Digital Archiving at the National Archives of Australia: Putting Principles into Practice

Michael Carden - August 2012

Introduction

The National Archives of Australia has been engaged in digital preservation work for over a decade. In that time we have completed some fundamental research, developed software and hardware systems based on that research, developed our skills as digital archivists and implemented a digital archive that has been securing Australian Government records for over six years. Preserving digital records is one of the core business activities of the archives.

What I will do in this brief presentation is share some of that work with you, take a look at how relevant the research we conducted ten years ago remains today, and consider some of the future challenges that face us. Along the way, I will dip briefly into some of the software we have created and the principles it is built on. There are many aspects to managing digital records, so on this occasion I will be concentrating only on the internal processes that we use to secure, preserve and control digital records.

The National Archives of Australia

The National Archives of Australia (NAA) 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 of the other Australian state or territory capitals.

The National Archives of Australia:

- helps Australian Government agencies create and manage their records;
- selects the most significant 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 collection that are over 20 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. We also maintain the administrative history of the Australian Government so that records can be related to their original context. There are conservation laboratories in Canberra, Sydney and Melbourne.

Motivation

Before I delve into our work in digital preservation, I would like to take a moment to discuss the motivation for investing time and resources into *proactive* digital preservation.

Digital records are inherently fragile and are very easily altered or destroyed. Changes to digital material can be large or small, they can be deliberate or accidental and they can even occur through inaction rather than through any external influence. An example of change through inaction is where digital media like tape or optical disk deteriorates while in passive storage. Usually the first indication of this type of problem is when an attempt is made to read the media.

The effect of a change in a digital record can vary from the trivial to the catastrophic, mostly depending on the nature of the digital material affected. As we all know, digital material at its heart is merely a collection of ones and zeros. Changing any single one or zero may result in an invisible change in a digital video, may perhaps result in the alteration of a single character in a text document and if it's in a critical location, may result in the utter destruction of a digital image.

A consequence of this malleability in digital materials is that there is no such thing as “benign neglect” in the digital domain. Unlike a paper record that may spend decades untouched on a shelf and remain perfectly readable, a digital record is much more likely to have undergone change in that time.

In the domain of paper records, we preserve original documents to ensure that they are trustworthy. The concept of originality is challenged by the fragility of digital records, so we have had to design software and systems to assure that we can provide access to trustworthy digital records.

Research Foundations

The National Archives of Australia commenced research into the preservation of digital records in the year 2000 with a project that in December 2002 produced our foundation document for digital archiving; *An Approach to the Preservation of Digital Records*[1], better known internally as The Green Paper. At only 24 pages long it's quite a brief paper, but it does a good job of explaining technical concepts using non-technical language. I urge anyone with an interest in digital preservation to follow the link to our web site and download a copy.

A key conclusion of the research explained in the paper 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 supports a record at its

1 http://tiny.cc/green-paper

creation. This fundamental idea, now adopted by a number of archival institutions around the world, remains central to everything we have done since then.

Other core concepts exposed in the paper include explorations of the *essence* of digital records, the value in employing *Open Standards* when preserving digital records and the benefits of developing *Open Source* software to deal with digital records.

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 digital records into standards based open formats that we believe offer the best opportunity for preservation into the future.

Our 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.

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 the Australian Government.

2 http://xena.sourceforge.net
3 http://dpr.sourceforge.net

Xena Software

The National Archives' Xena software is designed to automatically determine the file formats of digital records and to perform conversions into appropriate open formats while adding some preservation metadata. Xena 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 helps to manage the code we write through a distributed source code management system and provides bug and feature trackers for each of our software projects. 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.

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 into numerous preservation workflows addressing a range of business needs.

Plug-ins

Our software has been designed with a 'plug-in' architecture in which we create an expert plug-in for each genre of file formats. 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 Help[4] 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 LibreOffice free and open source office suite. The LibreOffice developers have done an excellent job of interpreting a range of proprietary

4 http://xena.sourceforge.net/help.php?page=normformats.html

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 for some more obscure image formats, we have had to write the code ourselves. 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. 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.

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 at the stage 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.

Open Formats

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.

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 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.

Although 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 or 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 digital record.

