Assessing Digital Preservation Strategies

Jan Hutař, Senior Advisor, Archives New Zealand, Wellington, NZ

Currently digital preservation practitioners have two primary strategies for use in preserving digital objects: Migration and Emulation. Archives New Zealand has begun investigating the impacts of implementing each of these digital preservation strategies. We are implementing the Rosetta long-term preservation system in our Government Digital Archive Programme (GDAP). We share Rosetta with the National Library of New Zealand on consortia level and therefore we need to understand commonalities and possible differences in digital preservation strategies, needs and implementations between archives and libraries in general. We have been conducting research into the successes and pitfalls of different digital preservation strategies to test the practicalities of the different approaches; results and their meaning for large scale institution are briefly described in this paper.

1 Introduction

It is well known that archives and other memory institutions, and our lives in general, have been through quite a big change last few decades. Information, and the way it is kept, shared, created and stored has changed and it is changing at this very moment. Physical carriers like paper are not necessary anymore for all those activities. From the late nineties archives and libraries began to encounter more and more information in digital form, first saved on different portable media, later coming from the web and via hard drives. In this period the entire concepts of acquisition, processing, storage and preservation have irreversible changed.

I think we can agree on that perception and adoption of this evolution has been quicker in libraries. Libraries began with digitisation in the middle of nineties and then moved to acquiring born digital material. In New Zealand it was the National library (NLNZ) which started with the National Digital Heritage Archive programme (NDHA) in 2005 and went live with their first version of the digital preservation system in 2008. Archives around the world in many cases joined this trend a bit later, when it was obvious that the increasing shift to digital technology in the Public Sector has created new challenges and environment for the management and preservation of government information. Archives New Zealand (Archives NZ) identified this issue and proposed the Digital Continuity Action Plan (DCAP), which Cabinet endorsed in 2009. Crucial part for fulfilling statutory responsibilities for the long term preservation and accessibility of digital data is to build robust digital archive system and processes. The Government Digital Archive Programme (GDAP) has been set up in 2010 to manage this work [1]. GDAP includes 3 phases each ending with major Release. At the time of final Release 3 in July 2013 Archives NZ will be able to conduct digital preservation processes on daily basis; agencies will be able to access their own restricted digital material and Archives NZ should be able to process all complex digital transfers from agencies.

As a part of the streamlining of government agency structures, The National Library of New Zealand and Archives New Zealand have been formally incorporated into the structure of the Department of Internal Affairs (DIA) on February 2011. One of many consequences of this major change was the decision that Archives NZ will leverage from previous government investment and research and GDAP will share the

already existing digital preservation system at NLNZ – Rosetta. This decision has influenced and still is influencing the work of both institutions. The aim is to share as much as possible of policies and thus processes, with existing points of divergence where necessary. Archives NZ is trying to understand all details and possible benefits coming from different preservation approaches (emulation and migration).

2 Digital Preservation Strategies

The preservation of digital objects is a process that is very different to the preservation of traditional documents, for example on paper. Paper documents are stored under special climate conditions and the aim is to keep them unchanged. This is intended to ensure that future users will be able to read the document in the future (assuming they know how to). The preservation of digital objects in digital form has the same aim: to keep the objects content unchanged and provide the users with the possibility to read and understand their content. But the options available for how to achieve this are different. For example it might be necessary to change the technology used to present the digital object's content to users via a migration preservation action that changes both the file or files that store the basic constituents of the content and the software (SW) used to add to and render that content in order to display or convey it to users. In this situation information about all changes (events) have to be kept in metadata. Alternatively emulation can be used for providing access to "not touched" files in the future using the original SW technology.

Digital objects are endangered for two main reasons, the first being the instability of physical media carrying the object, and the second being related to the rendering of digital files and the associated dependency on SW applications or relevant hardware (HW). Digital Preservation Strategies can be divided into groups: 1) preservation of technology (computer museum, HW preservation); 2) technology emulation (creating emulation SW; or digital archaeology); 3) information/data migration (physical migration to different media, data format migration and normalisation); 4) other approaches like encapsulation, migration to classic materials (paper). Physical migration of digital objects to the new (or redundant in case of backups) media or HW is simple bit stream preservation and thus precondition to the logical preservation which aim is to preserve the content, readability, usability, understandability of the object, to track all the changes during its lifecycle and to be able to actively verify the integrity and fidelity of the preserved digital object. Currently digital preservation practitioners have two primary strategies for use in logical preserving digital objects: Content Migration between formats and their associated rendering technology and Emulation.

