



Digital Preservation Testbed White Paper

Migration: Context and Current Status

The Digital Preservation Testbed was founded by the National Archives and the Ministry of the Interior and Kingdom Relations. It was established in October 2000 to research different methods of digital preservation over the long term. The Digital Preservation Testbed is part of the ICTU, a government initiative which houses different research projects concerned with varying aspects of E-government.

ICTU
Nieuwe Duinweg 24-26
2587 AD Den Haag

Tel. 070 888 77 77
Fax: 070 888 78 80

Email testbed@ictu.nl
www.digitaleduurzaamheid.nl

Publication of Digital Preservation Testbed White Paper *Migration: Context and Current Status*
The Hague, December 5th 2001

Contents

1

Introduction 4

- 1.1 *Digital Preservation Defined 4*
- 1.2 *The Task of Achieving Digital Preservation 5*
- 1.3 *Different Approaches 6*

2

Migration 9

- 2.1 *Definitions 9*
- 2.2 *Issues and Concerns 10*

3

Migration In Practice 13

- 3.1 *Current Research and Knowledge 13*
- 3.2 *Frameworks and General Guidelines 15*
- 3.3 *Current Use 16*
- 3.4 *Conclusion 17*

4

Bibliography 18

5

Websites 21

1

Introduction

Digital Records are fragile. The debate as to the best means of preserving digital records over the long term has been underway for many years and will no doubt continue for years to come¹. Various theoretical solutions have been proposed, and research is currently underway around the world to identify ways in which digital records can be authentically maintained whilst remaining accessible and usable over the long term. This paper focuses on Migration, currently the most widely used approach. We place migration in context with contemporary thinking and practice about digital preservation, identify the issues involved, and provide a summary of current knowledge and research into migration.

1.1

Digital Preservation Defined

Digital Preservation is concerned with ensuring that records which are created electronically using today's computer systems and applications, will remain available, usable, and authentic in ten to one hundred years time, when the applications and systems which were used to create and interpret the record will, more likely than not, no longer be available. Digital preservation consists of preserving more than just the record's bit stream. We must also be able to *interpret* the bit stream in order for the *record* to survive. Without interpretation, the bit stream is nothing more than a meaningless series of 0's and 1's. During preservation, questions of record context, content, structure, appearance, and behaviour must also be taken into account. Archivists are well versed in dealing with the first three elements, but appearance and behaviour are additional aspects that are peculiar to digital records. These may therefore require the most attention to authentically preserve the record over the long term.

There is a wide range of digital formats available and, to make matters more complicated, different digital objects have different preservation requirements. These can depend on the reason the record is being preserved, how long it needs to be preserved, the context and history of the record, and its original format². Digital Preservation does not mean the same thing for each digital object. Whilst it is often considered that digital preservation means preserving the object so that it is identical to its original format, this is not always required. It is not always necessary to preserve every aspect of a digital record, and thus research is underway to define the essential aspects of records and their authenticity requirements. In all cases, however, the record must be preserved so that it retains its integrity and is authentic and usable. This presents interesting challenges .

¹ The phrase long term can mean fifty years or more, as indicated by Bennett in *A Framework of Data Types and Formats, and Issues Affecting the Long Term Preservation Of Digital Material*(1997). This appears to us to be a reasonable amount of time for which to prepare a long term preservation strategy. Technological changes after fifty years may well exceed expectations and limit the validity of a well-designed strategy. However, we must bear in mind that Dutch National Archives Regulations, specifically article 11 (1995), speak of a time scale of at least one hundred years.

² Dutch regulations (article 8) stipulates that what needs to be preserved depends on the requirements of the working process to which the record belongs.

Digital preservation is not only an archival concern. Libraries and other institutions that are required to retain data for long periods of time in a digital form are also currently tackling the problem, albeit from a slightly different angle. The various communities involved have much to learn from each other, and co-operation and contact among them has so far proved to be valuable. Indeed, a large portion of this paper relies on work produced in other sectors, including libraries and museums.

This paper stems from the work of the Digital Preservation Testbed (DPT) and the scope of this paper is limited to the scope of the DPT. Our research focuses on four archival record types – text documents, spreadsheets, e-mail messages, and databases – and this paper is restricted to the preservation of these record types for archival reasons. Digital Preservation within the context of this paper does not include the preservation of artefacts through digitisation or digital imaging. Also out of scope for the project are the migration of storage media, advanced multi-media objects, and the migration of legacy systems.

1.2

The Task of Achieving Digital Preservation

There is a difference between paper and digital records. Any paper record can be perceived through the five human senses; no digital record can be perceived without going through computer hardware and software. For this reason, the speed of technological obsolescence makes digital preservation an important issue for everyone.

