APPLYING THEORETICAL ARCHIVAL PRINCIPLES AND POLICIES TO ACTUAL BORN DIGITAL COLLECTIONS

Leigh Rosin
National Library of New Zealand

ABSTRACT

Few archivists would dispute the need to maintain archival integrity and to adhere to the principles of provenance when working with heritage collections. Understanding a collection's origins and ownership, retaining its original order and context and documenting the collecting institution's decisions as regards to arrangement and treatment are several well-established archival principles, which are applied by most archivists on a daily basis.

So well-established and inherent are these principles, that it seems to have been simply assumed that archivists would apply them to ALL formats in their collections, including those that are born digital. However, what happens when you try to apply such traditional archival principles to actual born digital collections? This paper will demonstrate my experiences of doing exactly that; Applying the National Library of New Zealand's (NLNZ) archival and digital preservation policies on one particular unpublished born digital collection of theatre and performing arts papers.

This particular collection has a number of features which made the application of traditional archival policies challenging: Its 1600 files were spread out over a number of different folders and sub-folders, which represented a context and original order that needed to be retained; Almost every file had some technical issue which needed to be resolved in order to easily ingest the files into the preservation repository; The library had in place a "preconditioning" policy which stated that all actions performed on files must be reversible and clearly documented; The library's ingest tool requires that all files loaded in a single deposit share the exact same metadata.

The features above resulted in this particular collection being a "perfect storm" of theory clashing with practice. It is an excellent example of what happens when you try to apply traditional archival theory in the digital world. The paper will outline the process, the challenges and lessons learned.

KEY WORDS: Digital preservation, Authenticity, Digital policy, Ingest

---

1. INTRODUCTION

1.1 THE ORGANISATIONAL CONTEXT

The National Library of New Zealand (NLNZ) has been actively collecting born digital heritage collections since the mid-1990s and in 2008, the library went live with its National Digital Heritage Archive (NDHA). The NDHA

encompasses several ingest and access delivery components and utilises the Ex Libris digital preservation software, Rosetta. The majority of the digital collections the library produces and acquires are stored within the NDHA. The NLNZ collects a wide variety of digital items including unpublished materials (photographs, cartoons, manuscripts and private papers, oral histories and interviews) and published materials (websites, electronic publications and music). Digital items are collected under the National Library Act (2003) and are also acquired by purchase or donation.

This paper will focus on unpublished collections. For such collections, the ingest and technical analysis activities are performed by Digital Archivists. The work is highly collaborative in that it often requires input from several stakeholder groups across the library, including digital preservation analysts, policy analysts, curators and arrangement and description librarians.

Following an initial technical analysis, unpublished digital collections are appraised by curators and, if a decision is made to retain, they are arranged (if necessary) and described by digital materials librarians. The collections are then ingested and stored in the NDHA for longterm preservation.

1.2 THE POLICY CONTEXT

In 2011 the NLNZ, in collaboration with Archives New Zealand (ANZ), began the process of establishing a set of policies to support its digital ingest and preservation activities. The policy which had the greatest impact and relevance for the Digital Archivists was the Preconditioning Policy, which states that “[t]he diverse nature of digital content means that there are times when it is desirable to make changes to it before it is ingested into the preservation system. These changes are classed under the term ‘preconditioning’ ”. (Joint Operations Group, Policy [JOGP], 2012)

The goal of the policy is to “describe the limits of change that can be introduced to digital content from the time it is brought within the control of Archives or the Library to its acceptance into the preservation system” (JOGP, 2012). These changes can include an alteration to (or the addition of) file extensions in order to better facilitate file format identification and/or the removal of unsupported characters in the file name to enable more stable storage of the files in the preservation system's storage database.

Regardless of the change being imposed, one must be able to demonstrate that the action will not affect the intellectual content of the file (JOGP, 2012). What is more, there are two key operating rules which must be followed: “All changes must be reversible” (JOGP, 2012) and all changes must have sufficient documentation to demonstrate the reasons they were undertaken as well as “a system-based provenance note that clearly describes the change that has been made to the file” (JOGP, 2012).

