A Performance Model and Process for Preserving Digital Records for Long-term Access

Andrew Wilson
National Archives of Australia
Canberra, ACT, Australia

Abstract

The National Archives of Australia, as the archives and records authority for the Government of Australia, has a requirement to ensure that high value digital records created through the business activity of Australian Government agencies are accessible indefinitely. However, indefinite preservation is extremely difficult when dealing with digital records encoded in proprietary data formats.

The National Archives of Australia's digital preservation project, which has been underway since late 2000, aims to develop a methodology for preserving digital records so they will remain accessible over time. The National Archives approach is focussed on the centrality of data formats as the key to viable long-term preservation of digital records. To implement the approach the Archives is developing or adopting a range of open data formats in XML which will be used as schema to transform original digital objects into XML data formats, a process we refer to as 'normalisation'.

The first section of the paper provides the context for the NAA digital preservation project. It will discuss the policy approach developed by the National Archives and the performance model adopted. The second half of the paper will describe briefly the preservation process being trialled at the National Archives.

Introduction

This paper deals with an approach to the preservation of digital records that has been developed by a small research and development team at the National Archives of Australia (NAA). What the paper will cover is the conceptual model developed at the National Archives that has allowed us to implement an approach to preserving digital records which we think will enable us to make them available over time as authentic, reliable and immutable archives. We call this conceptual underpinning of our approach the "essential performance" model.

This paper will discuss:

- the nature of the problem and some introductory concepts;
- the preservation process, from agency to researcher, that we are trialling at present.

Because information management specialists internationally don't always talk about the 'stuff' that is the stock of their trade in the same way or with the same meanings, I should first set out some definitions:

- records are recorded information created or received and maintained by an organisation in the transaction of its business
- digital records are records in digital form processed by computers
- digital records do NOT include computer systems or working applications.

A Performance Model for Digital Records

Digital records challenge the idea that records are essentially objects for archivists to preserve, arrange, store and make accessible. As archivists, we are very comfortable with the concept of records as paper objects, as original and unique physical artefacts. A paper record can only be experienced at one place in time. Researchers can experience paper records directly if they can read the language of the record. For archivists, the problem of preserving physical records centres on the object: once the object is preserved, the record is preserved.

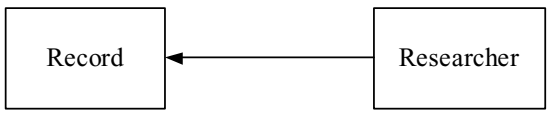

Figure 1. Traditional experience of a physical record

Digital records, while fulfilling the same general business purpose as paper records, are inherently different from their paper counterparts. The most obvious difference is that digital records are mediated by technology, which means that to experience digital records a person must have the right combination of hardware and software.

Digital records thus cease to be physical objects and are, instead, the result of the mediation of technology and data. The experience of the object only lasts for as long as the technology and data interact. As a result, each viewing of a record is a new 'original copy' of itself – two people can view the same record on their computers at the same time and will experience equivalent 'performances' of that record.

An important step, in our view, is accepting that preserving the object is meaningless. Instead, what we think needs to be concentrated on is:

- the interaction between the data itself and the technology that interprets that data;
- determining what is important about the record and the interaction that helps determine the record as an archive (that is, the essence of the record)
- creating and maintaining the ability to repeat that essence, on demand, and in a sustainable manner.

Because researchers experience the record through a 'performance' of these various components, all digital records are, in this sense, inherently different from paper records. The importance placed on originality, in relation to paper records, cannot apply to digital records, where many users can experience exactly equivalent copies. In the case of digital records, archivists should not be interested in the 'original' record but in capturing and recreating the fleeting and temporary performance of that record on the screen where it was viewed (or in other forms such as paper printouts).

The performance model breaks down the concept of a digital record into components that help explains the fundamental nature of records in digital formats. The *source* of a record is a fixed message that interacts with technology. This message does indeed provide the record's unique meaning, but by itself is meaningless to researchers, since it needs to be combined with technology in order to be rendered as its creator intended. The *process* is the technology required to render meaning from the source. When a source is combined with a process, a *performance* is created and it is this performance that provides meaning to a researcher. When the combination of source and process ends, so does its performance, only to be created anew the next time the source and process are combined. Unique combinations of processes (such as a specific computer architecture and a version of a software program) create a specific process platform. A source may be mediated by many different software platforms, and each combination of source and specific process platform may produce a slightly different performance.