Digital records are software dependant. They rely upon the software that was originally intended to interpret (or display) them. When that software becomes obsolete, perhaps within the space of a few years³, the problem arises of how to read that record without its original software application. It is unlikely that different versions of the application will read the file in the same way, and this may well result in a change in the interpreted record (the visible or available view of the file) that affects its archival integrity. Some data may be lost altogether; in other areas, data may be gained. There may be no way to compare a new version with the original, so changes may go unnoticed. Any changes to the record may affect its authenticity and integrity, which in turn may affect its archival and legal status. Depending on the nature of the record and its use, this can cause problems, not least that of losing or misrepresenting history.

Even a simple office computer system uses several different software applications. For each application, there may be several software manufacturers offering their own products. The rate at which new versions of software are released, with extended and (not necessarily backward compatible) new features, adds to the problem. Take Microsoft's Word® application as an example. The past six years have seen 4 versions released: Word 95; Word 97; Word 2000; and Word 2002. Two other versions of Word have also been produced for non-Windows Operating Systems – Word 98 Special Edition for Apple and IMac, and Word 2001 for the Mac. There are also different releases of Word within these versions. These releases have fewer

³ Gail Hodge, in *Best Practices for Digital Archiving* (1999), states that “new releases of databases, spreadsheets, and word processors can be expected at least every two to three years, with patches and minor updates released more often”. The Public Record Office at Kew state that it would be unusual if migration occurred more frequently than every three years, in their *Guidelines on the Management, Appraisal and Preservation of Electronic Record* (1999).

differences among them, but each has the potential to affect a record's integrity or authenticity.

There are already many examples of how quickly digital records and data can become inaccessible. The specific details concerning who sent the first e-mail communication in the 1960's are no longer available. Some records from the old East German Republic have been lost forever, through technological obsolescence. A recent communiqué on the Joint Information Systems Committee (JISC) listserv revealed articles describing NASA's loss of data from the Viking probes sent to Mars in the mid 1970's⁴.

There are several strategies for digital preservation. The following section provides a brief analysis of seven preservation approaches.

1.3

Different Approaches

The main preservation strategies are: *technology preservation; printing to paper; emulation; encapsulation; virtual machine software; XML; storage in standard formats; and migration*. These strategies have different technical requirements and costs. They also have different preservation metadata requirements.

1. *Technology Preservation*. One of the first options to be used was to preserve the technology required to access original records for as long as those records are required. However, this is costly and technologically complex (although in practice, some large corporations continue to employ this approach). Support for the software and hardware eventually ceases and the parts required to maintain the hardware become more and more scarce as manufacturers discontinue obsolete components. The number of machines available that are capable of reading old files continues to decrease, for computers do not last forever. The skills required to operate the hardware and software also become rare and eventually disappear.
2. *Printing to Paper*. This is another of the early approaches which is also still in practice. However, printing all records to paper is not a viable preservation method for the majority of records. Printing to paper loses functional or behavioural traits that the records had in their digital form. Certain information may also be lost. Embedded formulas in a spreadsheet, for example, will not print to paper. Databases were simply not designed to be printed out, and any printed version is only a selective view of the database, and not a preserved format.

Legal rulings have worked both for and against printing to paper⁵. As the NLA notes, 'flat data', such as text and some still images, can be printed to paper without loss of data but with some possible loss of functionality⁶. Printing to

⁴ JISC listserv, Friday 3rd August 2001, *A nice case study for digital preservation*.

⁵ The saga concerning NARA's GRS20 went on for many years, with the Judge initially ruling against GRS20, stating that an email document is not the same as a paper document, and 'that hard copy printouts of an email may omit important parts of the electronic version'. However, this ruling was later overturned by the Court of Appeal in favour of the Archivist.

⁶ National Library of Australia *Draft Research Agenda* 1998, p2.

paper is often employed as an interim approach to preservation whilst a digital solution is sought.

3. *Emulation.* The theory behind Emulation is that the only way to ensure the authenticity and integrity of the record over the long term is to continue to provide access to it in its original environment, i.e., its original operating system and software application. This can be done by preserving not only the record, but also an emulator specification, which contains enough details about the original environment for that environment to be recreated on a future computer when necessary.

Some people believe that emulation is too complicated with too great a potential for error. There is no guarantee that we will be able to recreate the full computing environment of the record on future computers, as we do not know what the future computers will be like. However, Emulation has been explored in other fields with some success, and it may be the only way to maintain complex databases or multi media objects.