At this point, it is worth highlighting the business and technological context, which made such a policy necessary for the NLNZ and ANZ. Files do not require file extensions to be loaded and stored in the preservation system, nor

do files names have to be entirely consistent and devoid of special characters. However, all files ingested into preservation system go through what is called the Validation Stack (VS) in Rosetta. This is a series of technical checks[1] run against all files to determine whether they have any issues. Files with missing or incorrect file extensions will fail the VS and be re-routed to a technical workbench area for assessment. Within this workbench area of the system, there are limits to the actions you can perform and the tools you can use. The NLNZ has found that for collections with large numbers of files requiring a complex series of fixes to be applied, it is easier and more efficient to perform these actions prior to upload. Another consequence is that the files in the preservation system will be in a relatively stable and "clean" state, which the NLNZ and ANZ hope, will make them easier to report on and identify later. Such preconditioning activities might also make future file format transformations for preservation, easier and less labour-intensive.

In addition to this digital preservation policy, the NLNZ also adheres to such well-established archival principles as "original order" (National Archives of Australia [NAA], 2013), "provenance" and "conservation" (Society of American Archivists, 2013).

The concept of original order whereby "[t]he order in which records and archives were kept when in active use" be preserved (NAA, 2013) is of particular relevance when processing digital collections. While there is, as yet, no explicit policy statement that this principle must be applied to *digital* collections at the NLNZ, business procedures seem to have naturally evolved to include it. The implications of this will be explored in the following case study.

Similarly the concept of "authenticity" is relevant when processing digital collections and had a particular impact on the NLNZ's approach. Maria Guercio summarises its relevance to digital collections in her article *Principles, methods, and instruments for the creation, preservation, and use of archival records in the digital environment*: "The authenticity of a record, or rather the recognition that it has not been subject to manipulation, forgery, or substitution, entails guarantees of the maintenance of records across time and space (that is, their preservation and transmission) in terms of the provenance and integrity of records previously created" (p.251).

All of the above stated preservation and archival policies and concepts govern how digital collections are ingested and managed at the NLNZ. As the following case study will demonstrate, such policies have a significant impact on workflows, staff resourcing and business processes.

2. CASE STUDY: PROCESSING A COLLECTION OF THEATRE COMPANY RECORDS

In 2008, the NLNZ received a donation of records from a local Wellington-based theatre company. The collection was made up of 7 boxes, covering a time period of 1990-2008. The donation included both digital and non-digital collection items and was made up of:

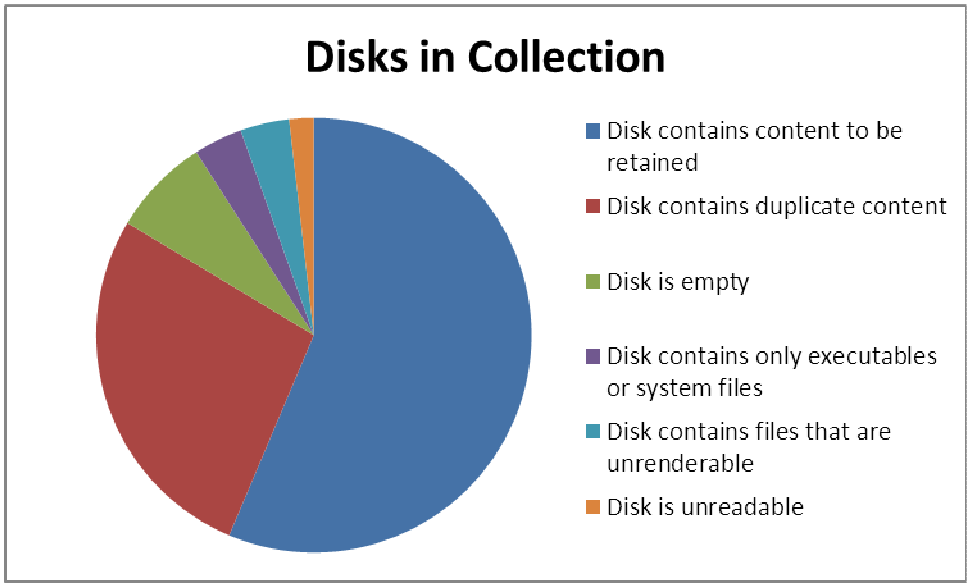
- Posters and production-related materials such as scripts and production notes
- Historical material including photographs of the theatre building and performers
- Administrative records including front-of-house box office sheets, correspondence including file of complaints, contracts with cooperatives, strategic plans
- Publicity materials including press releases, clippings and reviews

The digital component of the collection was made up of 55 disks, which included both Mac and IBM formatted 3.5 inch floppy disks, double density 3.5 inch floppy disks and zip disks.