The quarantine process begins with an automated check of the manifest to make certain that the transfer contains all of the records that it should and that no data has been corrupted in transit. This

is followed by an automated antivirus check and the records are copied to one of our carrying devices. In 2006 our carrying devices were 200 gigabyte external hard disks with USB connections. The size of digital records transfers has increased so much since then, that currently our carriers are external hard disk clusters with E-SATA connections and up to 10 terabytes capacity.

Once the contents of the transfer has been verified on our quarantine system, the carrying device is disconnected from that system and connected to the Preservation system.

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 make subjective judgements as effectively as a person can. We are conducting research in this area though, and we have some prototype software aimed at solving at least part of the problem.

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 records 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. The same data is copied to both systems so that in the event of a failure on either system, we have a full copy of 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.

Our digital archive is a secure facility only accessible to very few people, and not connected to any external networks. Again, this is about mitigating the risk of any changes to digital records.

Software Suite

Collectively, our Digital Preservation software tools are known as the Digital Preservation Software Platform (DPSP) and a single click installer can be downloaded from the DPSP web site on Sourceforge.[5] This single download will install Xena and DPR and other tools and documentation required to set up a full digital archive either on a single computer or as we do, in a larger scale data centre.

Other Approaches

Other institutions have explored, and in many cases have implemented, a range of other approaches to digital preservation. While there is likely to always be a diversity of methods for preserving digital records, we believe that we can all learn from the experiences of others.

One approach that we have observed in several institutions, is that of taking custody of digital records and storing them without any immediate preservation action. This has the advantage of being simple and quick, but also has a couple of key disadvantages. One is that if the digital data is either corrupted or encrypted when it is received, there will be no way of knowing this until an attempt is made to access a record – possibly some years in the future. By that time it is likely that the creator of the record will not be available to aid in its access.

Another disadvantage of this approach is that it's usual to set up project teams to analyse the formats being stored, and to eventually suggest a preservation action when a risk assessment indicates that something should be done soon. This involves people, time and decision making – all of which are expensive. In the digital domain, we should be looking at the ways in which we can reap the benefits of automation and to the largest extent possible, engage computers in minimising our workload.

Our process involves our software working at the time of transfer, with every individual record that we receive. This automation gives us the benefit of an instant alert in the case of a corrupted or an encrypted file and importantly makes an automated preservation decision about each record. Of

5 http://dpsp.sourceforge.net

course, computers are not perfect and it is possible that a preservation decision may turn out to be the wrong one, but we mitigate that risk by storing the original data object in our archive alongside the preserved copy.

In Retrospect

Software, Systems and the Performance Model

Looking back at the key elements of the research documented in our Green Paper, our current digital archive is the result of software and systems developed to support the Performance Model of digital preservation. In this we have succeeded. The software is in daily use and the systems are in place and we are actively preserving digital records. We are able to locate, export and 'perform' any of the digital records in our custody.

Importantly, none of this work has been outsourced. Just as we built our expertise in paper records over the decades, we have built the expertise in software and systems development in house, recognising that these skills form part of the core of any contemporary archival institution.

Essence

The other main thread of the Green Paper is that concerning the 'essence' of digital records, sometimes described as the 'essential characteristics' of a record. Our stated aim in 2002 was to produce extensive documentation defining the essence of each format of digital record. What makes a text file a text file? What makes a TIFF image a TIFF image? What makes a Word document a Word document? And so on. There has been some excellent work done around the world in this area, particularly in the UK, but it is not something that we have solved. Preserving the essence of digital records remains important to us, but we have not done it by documenting what we consider to be the essential characteristics of every digital format.

Instead, we have employed a two pronged pragmatic approach to ensuring that we capture the essence of each format that we preserve. First, during our development of Xena plug-ins, we test each conversion against a known corpus of test data. This gives us a high level of confidence in our software before it goes into production. Next, the Quality Assurance process built into the DPR gives us the opportunity to evaluate the performance of our software against the records we receive in real transfers. We think that these two overlapping assurances - off-line testing and in-process

validation - give us the confidence we need to ensure we are capturing the essence of digital records.

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, we release all of our software under the General Public License version 3 (GPL3). 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.

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 the 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 test 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. This resulted in some innovative development work to completely change the way that metadata is managed through the process and a tenfold improvement in processing speed.

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 will need to implement a secure, network based means of transfer.

Finally, although 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 is possible, and perhaps I can talk about that at the next congress.

For now, we have systems in place to secure, preserve and control digital materials so that in the future we can be sure of delivering reliable, authentic and trustworthy digital records.