4. *Encapsulation.* In contrast to the migration approach, the encapsulation approach retains the record in its original form, but encapsulates it with a set of instructions on how the original should be interpreted. This would need to be a detailed formal description of the file format and what the information means. This encapsulating layer could be expressed using XML, for example. If the original software used to interpret the data file is complex, then the description must also be complex and care would need to be taken to ensure that it was sufficiently complete. An extension to this idea is to create this description with an executable program: that is the subject of the section "Virtual Machine Approach".
5. *Virtual Machine Software.* A variant of the emulation approach has been proposed by Raymond Lorie of IBM⁷. This addresses the problem of interpreting data files in the future by writing a program to carry out this interpretation in the machine language of a "Universal Virtual Computer" (UVC). This program would be written at the time the record was archived and would be preserved together with the record. This program runs on what Lorie calls a UVC Interpreter, i.e. a virtual machine. In order to interpret the record on a future computer, a UVC Interpreter would be required and this could be produced from the specifications of the UVC. This approach is similar in principle to that used by the Java™ platform to achieve present day interoperability of Java programs. To make this efficient and achievable, the key features of the proposed UVC language are that it should be simple enough that it is relatively straightforward to produce the future virtual machines, and it should be general enough that it can be widely used for archiving purposes, so that it is cost-effective to produce the future virtual machines. With this approach, the data can be stored in any format and the knowledge required to decode it is encapsulated in the UVC program.

The approach can be extended to apply to the archiving of a program: this is more like the full emulation approach. It allows the emulator to be written in

⁷ Raymond A. Lorie, *Long Term Preservation of Digital Information* (2000)

the UVC language at the time of archiving, without requiring any knowledge of the future target machine.

6. *XML*. XML stands for eXtensible Markup Language. It is a text-based markup language for describing the structure and meaning of data. Because it is text-based, it is human readable, but it is designed primarily to be easy to process using computers. It is an open standard defined by the World Wide Web Consortium and is not tied to any particular type of hardware or operating system. Conversion of records to XML format can be seen as a particular type of migration approach, as discussed in more detail below. However, it is often regarded as the most promising present day data format for archiving and interoperability and so deserves to be considered as an approach in its own right.

There are a variety of ways in which XML could be used in electronic archiving. XML could be particularly useful in storing metadata and linking that metadata to the data files making up a record. The XML Stylesheet Language (XSL) is a part of the XML standard and is a way of defining the appearance of an XML document. The combination of XML and XSL is a promising method for defining both the content and appearance of document-based records. This is one aspect of the 'storage in standard formats' approach, included here with Migration.

7. *Migration* (including *Storage in standard formats*). As recent reports indicate, this is the most familiar and most widely-implemented preservation approach⁸. Migration is the focus of this paper and is now discussed in more detail.

⁸ The InterPARES report *Preservation Strategies for Electronic Records, Round 1 (2000-2001) Where We Are Now: Obliquity and Squint?* (2001) features the results of a 2000-2001 survey of recordkeeping institutions, in which 4/13 projects identified migration as their preservation strategy. This was the most prevalent approach. See also Margaret Hedstrom, *Digital Preservation: Problems and Prospects* (2001); also Jeff Rothenberg and Tora Bikson, *Digital Preservation: Carrying Authentic, Understandable and Usable Records through Time* (1999).

2

Migration

2.1

Definitions

The most widely cited definition of Migration is from the 1996 report of the Task Force on the Archiving of Digital Information:

“Migration is a set of organised tasks designed to achieve the periodic transfer of digital materials from one hardware/software configuration to another, or from one generation of computer technology to a subsequent generation”⁹.

This is a widely accepted definition. However, it is also rather broad. Different groups and individuals have proposed different technical approaches to a migration strategy. The choice of approach largely depends on the file and its preservation requirements. The Task Force proposed five basic ways of applying a migration strategy:

- Change the media on which the file is stored
- Change the file format itself
- Incorporate standards into a preservation strategy
- Build migration paths
- Use processing centres to do it for you.

Margaret Hedstrom, a member of the Taskforce, later broke this down into eight more specific categories, including options to transfer to paper or microfilm, create surrogates, store in a software independent format, or store in more than one format¹⁰.

A more recent publication by the National Preservation Office breaks migration down into four different strands:

- Change media
- Backward compatibility
- Interoperability
- Conversion to standard formats¹¹.

There are many ways in which each strand can be employed. *Change media* refers not only CD ROMs and magnetic tapes, but also paper and microfilm. It is often necessary to change the file medium in conjunction with other preservation action to ensure long-term interpretability of the bit stream. *Backward compatibility* relies on the

⁹ CPA/RLG, *Preserving Digital Information: Report of the Taskforce on Archiving of Digital Information* (1996).

¹⁰ Margaret Hedstrom, *Draft Section of a Report on Migration Strategies* (1996, revised May 1997). However, in actuality, there are no truly software independent formats on computers, it is merely that some formats have simpler software than others.

¹¹ Mary Feeney (Ed), *Digital Culture: maximising the nations investment*.(1999) (This is a synopsis of 7 JISC/NPO studies on digital preservation).

creator of the software to make new application versions compatible with their predecessors. This may be in the best interests of the software creators, but compatibility generally only lasts for a few generations, and even then the software creator may alter some features. *Interoperability* requires the ability to move objects from one platform or application to another. Measures must be taken to ensure there is no loss of authenticity, information or functionality. This is similar to the *conversion to standards* technique (also referred to as reliance on standards or *storage in standard formats*), which involves moving record objects from their original application-specific format to a standard, often non-proprietary, format.