2.1 Migration

Migration is widely used term for the process of moving content from one (or more) file(s) formatted with a no longer supported standard, to files formatted with a newer, still supported standard. Or more generally it is about periodic transfer of digital materials from one HW/SW configuration to another, or from one generation of computer technology to a subsequent generation [2]. The main purpose is to retain easy access to the content that is able to be migrated in future technological environments. Successful content migration means the content is easy to retrieve, easy to render and easy to use in that environment. If the migration is chosen as primary or one of preservation actions, it has to be inevitably repeated as new technology and file formats will be emerging. In OAIS this understanding of

migration is called Transformation and is described as: A Digital Migration where there is some change in the Content Information or Preservation Description Information bits while attempting to preserve the full information content [3].
For a long time migration has been seen as the only viable digital preservation strategy. However there is at least one well known issue with it, it can cause changes to objects each time it is applied. It should be noted that the only possible and acceptable change of the file (significant) properties, is such a change which we (as preservation managers) know about and has been accepted as not important or influencing the file's content perception. There shouldn't be any change or even alteration of the content we are not aware of. It might be acceptable to have some of the significant properties changed, for example change of colour, different "look and feel", ability to edit the document etc. It all depends on institutional preferences. Preservation actions (i.e. migration in this case) are assessed and reviewed on the basis of change to significant properties. Before proceeding with any migration action, practitioners should be sure about which significant properties they consider as important, less important and optional to keep. In order to do this it is in turn necessary to have significant properties expressed in our files associated metadata.
Migration might be considered difficult because it is complicated (or not possible) to decide what is the right moment for starting the migration [4] if we don't know what is the next direction of technology development. I think this argument is not valid anymore; there are systems able to help with that decision and alert the preservation manager about the existing threat. In majority of cases this ability is based on the rendering application availability, file format support, available documentation, list of set risk identifiers etc. In many aspects external services like DROID/PRONOM, UDFR help to enable this functionality in long-term preservation systems.
In general data migration is also used in many systems during ingest for creating the access derivatives (for example tiff>jpg) or normalisation (i.e. migration of the file to new file formatted in preferred file format; example of this would be XENA or Archivematica tool [5]).

2.2 Emulation

Emulation is the strategy of using original SW to open or "render" files formatted with a standard that isn't supported by modern SW. The original SW is run on a recreation of its original HW running on a modern computer and is then used to open or run the old files. Emulation has been seen for a long time as being impractical but recent developments in the Information Technology landscape have begun to undermine that assumption. Migration could work fine for simple individual digital objects, with not many dependencies and advanced functions. For complex objects it might be reasonable to think about emulation. There is general agreement in the digital preservation community that emulation is a necessity to preserve certain types of digital objects (e.g. games/interactive objects, obscure types of objects). Rendering digital objects in their original environment won't lead into changes of the object itself. In emulated Operating system on emulated HW it is possible to use original SW applications.
There is a visible movement for a more open perception of emulation and using emulation in daily preservation tasks. Emulation has been in use quite extensively in the enterprise context over the last 17 years (used to maintain old systems too costly to replace). The regular use of these technologies (including virtualisation) means that the understanding of the concepts that underpin emulation

