An Approach to the Preservation of Digital Records

Helen Heslop

Simon Davis

Andrew Wilson

December 2002

CONTENTS

INTRODUCTION

Background

The National Archives of Australia, in partnership with Commonwealth agencies, aims to ensure that full and accurate records are created and managed to support the business of the Australian Government and the interests of the wider Australian community. To help achieve this aim, the National Archives has developed a comprehensive range of policies, standards and guidelines under the collective title *e-permanence*. This approach to recordkeeping is based on the International Standard for Records Management (AS) ISO 15489.

The federal Government has a significant impact on the course of Australian history and the lives of individual Australians. The records of the Government are the principal means by which its actions and decisions can be studied and understood, now and into the future. The most important and the most interesting Commonwealth records are selected for preservation as part of the national archives, and the availability and quality of these archives depends on the quality of recordkeeping in Australian government agencies.

Over the last decade or so the rate of change in government administration has increased. The entry of computer systems into the work environment of organisations over the last two decades has dramatically altered the way in which employees work, communicate and share information. Email systems enable staff to communicate with immediacy and convenience. Shared folders on intranet systems mean that different workers can quickly and easily access and update the same copies of information resources over time and space. Database systems allow for the quick and easy retrieval of vast amounts of information. Corporate websites make organisations visible and allow them to communicate a diverse range of information to stakeholders, clients, customers and other interested people throughout the world.

These changes have made good recordkeeping both more difficult and more significant. For many years lack of attention to recordkeeping has been mitigated by the existence of long-standing, well known practices for the use of paper records. Paper records also have a robustness that enables them to survive long periods of neglect. In contrast, the sometimes haphazard use of electronic systems for communicating and storing recorded information is more fragile.

Although electronic systems offer many advantages, agencies must ensure that these records are captured, survive as long as they are needed, and can be read and understood. For example, important email messages must be captured into corporate recordkeeping systems where they can be preserved securely and found easily. Databases containing case records with long-term

value need to be migrated forward with hardware and software changes so that the records are still accessible. In addition, the move to doing government business online relies on the creation of accurate records to ensure that online business is conducted efficiently. Such records provide evidence that the transaction occurred and essential details about it. Nonexistent or poor quality records will prevent online business being conducted successfully and may involve loss of revenue and losses in court. Preservation of records that are 'born digital', ie. that come into being as electronic (digital) records, is an essential part of agency management of their records and should not be viewed as unnecessary. Agencies need to continue to manage records, regardless of format, that are critical to their day-to-day business. In order to help agencies manage digital records over time the National Archives commenced a project in 2000 to identify a viable approach to the long-term preservation of digital records.

This paper outlines the Archives' approach to digital preservation that is being developed as part of the Agency to Researcher Digital Preservation Project. The project is called 'Agency to Researcher' because it deals with digital records from the time they are transferred from an agency to the time they are used by the researcher. This paper is primarily technical in nature but explains the important concepts underlying the NAA approach to digital preservation in non-technical language.

Agency to Researcher Project

The introduction of computer systems has brought about an information revolution, and has resulted in a significant shift in the ways that Commonwealth agencies create and manage their records. One of the ways that the National Archives of Australia is responding to the challenges mentioned above is by changing its preservation methods to cope with a new breed of archival record. Over the years, the National Archives has developed a great deal of expertise in the preservation of paper records. It is now developing the same level of expertise for the preservation of digital records.

The Agency to Researcher Project also represents a significant shift in the National Archives' approach to digital preservation. For most of the 1990s, the Archives followed a distributed custody model, which meant that digital records of archival (ie. long-term) value remained in agency custody except in special circumstances. In March 2000, the Archives moved to a custodial model, which means it now takes into custody all digital records that are required to be retained as National Archives under approved disposal authorities. The Agency to Researcher Digital Preservation Project was initiated as a result of this policy change to ensure that digital records of long-term value will remain accessible for use over time. The project team is located within the Preservation Section of the Collection Management Branch of the National Archives.