The CLIR report by Lawrence et al, *Risk Management of Digital Information*, addresses the fact that the Task Force definition of migration is broad, and suggests that a more specific version would indicate 'that migration changes the structure of the original data file'¹². This retains the digital aspect and rules out the possibility of transfer to a non-digital format or to alternative storage media.

The recent publication by the CAMiLEON project offers yet another approach to classifying individual migration strategies¹³. It recognises that different digital objects have different preservation requirements, and thus proposes levels of migration, from 'minimum preservation', through 'minimum migration' and 'preservation migration', to 'recreation', 'human conversion migration', and 'automatic conversion migration'. These categories are a mixture of levels and methods of preservation and take us one step closer to fully defining migration pathways.

All this shows how much potential there is for confusion when identifying and undertaking a migration strategy, and also how rich and diverse a field migration is becoming. Research is focusing more closely on the specifics of migration, and the interpretations are becoming more evolved, more refined. We have certainly moved on since the Task Force Report of 1996, but still lack specific and detailed results of migration trials.

The Digital Preservation Testbed will contribute to this area, with recorded and published findings about the effects of specific migrations on the integrity and authenticity of archival records. The Testbed definition of migration is a relatively simple one: 'the transfer of records from one hardware and software configuration to another'¹⁴. This allows for research into the strands of migration identified above, including backward compatibility, interoperability, and the use of standards. We will also identify preservation requirements to help select relevant migration strategies for the record types under consideration.

2.2

Issues and Concerns

This section will examine the arguments for and against migration as a preservation strategy.

Migration offers the simplest technological way to ensure continued access to authentic digital records over the long term. It is particularly suited to documents

¹² Gregory Lawrence et al, *Risk Management of Digital Information: A File Format Investigation* (June 2000).

¹³ Paul Wheatley, *Migration – A Discussion Paper* (2001)

¹⁴ Nancy McGovern, *Digital Preservation Testbed: Research Framework 2001*. Dutch regulations define migration as the transfer of data and application software to another platform, platform being 'the whole of the hardware and operating software on which the application software runs'.

(records from word-processing software such as Microsoft® Word®), as our preliminary results have indicated. Conversion software is often available, and there will always be target file formats to migrate to. From a user's point of view, it can be more desirable to migrate than to emulate. If the record has been migrated, then the user can easily access, read and use the record with their familiar software tools. An emulation system may require that the user learn to use unfamiliar technology to access the record. Also, archives can develop methods to check that the record is still authentic and complete through migration cycles. Migration is currently the most 'doable' technological option for preservation. However, although it may be the simplest preservation option, it is by no means simple.

The main problem with migration is that it requires a specialised investigation for every pair of original document file format and target file format if the migration is to be carried out in a controlled and reliable manner. This is due to the different requirements of different file formats and record types. Migration to a standard format reduces the number of file format pairs for which investigations need to be undertaken, and many archives now restrict the number of formats they will accept¹⁵. However, a single migration cycle is unlikely to ensure that the record will last for as long as its retention period requires, and a typical record will need to be migrated more than once during its useful life¹⁶. This is especially true for records whose historical value means they should be preserved "forever".

It is also difficult to decide when a migration process ought to be initiated. Some archives now carry out a regular Technology Watch for developments which may affect their holdings. As the Task Force noted, there was (and still is) limited experience with the types of migrations that are needed to maintain access to records.

Migration is also labour intensive, unless it can be automated. Files have to be migrated each time a change in technology requires it, even though nobody knows whether anyone is going to use them or not. The metadata accompanying the file may also require a migration, depending on the format in which they are held, and the metadata must also be updated to account for the migration process itself. These problems mean that it is difficult to create a cost model for migration, but it is essential to have such a cost model so that archives can plan their budgets.

Migration has also been criticised because of the unknown effects the migration process may have on the authenticity of the record. Migration involves a transformation of the original bitstream. Even an exact copy process may cause corruption to the record through software bugs, mishandling of data, or mechanical failure of the input or output devices¹⁷. Migration adds further unknown risks of damage to the record. The Testbed experiments will quantify some of these risks.

Woodyard, in *Practical Advice on Preserving Publications on Disk*, makes the point that migration requires specific conversion software to convert the record from one format to another. If there is no such software available, developing a customised

¹⁵ The VERS strategy does this, accepting documents in PDF format and databases in XML. XML is also used to capture and store the required metadata. This is also the proposed policy currently under review at the Dutch National Archives.

¹⁶ Rothenberg and Bikson, *Digital Preservation: Carrying Authentic, Understandable and Usable Records Through Time* p51.

¹⁷ Lawrence et al, *Risk Management of Digital Information*, as cited in Hedstrom, *Digital Preservation: Problems and Prospects*. However, it is worth noting that these are problems that could also occur when utilising other preservation approaches,

migration system can be complicated. It can also significantly increase the costs of maintaining and preserving the records. Migration is easier and cheaper when the original software manufacturer provides the conversion software as part of their new version. An example is Microsoft, who include conversion from Word '95 and '97 in their Word 2000 software. This is the backward compatibility approach to migration.