More specifically, the *source* of a digital record is a data file or a recorded bitstream. This data file has a defined structure that varies according to different formats: a Microsoft Word document, a Microsoft Excel Spreadsheet, an Adobe Acrobat file and an HTML web page all use different data formats.

The *process* is the specific combination of computer hardware and software and the configuration needed to understand the file format of a source. A Word source, for example, requires the correct version of the Word application, using a Windows operating system, which is installed on a suitable Intel computer.

The *performance* is what is rendered to the screen or in any other output form, such as a paper printout or as sound.

Some archives (and a number of academics and software companies) for some time argued that it was probably possible to maintain the source and the process that created digital records.

This may have been possible with early and simple systems. Even today we can keep the source, and lots of archives and bureaux will do this for you. Indeed, for many years the National Archives of Australia stored over one million computer tapes which were the output of the search for oil in and around Australia. So, we were keeping the source of the information but when it came to providing information from the tapes we were more than embarrassed! We were not able to maintain and provide the original process.

Processes and operating systems become obsolete. For example, DOS systems transmogrified into Windows 3.1, then into Windows 95 and then into Windows NT (and 200 and XP, etc.). There is obviously a benefit to corporations that sell the software to ensure that there is marketplace obsolescence.

In addition, processes aren't interchangeable. Apple Macintosh operating systems are not compatible with Windows or Unix operating platforms, for example.

If you run Windows 3.1 you probably cannot access material created in Windows 95 or Windows NT.

Also, and most importantly, the engines that drive processes are usually proprietary. Access is governed by licence, not by purchase. And, again to protect the income stream of software corporations, software licences don't last forever.

So, the performance model for digital records, in comparison to the model for traditional records, looks like this:

Figure 2. Experience of a digital record

At NAA we looked at solutions using this concept. One possible solution involves keeping a master copy of every source we accept into custody. By doing this we can provide passive access, in which a researcher gets access to an exact copy of the original record as a series of zeros and ones, but not to the performance. If the researcher can reproduce the performance with software that allows recreation of the performance, then well and good.

But, to satisfy that majority of users who will be using current software, the Archives can undertake active intervention to recreate the performance. We can replace the source and the process and give active access to the essence of the performance.

To illustrate the point, this performance model can also be applied to audiovisual records, and in fact our experience with such records was an important foundation for our digital records performance model. In this case, the *source* is the film stock or videotape that has the image and sound recorded onto it. The *process* is the combination of projector and screen, or video and television, that is used to interpret or render the source. The *performance* is thus the collaboration of the source and process which produces the moving image and accompanying sound.

In the case of audiovisual material, the film is not generally valued as the archival record, since it is the moving image on the screen that interests researchers. Before nitrate film decays and turns to a brownish dust, conservators copy the film to a newer, more stable medium, such as polyester film. Conservators ensure that all the characteristics considered essential to the performance of the moving image are retained.

The unstable source, nitrate film, is migrated or copied to a more stable source, videotape. Converting to a new source also means changing the process from a projector and screen combination to a VCR and TV combination, which is capable of rendering the new source. While the source and process change, the performance remains equivalent. All the characteristics of the moving image from the film source that are considered important are preserved and retained on the videotape source. The researcher views an equivalent performance regardless of the source and process combination used to create the performance.

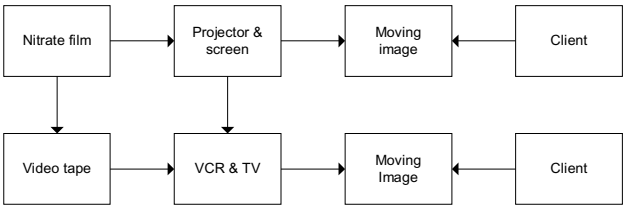

Figure 3. Example migration of film to video sources

This early experience and our analysis of the issues related to audio-visual records gave us some scope to apply similar principles to digital records.

The source object becomes far less important than the performance (unless you are a museum curator passively displaying the object). Also, you can change any of the components in the source and the process, and it doesn't matter. What does matter, however, is if the performance no longer keeps its archival value.

Problems of Digital Preservation

Although digital records are fundamentally performances and not objects, our first reaction may be to preserve both the source and process, and recreate the performance when it is required. However, just as it would be unrealistic to expect to watch an early 1900s film on nitrate film stock using a projector of the same era, it is equally unrealistic to expect to view a Word 2.0 file on an Intel 386 machine with a Windows version 3.1 operating system, even though this technology is less than 15 years old.