solutions and expertise is at an all-time high. If you browse through the iPRES conference proceeding, you would find quite a few papers describing emulation utilisation [6]. A great deal of information and tools has been published by the KEEP (Keeping Emulation Environments Portable) project [7].
Emulation is being attributed with couple of problems. First of all emulation is considered to be quite difficult. Deploying and using the emulator doesn't need to be difficult and might actually be quite easy, but to create the emulator, that is different story. To do that, it is necessary to have technical knowledge about the HW layer, OS layer. Related and often mentioned issue is that the end-user needs to know how to use the SW used for rendering the digital object. In cases where the SW is really old and there is not documentation available, this could cause lot of trouble. We can also say emulation is preserving the environment and its functionalities rather than preservation of digital objects themselves [8]. Emulators themselves will become obsolete in the future if dependent on certain technology. A bigger problem is that many SW applications are dependent on different types of activation, anti-copying protection, which are not possible to do in emulated environment. Another related problem is copyright and legal issues linked to activities like copying data from protected media, usage and tweaking of copyrighted SW etc. This has been known for quite a long time [9] and among others has been described lately in KEEP project [10]. All those above mentioned things have an effect that emulation is still not considered as good enough for assuring ongoing access to digital objects over the time without big effort and ongoing development but the huge potential benefits are seeing a lot of support for these efforts and developments to continue.

3 Preservation Strategies at Archives New Zealand

Some of the preservation activities conducted at Archives NZ have been driven by immediate urgency (copying data from physical media to our digital archive), some were more related to finding out the possible approaches to the data we are getting to the archives already, or we think we will be receiving.

3.1 Physical Media Migration

The immediate urgency apply to data from our collections stored on endangered (CD, DVD) or obsolete (different sizes and types of floppy discs) physical media. We started with the migration as secondary activity to our main BAU in early 2011 and proceed mainly with all different floppy (5,25"; 3,5") and optical discs. We are using the KryoFlux floppy controller [11] technology for copying the content. Total of 400 different floppy discs; around 150 optical discs; 30 Open Real Data Tapes and other individual items has been migrated. Mainly the digital document is only the digital instance of stored paper document. Sometimes we have digital documents as the only copy; very often media are technically obsolete or linked with unknown operation system or both. Interesting were 9 track tapes with the information about university grants or 62 Maori Land Court (MLC) 5,25" discs from the middle of 1980s with unusual Convergent Technologies OS. Images of MLC discs were created in early 2011 with KryoFlux, but we were unable to extract the files from them as they were structured using a very rare file system. Our colleagues in Germany developed file extraction tool for those discs in early 2012. We provided them with the file system structure that we have found on the internet. That documentation has since been removed from the internet. Discs content was mainly word processor files with Maori

judges notes [12]. All data acquired in migration from physical media are stored on SAN and will be ingested into our Rosetta long-term preservation system later on.

Number of disks imaged	345 disks
Number that could be read	190 disks
% of all disks that could not be read	45% of all disks
Number that couldn't be read due to file system	130 disks
% of total unreadable due to un-recognisable file system	38% of all disks
Number that couldn't be read due to disk/sector failure	25 disks
% of all disks that couldn't be read due to disk failure	7% of all disks
% of disks with a recognisable file system that were unreadable due to disk failure	12% of disks with recognisable file system

Table 1 - Disk Image Read Failures: Disks written between 1985 and 2000 [13].

3.2 Database preservation

Other interesting case is related to the preservation of databases as was part of one of your Work Packages. Land Information New Zealand (LINZ) transferred a large set of paper records to Archives New Zealand. Those records used to be indexed and made discoverable by a database system called SALT (Simple Access to Land Titles). Most of the data from this database was migrated out and used in other systems however the database was still a useful tool for making the paper records discoverable. The database consists of an MS SQL server running on Windows 2000 Server with a custom html front-end. Archives NZ was given the option of taking the original database HW to try to recover the database from by migrating the whole desktop to virtual HW [14]. The disc image has been created, converted into VMware disk image format. With some difficulties we were able to run the database in VMware, including cracking the passwords for the original OS. After that we were able to use original Windows 2000 environment and start the database and its web based GUI. To have more sustainable approach to preserving access to this database and its GUI, we decided to run it on fully emulated HW (only some of the VMware HW is emulated). We used one of the best emulation SW available for Windows 2000 Server - open source SW: QEMU [15]. Now original SALT database can be used in our reading rooms as the secondary finding aid as it was originally meant to be.