Relying on backward compatibility as a migration strategy involves a risk because current versions of software will usually accept files from two or three previous generations, but not from all previous versions. Also current software versions rarely operate all of their features in the same way as earlier versions¹⁸. This makes this strand of migration more suited to records with short retention periods. Finally, software developers are not required to make their programs backward compatible. However, it often serves their purpose to do so as they want consumers to buy the new version of their product, and to move to it with as few problems as possible.

Conversion to or reliance on standard formats as a migration strategy also has its problems. There are many standards to choose from. These include ASCII (for text), PDF, and XML. There must be convergence towards agreed standards if they are to be effective for preservation. The standards themselves are also changing and new versions are being released, which means that the migration cycle may still continue to be iterative. The proprietary nature of some standards, typically Adobe's PDF format, attracts criticism in that proprietary software cannot be guaranteed to continue for the long term. Note, however, that the PDF specifications and reader software are freely available, and that open standards are also not guaranteed over the 100 years for which we may wish to preserve records.

Despite these issues, migration remains a feasible option for long term digital preservation. Thanks to the research efforts of the Digital Preservation community over the past decade, the problems and risks encountered in migration have been mostly identified. It is now simply a matter of finding solutions to these problems. Many archival institutions are actively studying digital preservation, putting migration strategies in place, and issuing guidelines on the preservation of digital objects. A survey of current literature shows that, with an appropriate risk assessment in place, migration is believed to be an adequate preservation approach for the majority of record types. The report by the National Preservation Office concluded that some form of migration strategy was suitable for everything but the most complex data types¹⁹. The recent InterPARES report identified migration as the most prevalent preservation strategy currently in operation.

However, there remain some questions about the migration process and its effect on archival integrity. While the risks can be reduced by an accurate risk assessment and by investigation of the file and target formats, there is still the possibility that the migration may introduce a change in the bit stream that affects the record's archival integrity. It is also possible that the software used to interpret the new data format (after the migration) will introduce a change. The Digital Preservation Testbed will examine the effects of specific migrations and migration cycles on the authenticity of government records.

¹⁸ NLA, *A Draft Research Agenda for the Preservation of Physical Format Digital Publications*, p2.

¹⁹ Mary Feeney (ed), *Digital Culture: Maximising the nations investment*

3

Migration In Practice

The pros and cons of migration are being studied by groups or institutions world-wide. Knowledge has certainly increased when it comes to selecting an appropriate migration strategy, and interest in the viability and suitability of migration as a preservation strategy is growing. The projects and institutions featured below do not include every migration preservation strategy in operation. They are a selection from current work.

3.1

Current Research and Knowledge

There are currently no research projects that we know of focusing exclusively on migration; migration normally appears as a part of a wider research project. One of the current projects which contains migration research is the CAMiLEON project. The CAMiLEON Project is a joint venture by the Universities of Michigan and Leeds, funded by the National Science Foundation (NSF) and the JISC, concerned primarily with Emulation. Part of their work includes testing user preferences for different versions of 'preserved' objects, comparing those which have been emulated with those which have been migrated. Hedstrom cites the following example: "In a recent experiment, users who compared a migrated or emulated version of a computer game to the same game running in its original native environment, generally found the migrated version to be more like the original, in part because a thorough migration (complete re-write of the original code to run on a current platform) performed more like the original program than a poor emulation". Although this refers to console gaming, it is still a working example of migration versus emulation in practice. These findings may not be reproduced for archival records, where the criteria for success are different. A console game is far more interactive and filled with behavioural traits than a Word 97 document or an Excel spreadsheet. Preserving a record requires attention to aspects of authenticity and integrity which are not required for a game. However, the research on games may well be relevant when considering the preservation of databases or more complex digital objects.

The InterPARES project has recently published Draft Final Reports from each of its three Task Forces. The reports by the Preservation and Authenticity Task Force are of particular interest to the DPT. The Preservation Task Force final report includes a survey of Preservation Practices and Plans. This indicates that migration is the most common approach in use. It also includes a high-level Preservation Action Model. The Authenticity Task Force Report includes a set of baseline and benchmark requirements to ensure the authenticity of the record over the long term. While neither of these focuses specifically on migration, the guidelines and models presented can be applied to migration (among other preservation approaches).

Research is currently underway in Australia by both the National Archives and the National Library. The National Archives were due to begin a project in 2000 to develop advice for Commonwealth Agencies on using migration as a preservation

strategy over the long term²⁰. The National Library has conducted trials on various types of transfer, migration, and emulation²¹. Their migration trials focused on text files in the National Libraries Manuscripts collection and HTML files in the PANDORA collection (for online publications). The Library acquired a number of documents stored in proprietary formats which they did not have the software to access²². Through migration to a common format, the Library is now able to view the documents using the standard word processing software in the Library, Microsoft Word. They have also considered another migration option: to migrate each item to a current operating environment, but to maintain a 'preservation' master copy of the item in its original format so that it can be emulated when losses through migration become unacceptable. This option displays how a practical preservation approach can often rely on more than one strand.