While preserving the source is indeed possible, preserving the process is unrealistic because of the dynamic nature of the IT industry. The industry has been rapidly expanding and developing over several decades, with huge changes in hardware and software capabilities and the infiltration of computers into work and home life. Technology cycles are short; therefore product lifetimes also tend to be short. The implications of this largely market-driven instability are two-fold: rapid decay and technological obsolescence.

Storage media, such as disks, tapes and cartridges, decay relatively rapidly compared to other media. They are not designed for long term use and are therefore extremely susceptible to short and medium term decay. The short lifetime of contemporary storage media means that a constant media refreshing program is the only way to ensure the survival of digital material.

More serious than the decay of storage media is the issue of technological obsolescence. New advances in computer science mean that both hardware technologies and software data formats are superseded over time. Furthermore, market-driven innovations mean that manufacturers update and release new systems, software applications and hardware technologies at a rapid rate. In terms of the performance model described above, the structure of the source object and the process that these structures depend on are in a constant state of development and change. As a result, without intervention by archivists to preserve the source and process, the performance cannot be guaranteed.

The problems of decay and obsolescence do not make the job of preserving digital material impossible. The performance model shows that neither the source nor the process need be retained in their original state for a future performance to be considered authentic. As long as the essential parts of the performance can be replicated over time, the source and process can be replaced.

Other Approaches to Digital Preservation

The National Archives is by no means the first institution to tackle these issues of digital preservation. Two long-term preservation approaches often advocated within the archival and library preservation communities are *migration* and *emulation*.

Migration is the process of converting a digital object from one data format to another, for example from Word v8.0 to Adobe's Portable Document Format (PDF). Generally, archivists use migration as a way of ensuring the accessibility of a digital record when the software it depends on becomes obsolete. In performance model terms, migration converts a source object from an obsolete format into a current format so that a current process (the hardware and software combination) can render the new source.

Some attributes of the digital object may be lost during the conversion process, therefore the performance may not be equivalent after migration. The level of data loss through migration depends on the number of preservation treatments applied to the record, the choice of process, the new data format, the level of human intervention and post-migration descriptive work.

Emulation is an approach which keeps the source digital object in its original data format but recreates some or all of the processes (for instance, the hardware configuration or software applications such as operating systems), enabling the performance to be recreated on current computers. An example of emulation is writing a program for a Macintosh operating system to run on a Linux operating system. Advocates of the emulation approach often maintain that the exact 'look and feel' of the record must be preserved, and that recreating the exact functionality of the original processes is the best way of doing this. The look and feel includes not only the content of the record, but also the tangible aspects of its presentation, such as colour, layout and functionality.

Both approaches have been applied to digital preservation and have been proven to work, yet both approaches have a number of limitations that must be considered carefully: sustainability, 'look and feel' and accessibility.

Migration and emulation require a large commitment in resources up-front and over a long term. Ongoing migration requires intensive cyclical work to convert objects in obsolete formats to current formats. The work increases as the digital collection grows. Emulation requires highly skilled computer programmers to write the emulator code and sophisticated strategies to deal with any intellectual property and copyright issues that may arise when emulating proprietary software. Both approaches, therefore, would place a large and perhaps unsustainable burden on an organisation the size of the National Archives, if adopted.

Both preservation methods involve decisions about how the look and feel of a digital record is to be preserved. For emulation, the aim is to ensure that as much of the original look and feel is preserved as possible. The migration method is generally based on the premise that content is more important than look and feel. This approach is reflected in the wholesale migration of digital objects from one format to another with little control over identifying or retaining look and feel elements of the original data object. Neither approach, however, has an informed, formal mechanism for capturing look and feel characteristics**.**

Migration and emulation also support different levels of accessibility to the records. Emulation, while recreating the look and feel of the original, makes access difficult for those who do not have access to an appropriate emulation environment on their local computer. Furthermore, it requires those researchers who do have access to learn the original computing environment. For example, a researcher in 2050 may have to learn commands for a DOS system to access records from the early 1990s or to recognise the 'mouse clicks on icons' for a Windows system to access records from the late 1990s! Migration, on the other hand, relies on current data formats and current processes and thus requires fewer specialised skills or software to make records accessible. Researchers can access the migrated records through the web or email.

The lessons we learned from the two preservation approaches are that:

- most of the preservation effort needs to be invested at the beginning, not in continual emulator maintenance or data conversion;
- the preservation approach should impose minimal requirements on researchers to install and learn new software applications;
- preservation treatments must be accountable through documentation available to future users of the records; and
- formal mechanisms must be created for controlling and preserving the look and feel characteristics that are considered essential to the record's meaning. The preservation of these essential characteristics cannot be left to chance.