3.3 Rendering Matters Report

Another very useful effort has been spent on the Rendering matters report conducted during 9 months in 2011 [16]. In general the report outlines the results of research investigating whether changes are introduced to the information that is presented to users when files are rendered in different HW and SW environments. The sample test set consisted of 110 office administration files including word processing document files, presentation files, spreadsheets and databases. The basic methodology involved: first opening each file in SW that was chosen as the original rendering SW and which was running on original HW from the era. Various aspects of this “control rendering” were then documented using a survey tool. These “attributes” included such things as: whether metadata was embedded in the file, how images and diagrams were displayed, what word count the SW gave, and whether various formatting aspects or

fonts were included. After that each file was opened in the same SW running on emulated HW and the same attributes were evaluated to check for any changes, and documented again. Each file was then systematically opened in a number of modern office SW suites and the same attributes were evaluated and the results documented [16]. Some of the results were expected, for example that the choice of rendering environment (SW) used to render an office file invariably has an impact on the information presented through that rendering. When files are rendered in environments that differ from the original then they will often present altered information to the user, in some cases the information presented differs from the original significantly. Other findings showed that the emulated environments, with minimal testing or quality assurance, provided significantly better rendering functionality than the modern office suites (60-100% of the files rendered using the modern office suites displayed at least one change compared to 22-35% of the files rendered using the emulated HW and original SW). In spite of that many files may not need to have content migrated at this stage as current applications can render much of the content effectively (and the content's accessibility will not be improved by performing this migration).

3.4 File Format Migration

File format migration research at Archives NZ hasn't been that extensive. The Rosetta long-term preservation system, which we share with the National Library, has a Preservation module at the moment designed mainly for data migration as being main digital preservation approach. As we are in the first phase of deploying the system, we haven't used this functionality to do real migration of big amount of data yet. Extensive preservation functionality testing at Archives NZ is planned from the end of 2012.

The digital preservation objective for the Rosetta is to maintain the ability to render the content that is intended to be preserved. In this context, risks are potential issues that might cause there to no longer be a currently supported application associated with a format for rendering. This may include obscurities in files that mean currently supported applications cannot properly render them or may include a lack of applications currently easily available to render objects. Risks can be removed at least in two ways: 1) the files associated with the format that has a risk associated with it can be migrated to new "formats" that do not have a risk associated with them; 2) the format that has a risk associated with it can have a new application associated with it that is currently supported. Both of these options rely on a great deal of technical information in order to be achieved effectively. This and other information is extracted by the tools like DROID, JHOVE, NZME included in a Rosetta installation. This information is what the "risks" in Rosetta are based on/populated from [17].

Rosetta's Preservation module provides an environment for supporting all activities related to preservation, including possibility to describe the risks in the repository, locate the files at risk, create preservation plans to mitigate the risks, automatically or manually compare different alternatives of the plan and execute the plan. Rosetta not always does the actual migration however. This is because each migration combination of source to target format and rendering environments may require a separate tool and therefore it would not be suitable for a generic preservation suite to include such specialised and variable functionality. It provides the environment to keep track of what was happening. A simple

description of the common business workflow ending with preservation action (migration) would be following:

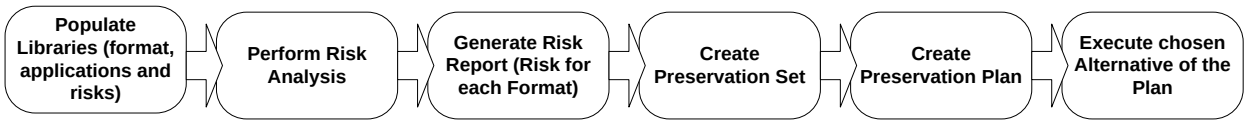

Figure 1: Simple preservation workflow [18].

A necessary pre-requisite to perform such a migration is to know what file formats we store in the system, what their characteristics, dependencies, risks and rendering applications/environments are. This information is kept in the Format library for each file format. This allows the preservation analyst to run risk report on the stored data and see the risk status for them. The risk analysis process itself is part of the SIP processing; each file ingested is checked for risks. A risk report will identify the set of endangered files which could be used as Preservation set for testing different alternatives of Preservation plan. A preservation plan would normally include a test on small portion of files from the set and comparison of different alternatives. After the alternatives of the plan have been evaluated, the final step is to execute the chosen plan – the new risk-free version (representation) of Preservation master will be created. A Preservation analyst has to ensure to know what will happen with the information content of the files, their significant properties etc. Migration itself could be done by internal transformation agents or with the help of external tools (export>transform>import). Information and details about the event has to be kept in metadata to provide the possibility to track back what happen. This above described process will be tested in detail later this year. Significant amount of effort has been put into similar testing at the NLNZ already.