Within the NLA, Collection Areas and Preservation Services (CAPS) maintain a list of hardware and software used. This is regularly examined and used to flag potential changes in technology. They advise which hardware or software should or should not be disposed of, and when. They also advise the Information Technology section to ensure they are aware of collection needs when planning a change of hardware or software within the Library.

The Library is experimenting with removing 'dead' source code from selected HTML pages and replacing it with current tags. This will enable the team to assess the effect and the potential for continued use of making changes to the HTML code, albeit small changes. The NLA's approach is that migration does not necessarily have to maintain the look and feel of the item. This may not be a problem for the Library, but may incur problems with archival records.

There are still surprisingly few hard facts available about the effects of specific migrations upon archival records. This is due in part to the plethora of document and format types that can qualify as records. As Eiteljorg points out, "Although the problems of migration are well understood and do not represent a significant intellectual burden [documented] practical experience is very limited"²³. There are some examples of the migration of legacy systems, but such examples are specific to the systems involved and are focused on the migration not only of the contents but also the infrastructure of the system which supports and stores them²⁴. The CLIR Report *Risk Management of Digital Information* is one of the few documents which provides specific details, providing details of two case studies, one for an image file format (TIFF) and one for a spreadsheet file format (Lotus 1-2-3).

Snippets of information are also available. Juha Hakala of the Helsinki University library recently used the following example: if a mathematical dissertation in LATEX is converted into HTML, then all of the formulae will be lost unless they can be converted to images²⁵. An article by Al Klein in *inform* magazine notes that "Microsoft Draw is a graphic application that was available in Microsoft Word 2.0, but is inoperative in Word

²⁰ National Archives of Australia, *Electronic Records: Preservation and Migration of Electronic Records*, as cited in Woodyard, *Digital Preservation: The Australian Experience* (2000)

²¹ Woodyard, *Digital Preservation: The Australian Experience* (2000)

²² Deborah Woodyard, *Practical advice for preserving publications on disc* (1999)

²³ Harrison Eiteljorg II, *Electronic Archives* (1997).

²⁴ For a discussion of the Migration of Legacy Systems, an excellent source is Michael Stonebaker and Michael Brodie's *Migrating Legacy Systems*. Such systems are, however, out of the scope of this paper and are thus not considered in detail here.

²⁵ Juha Hakala, *Metadata for Referencing and Archival Usage* (2001).

6.0 or later"²⁶. This means that any records created in Word 2.0 which used MS Draw® will no longer be complete if accessed in Word 6.0 or a higher version. Identifying these specific areas of concern now will enable us to formulate ways around them whilst retaining the authenticity of our records.

3.2

Frameworks and General Guidelines

Three models in particular have appeared in recent years, suggesting a well-developed framework for a migration strategy.

Risk Management of Digital Information: A File Format Investigation was published in June 2000. It uses Risk Assessment as a tailored approach to choosing an appropriate migration strategy²⁷. It is centred around the concept that migration involves a translation of the file bit stream. If a Risk Assessment is done before the migration takes place, using certain practical tools, the chance of damage to the file can be significantly reduced. A Risk Management Scheme separates the process into steps that can be described and quantified, and is divided into three categories: risks associated with the general collection; risks associated with the data file format; and risks associated with the file format conversion process. By analysing both the original and the target file format, it is possible to estimate where and to what extent risk may occur. However, as the introduction notes, 'the difficulty of course is determining when risk is acceptable and when it is not'.

A similar approach is proposed by John Bennett in a JISC-funded British Library Report²⁸. This proposes the development of a 'Scorecard' approach that measures preservation complexity levels. Preservation issues are identified and records are scored against them to ascertain a rating level. These levels then equate to appropriate preservation approaches, which are not explicitly developed, but which indicate when and where further intervention is required to ensure long term preservation.

Alternatively, another JISC funded study by Tony Hendley offers a framework that allows the user to choose the most appropriate strategy for them, based on the category of digital resource under consideration. Two models are provided, a cost model and a decision model, both of which assist the record keeper in judging which is the most appropriate preservation strategy for the records concerned. Hendley's report also recognises that the cost of digital preservation can play a decisive factor in choosing a preservation strategy.

General good practice issues are also beginning to emerge. A variety of advice is now available on building migration paths and on identifying migration issues. The National Library of Australia's research into migration has been described above. The National Archives of Australia also offers advice on migration in its publication 'Managing Electronic Records' (Appendix 3), noting that 'migration becomes the focus of the preservation of continuing accessibility to electronic formats rather than the preservation of individual items or formats'²⁹. They recommend building migration

²⁶ Al Klein, *Data Migration: Issues and Strategies* (1999).