Concept of Essence

The National Archives project team developed the concept of a record's 'essence' as a way of providing a formal mechanism for determining the characteristics that must be preserved for the record to maintain its meaning over time. The performance model demonstrates that digital records are not stable artefacts; instead they are a series of performances across time. Each performance is a combination of characteristics, some of which are incidental and some of which are essential to the meaning of the performance. The essential characteristics are what we call the 'essence' of a record and it is these essential characteristics that we will preserve and make accessible over time.

The essential characteristics of a word processing document, for instance, may include the textual content; formatting such as bolded text, font type and size; layout; bulleting; colour and embedded graphics. These characteristics are devices deployed by the creator to emphasise the message or assist with its comprehension. Since it's the message that provides evidence of business activity, this message and the characteristics of the document that qualify this message comprise the essence of the record.

The characteristics that are not essential to the meaning of a document's message are not essential to the document's meaning as a record. These might include characteristics of the application program that created the document, such as the toolbars, button functionality and colour in the user interface. Other non-essential characteristics might include the ordering of bytes in the document's data file or the specific data format of the document – since, as we have already seen, as long as the way the document was rendered can be recreated, the actual structuring of data is not essential to the record's performance.

Preserving all the characteristics of a performance can result in a large amount of resources being spent on preserving elements that are inconsequential to the record's

archival meaning. To avoid this, archivists need to determine which elements of a performance are essential for the record to retain its meaning, and to focus on preserving **them**. Identifying at the beginning what we want to preserve over time also gives us control over the preservation process – we do not need to rely on preserving only what software vendors allow us to preserve. Such a reliance would be a problem if we moved from one proprietary format to another.

We can use the example of Australian census records to illustrate the point here. The essence of the record – what the researchers want – is reliable personal information. Microfilming the original paper documents, which is what the responsible government agency does, captures the essence. From a preservation perspective, then, what we need to ensure is that the much smaller volume of the microfilm is preserved and kept accessible rather than the shelf kilometres of the paper originals.

Determining the essence of records is not a science and is open to subjectivities and archival interpretation, but it is essential to an efficient, effective and accountable preservation program. Focusing on the essence of a record allows us to clearly state our archival requirements for the preservation of that record and to be held accountable against those requirements. It means that researchers in the future can have access to the archival decisions that were made about a record's essence when it was preserved.

Preserving the Essence of Digital Records

Using the essential performance model, there are five things we need to do to be able to preserve the essence of archival records that were created in a digital format. We need to:

- Define the essence
- Acquire our own processes to suit this essence
- Normalise the source to suit our processes
- Maintain the source and processes over time and
- Recreate the performance as we need it

As part of defining the essence, there will be different approaches to different types and purposes of records. We need to ask ourselves how important the various elements are in any performance.

1. Define the Essence

For example, there are different presentations of the record which may or may not effect the meaning or value, or essence, of a record. The formatting or highlighting colour may be particular in a record, and change its meaning. Or, other embedded information may determine the evidential value or currency of a record, data such as metadata embedded in an e-mail, or formulas in spreadsheets.

But, perhaps it is just the content of the record that needs to be preserved. An assessment of the different genres of records and contexts in which they are created will determine which rules they need to operate under.

2. Acquire Our Own Processes

Experience has let us to the view that it is impractical to preserve original processes.

The cornerstone of our approach is the use of archival data formats that are non-proprietary and specifically designed for long-term access across different computer platforms. Archival data formats are formats that digital data objects are converted into for preservation purposes. For the National Archives purposes, the archival data formats we choose must also be able to carry the essence of the particular record or record type being preserved.

Within the archival and digital library communities there have been many candidate archival formats suggested over the last decade. Adobe's Portable Document Format (PDF), for instance, is often nominated as an archival format for typical office documents. PDF presents the digital record as if it were a printed page. This means that for any digital record saved to this format, its look and feel is fundamentally one of text and images designed to fit a particular page size. However, proposals to use formats such as PDF normally suppose that the entire range of preservation requirements for digital records can be satisfied by a single data format. Our experience is that this is not so. To use earlier examples, embedded e-mail headers, or active formulas in spreadsheets are not recreateable in PDF.