4 Possible differences between archives and libraries related to digital preservation system deployment

As stated above, Archives NZ and NLNZ share the Rosetta long-term preservation system on consortia level. During our implementation and discussions about shared policies and the way of using the system, we came across topics and facts that originate from differences between both types of institution and therefore may create different approach to digital preservation and business as usual activities. If we exclude the libraries with archival function, i.e. serving as both as a library and archive, then the possible general fields of difference between archive and the library running the digital archive might include:

- data creation and transfer – data in government agencies may exist in many different file formats, often dependent on proprietary SW application or system used for creating or storing them (incl. databases, emails, old legacy files like SBF or Cal-Scan files we received in the pilot transfer from agency); data coming to the library might be (not necessarily) more homogenous (legal deposit, digitisation, web archiving); dealing with the agencies about what file format will they transfer seems to be crucial for archives in order to avoid the flood of different file in transfers and related work effort;
- appraisal – archives might need environment for appraisal of relevant documents before ingest; they might receive HDD or system dumps with system files etc.; libraries usually gets sets of

documents for direct ingest with no need for appraisal and sorting (except cases where library receive HDD or whole computer of some famous writer and is supposed to preserve the content);

- ingest – archives will likely encounter more files with unknown (not identified) file format, possible extension mismatch (many with the same problem at the same time, i.e. time of one ingest); due the variety of file formats archives might have more problems with the format validation, metadata extraction which poses a significant risk to the ability to undertake successful digital preservation;
- data management – bigger demand for privacy and data security in archives (unauthorized access); archives are sometimes committed to delete stored data after specific period of time (in general, LTP system shouldn't be used as the recordkeeping system, it brings the question if to store files with disposal timeframe in permanent repository under same conditions as long-term value documents);
- preservation - preservation planning and preservation action decisions could be different based on different needs for keeping significant properties or file features (acceptable outcome of the same preservation action then could be different); archives will put more accent on emulation (access to databases, compound objects, legacy systems etc.) in order to ensure high fidelity and object integrity for the legal records that archives preserve;
- access and delivery – archives likely to have different data types to provide access for with different features and thus different viewers or techniques for specific data; differences related to access restrictions coming from legislation and the nature of stored information; need for the solution for the documents not suitable for data migration (implement emulation framework); high demand to prove authenticity and integrity of the document in archives coming from legislation.

We think migration is the main current method we will be using. At the same time our opinion is that one of the important aspects for archives will be deploying the emulation as another valid preservation approach for specific types of documents. Undertaking emulation as a preservation action for certain specific data types would mean to support a particular emulation environment/application and then associating that environment/application with at-risk files so that they are no longer considered at risk. Emulators could potentially be added as viewers and associated with certain formats. To enable this, viewers would have to be able to be associated with representations based on rendering application information rather than file format information.

5 Conclusion

The paper summarized digital preservation activities at Archives NZ and showed future direction of our effort. We know similar activities could be found in many national institutions internationally. We think that migration and emulation are both valid approaches and we will continue with the research and using both methods in business as usual. It's worthy to stress that emulation, migration and other approaches are not competing each other in most cases; quite contrary the usual situation is they might complement each other and might be used by one institution for different purposes or different document types. Decision what approach is suitable for which document types is important to do as early as possible, because it help us to clear out what kind of metadata we would need, what would the

processes and policies will look like etc. We would like to see it in current LTP systems as one of the options.

Sharing one LTP system between the library and the archive brings many interesting and challenging situations and opportunities for further thinking about digital preservation. Librarians consider the content of the document as the main thing to preserve and migration very well fit to this concept. Archivists sometimes tend to think that the "original" document received from the agency in transfer is itself the subject to preserve. Unfortunately preservation of the content very often means changing the document (not its content). To preserve doesn't necessarily mean to keep the document untouched. The challenge in the digital age is defining what the digital object and its content are so either or both can be preserved without change indefinitely.