²⁷ Gregory Lawrence et al, *Risk Management of Digital Information: A File Format Investigation*.

²⁸ John Bennett, *A Framework of Data Types and Formats, and Issues Affecting the Long Term Preservation of Digital Materials*.

²⁹ National Archives of Australia, *Managing Electronic Records – Appendix 3* (1997).

paths, carefully selecting suitable target file software, and a full test of the migration process in advance, before the deletion or destruction of any records. The English Heritage Centre for Archaeology has also published guidelines for developing migration strategies. They also recommend a risk management approach to managing migration, and have issued guidelines to that effect³⁰.

Documenting the migration procedure and its results are vital if the records are to remain authentic. A full and complete account of the migration needs to be maintained, thus protecting the integrity of the records even if some features are affected. As Hedstrom notes, "Documenting information loss during a migration also offers an alternative to preserving an exact replica of the document and all of its associated functionality"³¹.

3.3

Current Use

As the problems of digital preservation become more widely known, organisations are beginning to turn to migration either as an interim solution, or as a total solution involving full file format assessments in advance. Many organisations routinely update their files from one version of software to another. This has previously been an unconscious and undocumented migration. Today, awareness and knowledge on migration matters are steadily growing. Several institutions are currently using migration for preservation, or are issuing guidelines on choosing and documenting the most appropriate migration strategy.

A recent article on the National Archives of Canada website provides details of a new Electronic Archives Preservation System (EAPS) they are implementing, which will automate the migration, survey, and retrieval functions of e-records in their collections³². Automated migration limits the amount of user intervention required, reducing the labour involved. Relevant technical information is captured as metadata during preservation processing, thus ensuring that both old and new file format details are retained. Indeed, migration is not practical without automation for any but the smallest data collections or organisations.

Many archives are implementing procedures which limit the formats of records they will receive. This reduces the number of file format investigations which are required. The Dutch National Archives is currently drafting legislation limiting document formats to PDF and XML, and the Australian VERS Electronic Record Strategy has resulted in a similar strategy and legislation³³. The UK Public Record Office limits its formats for transfer into the archives to Postscript, TIFF, SGML and PDF. Limiting the number of formats that an Archives will accept may require pre-archival conversion by the originating agency, but if the process is controlled and fully documented at all stages, and the target file format has been well chosen, this should not be a problem. The selection of appropriate file formats corresponds with the use of Standards as a migration pathway. Using PDF as a Standard has raised concern because it is a proprietary format, but Adobe have alleviated much of that concern by placing the specifications in the public domain.

³⁰ English Heritage Centre for Archaeology, *Digital Archiving Strategy, Version 1.0* (2000).

³¹ Margaret Hedstrom, *Research Issues in Migration and Long Term Preservation* (1997).

³² http://www.archives.ca/13/130103_e.html National Archives of Canada site. Details of their Electronic Archives Preservation System.

³³ The VERS Strategy has been widely publicised. Essentially, documents are converted into PDF which is then converted into Base 64 and wrapped in XML. The XML wrapper includes the records metadata.

3.4

Conclusion

It is gratifying to see attention turning away from theoretical problems and potential drawbacks and focusing instead on practical applications, advice and experience. Migration has much potential for the preservation of different record types using different applications and with different requirements. A suitable investigation of the requirements and ways to meet them in advance can satisfy many of the issues and concerns that have been raised.

The Digital Preservation Testbed project will identify file formats and record types suitable for migration. We will provide recommendations and advice to the government and National Archives of the Netherlands to help design and build migration pathways. This advice will also be used in the maintenance of electronic records while they are still in active use. We anticipate that migration alone will not solve the problem of digital longevity. File formats and preservation requirements differ so widely that it will not be possible to develop a 'one size fits all' approach. However, migration will almost certainly form part of a wider and more pragmatic strategy for long term preservation of digital objects and archival records.