Audience and Scope

This paper is primarily aimed at anyone with an interest in approaches to preserving digital records. The approach described in the paper is principally intended to be applied to digital records required to be retained as National Archives. However, there is no reason why agencies cannot adopt components of the approach to ensure that digital records required by agencies for longer-term business needs remain accessible for as long as they are needed.

The project team began by surveying approaches to digital preservation adopted by other archival and custodial organisations throughout the world. We then developed the performance model to help clarify our understanding of the fundamental nature of digital records and to explain it to archivists and conservators using familiar examples. We also developed the concept of the 'essence' of a record – those characteristics of the context, rendition and structure of a digital record that must be preserved together with the content to give the record meaning – which forms one of the foundation stones of the approach.

Also underpinning the approach are principles developed to ensure that the performance model supports the Australian Government policy of comprehensive, equitable and sustainable access to the Commonwealth's archival resources.

The first four sections of the paper provide the basis for the principles that underpin our approach. These principles are outlined in section 5. Section 6 explains our use of XML-based archival data formats and the general preservation process we are developing. Section 7 summarises the key products the Agency to Researcher Project must deliver and scopes out the work the project team expects to complete in the next two years. This work includes the development of software applications and further development and testing of the approach, which will ultimately form the National Archives digital preservation program.

The National Archives would welcome feedback on both this paper and the approach we are developing for long-term preservation of digital records. Comments, suggestions and queries may be sent by email to andreww@naa.gov.au, or by post to:

Andrew Wilson
Project Manager
Digital Preservation Project
National Archives of Australia
Box 7425
Canberra Mail Centre ACT 2610
Australia

1. Performance model and fundamental nature of 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. For researchers, paper records can be experienced directly if they can read the language (see figure 1). For archivists, the problem of preserving records centres on the object: once the object is preserved, the record is preserved.

Record	←	Researcher

Figure 1: Direct experience of a paper record

Digital records, while fulfilling the same general business purpose as paper records (reports, letters, memos), 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.

The importance placed on originality, in relation to paper records, does not apply to digital records, where many users can experience equivalent copies. In the case of digital records, archivists are not 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.

The performance model breaks down the concept of a digital record into components that help explain their fundamental nature. The *source* of a record is a fixed message that interacts with technology. This message provides 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.
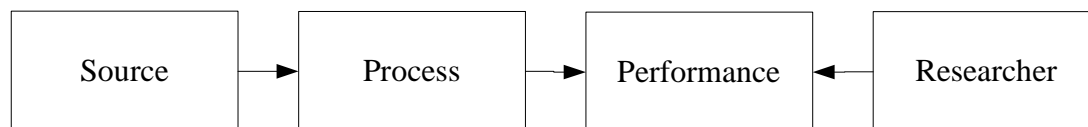
| Source | → | Process | → | Performance | ← | Researcher |

Figure 2: Performance model – source and process components

More specifically, the *source* of a digital record is a data file. 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 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 to any other output device.

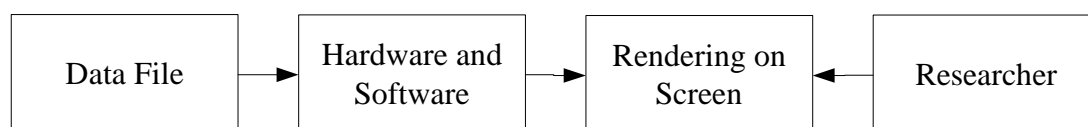| Data File | → | Hardware and Software | → | Rendering on Screen | ← | Researcher |

Figure 3: Performance model – digital records

The performance model can also be applied to audiovisual records. 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 the collaboration of the source and process: the moving image.

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.