6 References

[1] FLEMING, Alison. 2012. *Government Digital Archive Programme: Programme Organisation*. June 2012. p. 5. Internal Document of Archives NZ. CMS ID A654810.

[2] GARRETT, John and Donald WATERS. 1996. *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information* [online]. The Commission on Preservation and Access and The Research Libraries Group, 1996 [cit. 2012-11-07]. p. iii. http://www.clir.org/pubs/reports/pub63watersgarrett.pdf

[3] GIARETTA, David. 2011. *Advanced Digital Preservation*. Heidelberg: Springer, 2011. pp. 171 and 200-201. ISBN 9783642168086.

[4] ROTHENBERG, Jeff. 1999. *Avoiding technological quicksand: finding a viable technical foundation for digital preservation.* Washington (DC): Council on Library and Information Resources, 1999. p. 14. ISBN 1-887334-63-7. http://www.clir.org/pubs/reports/rothenberg/pub77.pdf

[5] XENA Tool http://sourceforge.net/projects/xena/;
Archivematica https://www.archivematica.org/wiki/Main_Page

[6] BORBINHA, José (ed.), et al. 2011. *iPRES 2011: 8th International Conference on Preservation of Digital Objects, 1.-4.11.2011, Singapore*. Singapore: National Library Board Singapore & Nanyang Technology University, 2011. 287 p. ISBN 978-981-07-0441-4. http://tinyurl.com/d3mxw56

[7] http://www.keep-project.eu/ezpub2/index.php

[8] BEARMAN, David. 1999. Reality and Chimeras in the Preservation of Electronic Records. *D-Lib Magazine* [online]. April 1999, vol. 5, no. 4 [cit. 2012-06-19]. ISSN 1082-9873. DOI 10.1045/april99-bearman. http://www.dlib.org/dlib/april99/bearman/04bearman.html

[9] GRANGER, Stewart. 2000. Emulation as a Digital Preservation Strategy. *D-Lib Magazine* [online]. October 2000, vol. 6, nr. 10 [cit. 2012-07-27]. ISSN 1082-9873. http://www.dlib.org/dlib/october00/granger/10granger.html

[10] HOEVEN, Jeffrey van der, Sophie SEPETJAN and Marcus DINDORF. 2010. Legal aspects of emulation. In: RAUBER, Andreas, et al. (eds.). *IPRES 2010: Proceedings of the 7th International Conference on Preservation of Digital Objects, Vienna, Austria, September 19-24, 2010.* Vienna: Oesterreichische Computer Gesellschaft, 2010, pp. 113-120. ISBN 978-3-85403-262-5.

[11] KryoFlux http://www.kryoflux.com/

[12] more about this research at http://www.openplanetsfoundation.org/blogs/2012-01-03-digital-archaeology-and-forensics or http://www.openplanetsfoundation.org/blogs/2012-07-17-conclusion-ctos-forensics

[13] COCHRANE, Euan. 2012. *Disk Image Read Failures*. 2012. p. 1. Internal document of Archives NZ, CMS ID A647018.

[14] COCHRANE, Euan. 2012. *Land Information New Zealand (LINZ) SALT Database: Migration from original HW to virtualised HW*. 2012. 13 p. Internal Document of Archives NZ, CMS ID A641672.

[15] QEMU emulator http://wiki.qemu.org/Main_Page

[16] COCHRANE, Euan. 2012. *Rendering Matters: Report on the results of research into digital object rendering*. Archives New Zealand: Wellington, 2012. p. 4. http://archives.govt.nz/rendering-matters-report-results-research-digital-object-rendering

[17] COCHRANE, Euan. 2012. *Recommendations for the Use of Rosetta For Digital Preservation at Archives New Zealand*. 2012. p. 7. Internal document of Archives NZ, CMS ID A574634.

[18] EX LIBRIS. 2012. *Rosetta Preservation Guide, v 3.0.1*. June 2012. p. 14. Internal Document of Archives NZ.