The non-digital components of the collection were briefly assessed by curatorial staff and were deemed to meet the library collection policies. The digital components would require considerably more analysis to get the files into a fit state to be properly appraised, however based on the information provided by the donor and the labels on the computer disks, the collection in its entirety was formerly accessioned into the library's collections and was selected for permanent retention.

2.1 Media analysis and migration

All 55 disks were analysed, yielding the following results:



(31) disks contained files that met our collection policies and which we wanted to retain
(15) disks contained files duplicated on other disks, or files that did not meet our collection policies
(4) disks were empty
(2) disks contained only executables and system files
(2) disks contained files which could not be identified or rendered

(1) disk was unreadable

This resulted in 923 of the 1530 total files, being migrated from the original media and put into temporary storage on a secure server, awaiting further technical and curatorial analysis. In order to migrate the files from their original media, a number of different disk drives and operating systems were utilised.[2]

2.2 Technical analysis of the files

A technical assessment of the files was undertaken. This involved considerable liaison with the NLNZ's Digital Preservation Analyst, who used several tools[3] to identify the various file formats within the set of the files. This analysis revealed that, while most could be opened and read, there were various technical issues with the files, which included:

- Missing file extensions
- Incorrect file extensions
- Illegal or undesirable characters in the file name[4]
- Full stops in the file name

Of the 923 files being retained by the library, 700 (76%) were found to have some kind of technical issue.

Given the complexity of technical issues within this collection, planning began amongst the Digital Archivists, the NLNZ Digital Preservation Analyst and the Curator of the content as to how to address these problems. A number of activities were discussed and identified as preconditioning actions, to be performed before the files were ingested into the preservation system. Prior to each preconditioning action being conducted on the original files, tests were conducted on copies to determine if the fix applied would affect the intellectual content held within the files and would be reversible, in accordance with the preconditioning policy.[5] Staff also agreed how, and by whom, such actions should be documented.

The technical assessment, in addition to revealing technical issues, also found that the files were provided to the library in a particular folder structure created by the theatre company. These folders provided important context and arranged the files by theatre production, date and record function.

2.3 Applying Preconditioning actions

Copies of the 700 files requiring preconditioning activities were passed to the NLNZ's Digital Preservation Analyst for processing. A range of tools were used to add/change file extensions and remove/replace undesirable characters from the file name.[6]

A report was generated during this work to indicate the file original name, the new file name, the original file path and the change(s) applied to each file. A Python script was created to generate this detailed list. Based on the preconditioning action applied, the script automatically generated a brief

statement, describing the action taken. At the NLNZ, this brief statement is called a “provenance note” and forms part of a file’s core metadata. Provenance notes are created during the upload process using the NLNZ's ingest tool INDIGO[7].

2.4 Preparing to upload the files - Policy dependencies

As discussed, each preconditioning action must be reversible and well-documented. This means that each pre-conditioned file must be uploaded with enough metadata to describe every one of its preconditioning actions, in such a way that it could be undone, should it prove to be necessary to re-create the file exactly as it was when it arrived at the library.

Another policy dependency is an archival one, which concerns original order. The files were provided to the library in a particular folder structure, which provided important archival context to the collection. While the files would not be stored within the database in their original order, it was determined that the original folder path of each file was a key piece of information that should be captured as part of its object metadata.[8]

In the interests of better reporting on preconditioning actions in the future[9], a policy decision was made to create a different provenance note for each individual preconditioning activity applied. Take for example the following file from the collection:

Kieron doc. 5/5/92 [ORIGINAL file name]
Kieron doc-5_5_92.mcw [NEW file name]

This file was missing an extension and had characters in the file name, which were either undesirable or unsupported. Three preconditioning actions were applied to it, requiring three distinct provenance notes:

1. 002_File extension was added: The file was submitted by the donor without a file extension. An .mcw extension was added to this filename by the Preservation Analyst as it has been identified as Microsoft Word for Macintosh document version 4.0.
2. 003_special character(s) were removed from the file name: The file name was submitted with forward slashes as part of the file name. On the recommendation of the preservation analyst these have been removed and replaced with underscores.
3. 003_Special character(s) were removed from the file name: The file name was submitted with a full stop as part of the file name. On the recommendation of the Preservation Analyst this has been removed and replaced with a hyphen.