Figure 4 shows the migration of film to tape in terms of the performance model. 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.
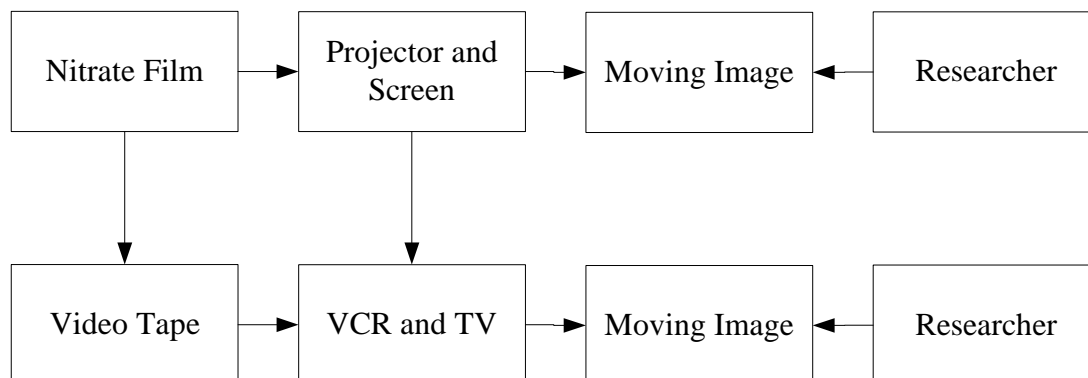

Figure 4: Migration of film to video sources

2. Problems of digital preservation

Although digital records are fundamentally performances and not objects, archivists' 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, 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.

3. 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 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.

4. Concept of essence

The 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.

For instance, the essential characteristics of a word processing document 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 agency 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.

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.

5. Principles

Taking into account the lessons learned from other digital preservation approaches, the project team developed a set of principles, which underpin the proposed approach. These principles are based on the concept of the fundamental nature of digital records as provided by the performance model. They are designed to ensure that the Archives' digital preservation program is consistent with its values of comprehensive, equitable and sustainable access to the Commonwealth's archival resources.

1. ***The digital preservation program must be able to preserve any digital record that is brought into National Archives' custody regardless of the application or system it is from or data format it is stored in.***

Digital records gain their archival value from the business context in which they are created and used, not from the recordkeeping processes designed to manage them. High quality recordkeeping makes it easier to identify archival material and thus increases the likelihood of such material coming into the National Archives' custody. However, for the digital preservation program to be comprehensive it must be able to preserve records that come from all types

of recordkeeping environments, which may vary according to how well they meet recordkeeping standards.

This principle means that our digital preservation approach cannot assume that all archival digital records will be accompanied by appropriate metadata, are well controlled, or use common, well-understood data formats. To meet the expectations of the National Archives, and of researchers, the program must be able to preserve any record that the National Archives has selected as a national archive, without exception.

2. The digital preservation program must determine and preserve the essence of the digital records in the National Archives' custody and recreate their essential performance over time.

Digital records are varied and complex. They do not all have the same essential characteristics, or 'essence', that need to be preserved over time. With this in mind, determining the essence of a particular genre or type of record, such as a word-processed document or email, before the application of any preservation treatment is a very important way to ensure optimal, sustainable and accountable preservation. This analysis work is fundamental, because we must be able to recreate the equivalent performance to ensure the stability of the record's meaning over time.

This principle means that archivists in the National Archives digital preservation program will need to spend time analysing genres of records in our custody to identify and document their essence. Any preservation techniques developed as part of the approach must ensure that the essential characteristics remain accessible to researchers over time.

3. The digital preservation program will be based on non-proprietary technologies.

Proprietary data formats are unsuitable for long-term preservation and accessibility of digital records, particularly for an organisation committed to free long-term access to digital records.

The IT industry is dominated by organisations that invest heavily in new product development and seek to recoup that investment by energetically protecting their intellectual property rights over their products. As a result, many of the information technologies used to create archival digital records are proprietary. Licences from the intellectual property owners are needed to use software applications, hardware components and to structure source objects. Indefinite access to technologies is not a standard condition of the licences, meaning that IT vendors can change licence conditions or withdraw products from the market without consultation in order to support other aspects of their business. Access to digital records in proprietary formats is

ultimately in the hands of the intellectual property holder, not the National Archives.