So, there is a need to be flexible in designing or acquiring our own processes to ensure they use open standards. The advantages of this are that the processes we use are not owned by anyone, are capable of being applied though multiple implementations, and are interchangeable.

Therefore at the moment we are saying that our standard platform of choice is XML. The idea of creating our own data formats to meet the preservation needs of many record types is not as daunting as it first seems. Mark-up language technology, and specifically XML, allows us to quickly and easily create our own non-proprietary archival formats that can preserve a record's essence.

Since the specification of the XML standard is freely available, the National Archives can create and maintain its own XML tools without dependence on a particular IT vendor and their proprietary knowledge. Our preservation program can thus use XML as its technology base indefinitely. Even if the IT industry replaces XML with another data format technology in the future, we will still be able to create our own XML tools for as long as we wish because all the information needed to construct XML tools is publicly available. Therefore, once the source objects of digital records have been converted into XML, the National Archives will not be forced to re-convert the data objects to another data format. Forced migration is avoided and preservation treatments can be minimised, thus reducing the long-term risk to digital records' integrity.

3. Normalising the Source

In our proposed preservation approach, when a digital record is transferred to the National Archives, it undergoes a single preservation treatment, called normalisation. Normalisation is the conversion of the source object from its

original data format into an XML-based archival format. The conversion work is automated by using specific software applications, called normalisers, that convert the original source object into XML. The newly created preservation master is then stored in a digital repository, along with the original transferred source object. The major difference between normalisation and many other forms of migration is that records are migrated only once into archival data formats, and do not enter into an ongoing, cyclical migration process. This is a risk assessment that we have undertaken, taking note of the view that the greater the number of migrations the greater the risk of loss of valuable data.

4. Maintain the Source and the Process

After normalisation we undertake to maintain the digital records for as long as they are required. We will maintain both the normalised source (plus the original bitstream) and the process. We place everything into a digital repository, maintained as just one component of a physical archival repository. Backups are also stored, and we will refresh the media at regular intervals.

We understand that technology will change and we will need to acquire new preservation processing platforms. Depending on how robust the market is, we can in future select a vendor product to manage the preservation processing platform, or we can build our own.

A principle remains, however, that researchers will never use a master copy.

5. Recreate the Performance

From the digital repository we need to recreate the performance of the record. But, the test of what we provide will be the same as for any other preservation program: that what we give to the researcher has to be both readily accessible, and authentic.

There are two ways we do this: we can make copies of the preservation master copies, so that these zeros and ones can be used to recreate the performance by use of a relevant browser, if that's what the researcher wants. Or, the most common form of access will be by providing researchers with the XML data object and access to an appropriate viewing software application (such as a browser or viewer application developed by the project team) to recreate the digital record's performance.

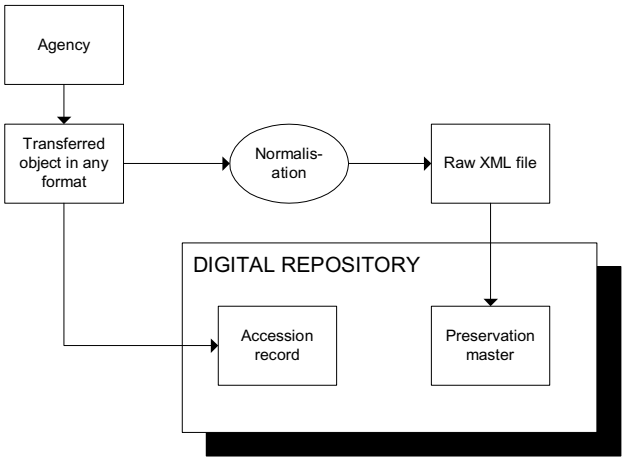

Figure 4. Normalisation Process

Conclusion

The challenges of digital preservation affect all major public archives, both in Australia and around the world. In the same way that the National Archives has developed recordkeeping standards to address government recordkeeping issues, it is committed to providing innovative solutions to the problems of digital preservation. We expect that the approach documented here will be of value to many other archival institutions within Australia and overseas. Ultimately, this work will preserve millions of Government records that will be of significant value to future generations.

Biography

Andrew Wilson is a Director in the Digital Government Branch of the National Archives of Australia. He is currently managing a project to develop the capability of the National Archives to manage digital records from transfer to access. Prior to that he managed the project developing an approach to the long-term preservation of digital records. Andrew is currently a member of the Dublin Core Usage Board and co-chair of the DC Preservation working group. He also represents the National Archives on the RLG/OCLC PREMIS working group.