4

Bibliography

- Bearman, David *Reality and Chimeras in the Preservation of Electronic Records* (1999)
<http://www.dlib.org/dlib/april99/bearman/04bearman.html>
- Bennett, John C *A Framework of Data Types and Formats, and Issues surrounding the Long Term Preservation of Digital Material* (1997)
<http://www.ukoln.ac.uk/services/papers/bl/jisc-npo50/bennet.html>
- Cloonan, Michele, and Sanett, Shelby *Preservation Strategies for Electronic Records, Round 1 (2000-2001) Where We Are Now: Obliquity and Squint?* (2001)
- CPA/RLG *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information* (1996)
<ftp://ftp.rlg.org/pub/archtf/final-report.pdf>
- Dollar, Charles *Authentic Electronic Records: Strategies for Long Term Access* (2000)
- Eiteljorg, Harrison II *Electronic Archives* (1997)
<http://intarch.ac.uk/antiquity/electronics/eiteljorg.html>
- English Heritage Centre for Archaeology *Digital Archiving Strategy, Version 1.0*(2000)
<http://www.english-heritage.org.uk/knowledge/archaeology/das1-0.pdf>
- Feeney, Mary (ed) *Digital Culture: maximising the nations investment* (1999)
- Hakala, Juha *Metadata for Referencing and Archival Usage* (2001)
<http://associnst.ox.ac.uk/~icsuinfo/hakalafin.htm>
- Hedstrom, Margaret *Digital Preservation: Problems and Prospects* (2001)
<http://www.si.umich.edu/CAMILEON/>
- Hedstrom, Margaret *Research Issues in Migration and Long Term Preservation* (1997)
<http://www.sis.pitt.edu/~cerar/s5-mh.html>
- Hedstrom, Margaret *Section of a Report on Migration Strategies prepared for the Experts Committee on Software Obsolescence and Migration(Draft)* (1997)
<http://www.sis.pitt.edu/~cerar/ftp-docs/Mig-Stra.doc>
- Hendley, Tony *Comparison of Methods and Costs of Digital Preservation* (1997)
<http://www.ukoln.ac.uk/services/elib/papers/tavistock/hendley/hendley.html>
- Hodge, Gail *Best practices for Digital Archiving: An Information Life Cycle Approach* (2000) <http://www.dlib.org/dlib/january00/01hodge.html>

- Hodge, G, and Carroll, B *Digital Electronic Archiving: The State of the Art and the State of the Practice* (1999) <http://www.icsti.org/icsti/99ga/>
- Klein, Al *Data Migration: Issues and Strategies, in INFORM, magazine of the aaim.* (1999)
- InterPARES *Preservation Task Force Final Report (DRAFT,2001)*
http://www.interpares.org/documents/ptf_draft_final_report.pdf
- InterPARES *Authenticity Task Force Final Report (DRAFT 2001)*
http://www.interpares.org/documents/atf_draft_final_report.pdf
- Lawrence, G et al *Risk Management of Digital Information: A File Format Investigation* (2000)
<http://www.clir.org/pubs/reports/pub93/contents.html>
- Levy, David *Heroic Measures: Reflections on the Possibility and Purpose of Digital Preservation* (1998)
- Lorie, Raymond A. *The Long Term Preservation of Digital Information* (2000)
<http://www.almaden.ibm.com/u/gladney/Lorie.pdf>
- McGovern, Nancy *Digital Preservation Testbed: Research Framework* (2001)
- National Archives of Australia *Managing Electronic Records* (1997)
http://www.naa.gov.au/recordkeeping/er/manage_er/
- National Library of Australia *Electronic Information Resource Strategies and Action Plan, 2001 – 2002*(2001) <http://www.nla.gov.au/policy/electronic/resourcesplan.html>
- National Library of Australia *A Draft Research Agenda for the Preservation of Physical Format Digital Publications* (1998) <http://www.nla.gov.au/policy/rsagenda.html>
- Russell, Kelly *Digital Preservation: Ensuring Access to Digital Materials in the Future*(1999)
<http://www.leeds.ac.uk/cedars/Chapter.htm>
- Rothenberg, Jeff & Bikson, Tora *Digital Preservation: Carrying Authentic, Understandable and Usable Records Through Time*(1999)
http://www.digitaleduurzaamheid.nl/bibliotheek/final-report_4.pdf
- Wheatley, Paul *Migration – a CAMiLEON discussion paper*(2001)
<http://www.personal.leeds.ac.uk/~issprw/camileon/migration.htm>
- Woodyard, Deborah *Digital Preservation: The Australian Experience* (2000)
<http://www.nla.gov.au/nla/staffpaper/dw001004.html>

- Woodyard, Deborah *Practical Advice for preserving publications on disc* (1999)
<http://www.nla.gov.au/nla/staffpaper/woodyard2.html>
- Woodyard, Deborah *Farewell my Floppy: A Strategy for the Migration of Digital Information*
1998/9)
<http://www.nla.gov.au/nla/staffpaper/valadw.html>

5

Websites

http://www.archives.ca/13/130103_e.html National Archives of Canada website.

<http://www.jiscmail.ac.uk/> JISC Listserv

<http://www.digitaleduurzaamheid.nl/> Testbed website

<http://www.interpares.org/> InterPARES Web site

<http://www.si.umich.edu/CAMILEON/> CAMiLEON website

<http://www.leeds.ac.uk/cedars/> CEDARS website

<http://www.nla.gov.au/padi/> PADI (Preserving Access to Digital Information) website

<http://www.dlib.org/> D-Lib website

<http://www.rlg.org/preserv/diginews/> RLG Diginews website

<http://www.prov.vic.gov.au/vers/final.htm> VERS Final Report site

<http://www.archivebuilders.com/whitepapers/index.html> Archive Builders website

<http://www.pro.gov.uk/recordsmanagement/eros/> UK PRO E -records site