Viable alternatives to many proprietary technologies, especially software technologies, are becoming available in the public domain or through open licences. Unlike proprietary technologies, these technologies would be available to the National Archives indefinitely. By using non-proprietary technologies, access to digital records in our custody would remain in the Archives' control and would not be dependent on a company's intellectual property rights. Other advantages of using open licence applications are that the financial cost of development can be spread between like-minded institutions, ideas and innovations can flow freely and other institutions may be encouraged to take up a similar approach and technology.

This principle means that, as far as possible, the National Archives will retain control over the technologies it relies upon for digital preservation by avoiding the use of proprietary technologies. In cases where there are no public domain or open licence alternatives, such as for many hardware technologies, the National Archives will favour proprietary technology that can be sourced from multiple vendors.

4. *To lessen the risk to the integrity of the records, the preservation program will minimise the number of preservation treatments applied to each digital record.*

Applying a preservation treatment to any type of archival record is both expensive and potentially harmful to the integrity of the record. Although the per record cost of treating digital materials may be less than for other record types, when using a batch treatment process (such as a mass migration from one data format to another), the risk to the integrity of a digital record posed by preservation treatment is extremely high as digital materials can be changed with little or no trace. Continual digital preservation treatments – such as regular, short-term migration – pose a great risk to the preservation of a digital record's essence. Testing that a digital record's performances still carry the record's essence after a digital preservation treatment is likely to be a costly and inexact activity.

Our digital preservation program, therefore, will minimise the number of preservation treatments applied to each digital record. Where a preservation treatment is necessary, we will always attempt to apply the treatment that will last the longest. All our treatments will be accompanied by full documentation.

5. *The digital preservation program will not limit the accessibility choices of the National Archives or of future researchers.*

No preservation approach would consciously set out to restrict accessibility choices for the users of the records. However, decisions made early in the development of any approach can greatly affect the flexibility of the access given in years or decades time. For example, the emulation approach requires access to the emulator applications and sometimes sophisticated knowledge of how to use them. Reliance on proprietary archival data formats means that a user must own or have access to a proprietary application in order to display the record. The applications themselves may not be free, which means a user would have to buy a licence to use the application before being able to read the record.

This principle means that the approach we develop must allow for equitable and sustainable access to preserved digital records. In part, this can be done through the use of open source software applications, which are available indefinitely to the National Archives and can be made downloadable to users. Also, the creation and use of platform-independent technologies means that users can access the records from any computer platform.

6. Approach

The use of archival data formats is gaining popularity in digital preservation practice. The National Archives' approach is based on the use of XML (eXtensible Mark-up Language) document formats as archival data formats (see appendix for more information on XML). The following sections discuss the use of XML as an archival data format, and outline the preservation process as a whole and its two core sub-processes of normalisation and transformation.

Archival data formats

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.

A word-processed document contains not only the visible text but also embedded metadata that may include information about document authorship or revision changes. In some business contexts, such embedded information may be crucial to the document's ability to act as a record; in other contexts, embedded metadata may be relatively trivial to the meaning of the document as a record. An archival format, such as PDF, that preserves only the visible characteristics of a document and not the embedded metadata may be suitable in some cases but completely unsuitable in others.

By focusing on records as performances and preserving the essence of these performances, we can make decisions about which characteristics, such as the visible text and metadata, are important and must preserved, and which characteristics do not need to be preserved. Once the essence is determined, we can then develop or select an archival data format to suit the preservation requirements. This may mean we must create our own data formats to suit the particular circumstances of unique records.

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. XML is our preferred archival data format.

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.

