always involves working with data in ways that expose records to the risk of unintentional change – even if only in copying from one location to another. Making our software open source allows others to critically inspect our processes and demonstrates the authenticity of what we do.

Licensing

When we commenced our software development seven years ago, we were somewhat naive when it came to license selection. Knowing that we wanted strong protection for our intellectual property, we released our code under the GNU General Public License 2 (GPL2). This was effective and the GPL remains a very useful license.

As our code grew over time and incorporated new features, we included many external open source software libraries to perform specific tasks. Eventually we would include nearly 100 external projects in our software. By some estimates, we may have written less than 5% of the lines of executable code in our finished products.

When selecting external libraries we would always look for those with an open source license and readily available source code. If a library met our needs and had an open source license, we would use it.

Last year, we decided to take a closer look at the licenses of the wide range of external projects that we depend on and we discovered with some surprise that many of the different open source licenses employed by these projects were either not compatible with the GPL2 or were not compatible with

each other. This triggered a major overhaul of our code in which we changed our own license to the GPL3 (or later), replaced incompatible libraries, carefully documented all licensing and wrote replacement code for those libraries that we could not replace. We now have a much better understanding of the nuances of open source licensing, and we would strongly advise that anyone doing open source software development should carefully check license compatibility before incorporating external code into a project.

Transformation on ingest

The process of file format transformation on ingest involves our software interacting with each individual data object as it passes into our repository. Doing so provides us with the great benefit of discovering immediately if a file is corrupt, encrypted, password protected or otherwise inaccessible. This gives us an opportunity to liaise with a record's creator to discover the cause of any problem while the creator is still available.

Contrast this with the approach of many archival institutions that accumulate digital records and only attempt to change their format at a future time when a format's obsolescence is predicted. If a file is found to be corrupt or inaccessible in five, ten or twenty years, there is a good chance that its creator will not be available to provide an accessible copy.

Digital records bring with them the opportunity for automation in ways that traditional paper records could never bring. We believe that it's important to take advantage of automated processes around digital records so that costs can be measured in very inexpensive CPU cycles rather than in relatively expensive staff hours.

Scale

Scaling a digital archive to manage the ingest of hundreds of thousands of digital records at a time is a challenge. The obvious parts of the challenge are the need for large and reliable data storage and the need for computing infrastructure to support processing software. What is less obvious is that as the number of digital objects in a transfer grows from hundreds to thousands and then to hundreds of thousands, the quantity of metadata to be managed grows at the same rate.

Our early tests on the first versions of DPR were done using fake records transfers of hundreds or thousands of data objects. These tests were very successful. Once we started using DPR in production with real transfers, we began to receive records in tens of thousands or hundreds of thousands at a time and we immediately encountered difficulty in processing the associated metadata. Resolving this issue has resulted in a significant change to the way that DPR manages metadata through our process.

The Future

The digital environment remains an interesting and dynamic place to work. New technologies, new file formats, new ways of creating records and new ways of interacting with systems are all going to change the digital preservation landscape as time moves on.

We expect to develop Xena plug-ins for more formats in the coming years, hopefully through collaboration with interested parties around the world.

We anticipate challenges in scaling our systems to cope with individual files sized in the terabyte range.

Our current method of receiving records on physical media will not be adequate once all government agencies are ready for regular digital transfers and we plan to implement a secure, network-based means of transfer.

Finally, though we have an impressive amount of automation in our current process, our goal is always to remove operator intervention wherever we can. A fully automated transfer and ingest process for digital records should be possible and we are working in that direction.