Preservation process

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 (see figure 5). 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.
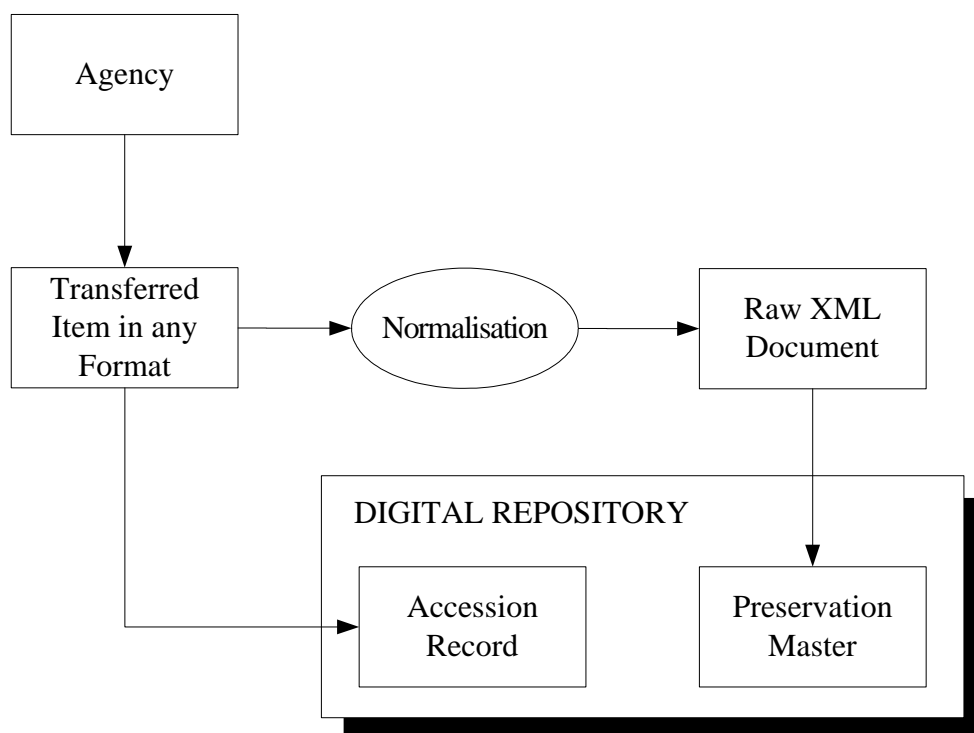

Figure 5: Normalisation process

The proposed preservation process involves two major processes: *normalisation* to convert the original source object into XML, and *transformation* to convert the XML into an accessible format.

The project team has started work on a prototype of a combined XML normaliser and viewer. The prototype can accept source objects in certain formats, normalise them into XML, then recreate the digital record's performance in a self-contained viewer. The prototype is being written in Java language and is designed as a cross-platform application able to operate (as a minimum) on Windows, MacOS, Linux and Solaris platforms.

Once digital records have been preserved and are in the open access period, they can be made available to researchers. 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. This process is called transformation (see figure 6).
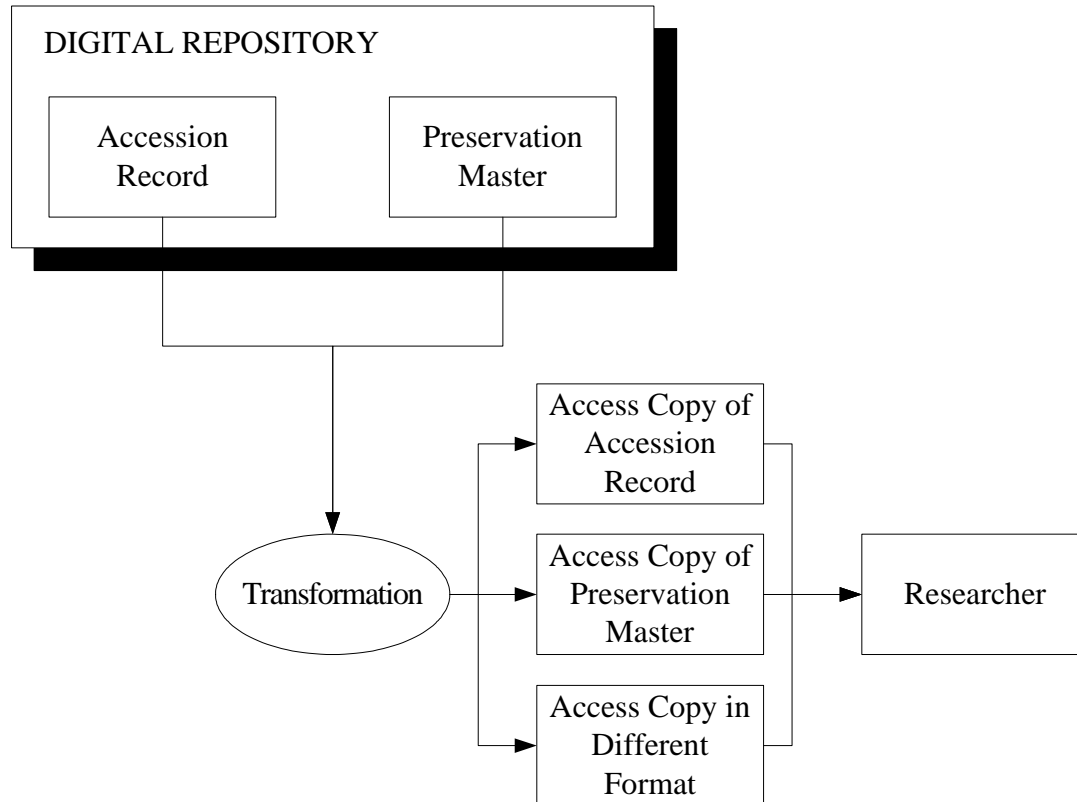
Figure 6: Transformation sub-processes

As both computing and service environments change, so will the opportunities for the National Archives to develop other types of access based on the transformation process. For instance, in the future the National Archives may want to offer a voice access service to digital records, where researchers could have a computerised voice read a record to them instead of viewing the record on a screen. Although we can speculate on the access services of the future, access will ultimately depend on the archivists and researchers of the time determining the best option. To ensure they will be able to do so, it is essential that the National Archives maintain the ability to transform the records from XML into other formats as needed. This means that the National Archives must maintain its XML competencies and XML software applications over time.

In addition to providing researchers with access to the transformed XML access copy the National Archives will also provide access to a copy of the transferred source object preserved in its original form. This form of access is called *passive access*, because we cannot guarantee that researchers will be able to recreate an accessible performance from the transferred source. As discussed earlier, the hardware and software required to read the original source object may be obsolete or unobtainable by the time a researcher requests the source object. However, for some sophisticated researchers this may be acceptable, since they will use their own software and hardware tools to process the source.

Preserving the transferred source object has another important purpose. It allows for the normalisation process to be repeated and verified at any later time. The repeatability of the normalisation process is a key aspect in ensuring the overall authenticity of the digital record. If, sometime in the future, researchers doubt the validity of the XML-based digital record presented by the National Archives, the transferred source object can be used to repeat the normalisation to satisfy their doubts.

7. The way ahead

The Agency to Researcher Project is a multi-million dollar project that will continue until mid 2004. The project team will use the concepts and approach outlined in this paper to develop a sustainable digital preservation program at the National Archives. To do this, we will need to develop a number of studies and major products, as follows:

1. A set of *research studies* that will inform our overall digital preservation approach. As a way of testing our assumptions and to fully understand the environment we are working in, we have already scheduled reports on:

    *   researcher expectations of preserved digital records;

    *   the production of digital records in the Commonwealth;

    *   digital authenticity issues; and

    *   pilots to preserve transferred records.

    These studies will result in a 'white' paper that expands upon the concepts and approach outlined in this paper. The studies are expected to take about 12 months to complete.

2. The *preservation software platform*, which contains the XML software tools and other computer programs designed to carry out the normalisation and transformation processes described above. The project team is currently developing a prototype of a combined XML normaliser and viewer. All components of the preservation software platform will be released under

an open source licence. This means that other archival institutions can not only use components of the platform but can also modify the components as they see fit. Work on the platform will continue for the duration of the project, and beyond.

3. The *digital repository*, which is the system that will manage the long-term storage of digital records within the National Archives. Closely associated with the digital repository is the *workbench* which consists of the network, and associated tools, in which the digital repository and the preservation software platform operate.

4. A *digital preservation organisation,* which will take over the operational work of the digital preservation program once the Agency to Researcher Project has ended.

The project team will test the preservation process through specially selected pilot transfers of digital records from Commonwealth agencies and personal records depositors. Some 20 series of digital records transferred into the National Archives custody during the 1970s and 1980s will also be preserved using the new approach.

The National Archives will also release the first version of the prototype XML normaliser/viewer application. The archival community will be invited to test and use it as well as to contribute to the continued development of the application by writing new normaliser and viewer modules.

8. Conclusion

The challenges of digital preservation affect all major public archives, both in Australia and around the world. In the same way that the 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 solutions developed will be of value to many other archival institutions within Australia and overseas. Ultimately, this work will preserve millions of Commonwealth records that will be of significant value to future generations.

Appendix: What is XML?

XML is a variant of Standard Generalised Markup Language, an international standard for structuring digital documents ratified by the ISO in 1986.

XML is not so much a data format or 'language' as a set of universal rules for describing data and documents. It does this by providing *elements* that identify unique sections of data within a digital document. These elements are separated from each other by start and end *tags*. Each element can also have attributes associated with it that provide further context to the element's enclosed data.

XML is an open standard maintained by the World Wide Web Consortium (W3C). The W3C is the standards-setting organisation for web data standards such as HTML (HyperText Markup Language, the document format used to encode web pages). The W3C specially developed XML to be cross-platform, so that a document structured using XML can be read on a wide variety of computing platforms (including all Windows platforms, MacOS, and all variants of UNIX and Linux) using a wide variety of software tools. Many recent browsers, such as the latest versions of Microsoft Internet Explorer and Netscape Navigator, can read and display XML documents.

Although XML sets out the rules for creating elements and attributes (as well as other aspects of document structure), it does not set out what particular elements or attributes should exist in any particular document. Instead, archivists can use the XML rules to construct their own document types that meet the particular needs of a digital record type.

For instance, if we needed to preserve a digital address book (say, from an email system), we may construct an XML document type that consists of a contact element that contains elements for name and email. The XML version of such an address book may look like this:

```
<addressbook>
<contact>
<name> Simon Davis</name>
<email>simond@naa.gov.au</email>
</contact>
<contact>
<name>Helen Heslop</name>
<email>helenh@naa.gov.au</email>
</contact>
</addressbook>
```

The text enclosed within the angle brackets (like <name> and </contact>) are start and end tags that separate the data (like 'Simon Davis' and 'simond@naa.gov.au') from each other. Individual pieces of data within the address book, such as a specific name or email, can be easily and clearly identified and processed by an XML-aware software application.

If we had to preserve a minute, the elements we created for the address book may be meaningless. XML lets us define new elements to meet the requirements of specific document types. We could create a different set of tags so that a minute was described like this:

<minute>
 <to>Stephen Ellis</to>
 <from>Simon Davis</from>
 <subject>Project progress</subject>
 <paragraph>The Agency to Researcher Digital Preservation Project is still on schedule.</paragraph>
 <paragraph>I will release a detailed highlight report next week.</paragraph>
</minute>

Instead of elements like <contact> or <email>, we now have elements that support the essence of an inter-office minute: to, from, subject, and multiple paragraphs. Note that even though the elements describe the data differently, the start and end tags follow the same basic syntax rules. Any XML-aware program that is capable of understanding element tags will be able to differentiate between the data in the 'to' element ('Stephen Ellis') and the data in the 'subject' element ('Project progress').

This is the true power of XML: different document structures can be easily created while the basic syntax rules remain unchanged. Applications that follow the XML standard can therefore read many different document types without special programming for each document type.

For the display of an XML document by an XML-aware software application (browser or viewer application), the browser must rely on an additional set of instructions to tell it how to format each tag set. Unlike HTML where the display is built into the tags (ie: <b> = bold text), XML separates the content and the formatting. In the minute example, a style-sheet may indicate that all the content in the <to> tag must be Times New Roman font, 12 point size and bolded, while the <paragraph> may be Arial font and 10 point size.

The flexibility of XML means that an archivist can tailor the style-sheet to capture the essence or 'look and feel' of any document.