2.5 Preparing to upload the files - Technological dependencies

Even though INDIGO does not impose limits on how many files can be loaded in a single deposit, there are several technological dependencies inherent in the ingest tool, which affects the manner in which the files can be uploaded:

- All files in a single deposit must share the SAME provenance, descriptive and administrative metadata.
- In order to capture a file's original path, the file must be loaded from within its original folder structure.

2.6 UPLOADING THE FILES

Under the above technological and policy constraints, work proceeded to upload the files into the preservation system using INDIGO. A great deal of planning was required to ensure that files were uploaded according to their original file path, as well as according to the preconditioning action that was applied. Take, for example, a group of 9 files from the \FRINGE\FRINGE FEST 93\PRODUCTION folder, as illustrated in Table 1:

Table 1: Analysis of the variety of preconditioning actions applied to a single folder of files and implications for uploading.

ORIGINAL FILE NAME	NEW FILE NAME	FILE EXTENSION WAS ADDED:.MCW	FILE EXTENSION WAS ADDED:.XLS	"/" REPLACED WITH "_"	FULL STOP REPLACED WITH "_"	DEPOSIT
Bar p.a. contract	Bar p_a_contract.mcw	✓			✓	Deposit 1
Drama School Assessment	Drama School Assessment.mcw	✓				Deposit 2
Fran Wilde letter 15/2	Fran Wilde letter 15_2.mcw	✓		✓		Deposit 3
Winding up letter	Winding up letter.mcw	✓				Deposit 2
CO-OP NUMBERS	CO-OP NUMBERS.xls		✓			Deposit 4
Prod. roster	Prod_roster.xls		✓		✓	Deposit 5
Programme List	Programme List.xls		✓			Deposit 4
Rehearsal Sched.	Rehearsal Sched_.xls		✓		✓	Deposit 5
Techno list	Techno list.xls		✓			Deposit 4

Due to the variety in actions applied, this small group of 9 files were required to be uploaded in five unique deposits.

In this way, the loading progressed and eventually the 923 files required a total of 118 unique deposits to be created. These were then loaded into the preservation system. As a result, this relatively small collection, took one staff member a total of 52 hours to upload.

2.7 CASE STUDY – ANALYSIS

By processing complex collections such as this, the library identified a number of technology areas in which its tools could be further developed to create greater efficiencies. For example, the NLNZ is at present re-developing INDIGO to be able to apply provenance notes and descriptive metadata at a more granular level and to do so in an automated way. It is also working towards enhancing INDIGO so that file path information might be provided as a piece of contextual metadata without the need to upload files from their original folders.

These enhancements will result in greater flexibility when building complex deposits with varied metadata requirements. When this functionality is in place, a collection similar to that described in this case study could be uploaded in a *single deposit*. This will allow the required provenance actions and contextual metadata to still be applied to each individual file, but without the need to load those objects separately, thus saving dozens of hours in staff resource time.

Collections such as this also greatly inform the NLNZ as to how theoretical principles and policies are applied during day to day business operations. A plan has been put into place to review the library's digital preservation policies every two years. It may be that practical workflow issues raised during the implementation of such policies, as have been illustrated in this case study, will have an impact on existing policies.

In the course of implementing the Preconditioning Policy and creating provenance notes, which form part of a file's core metadata, the NLNZ have had to take a close look at how preservation metadata is stored within the Rosetta system. Metadata in Rosetta largely conforms to the PREMIS[10] model. We have come to the conclusion that the information we require in the provenance note to satisfy the Preconditioning Policy is at a level that is more detailed than is allowed by our current event metadata elements. As a result, the NLNZ have made a recommendation to the Editorial Committee for PREMIS around a new Provenance aspect of the data model, in order that our requirements for documenting preconditioning actions can be more accurately satisfied.

In addition to the policy and technological constraints, there were also several staff-related factors which can be seen as having contributed to the complex and time-consuming nature of the processing of this particular collection. There were several staff changes during the period of processing this collection, due to a sabbatical, a maternity leave, a staff resignation and a staff secondment to another project. This meant that the processing of this collection was handed over to multiple staff members of the course of several years. Each time this occurred there was inevitably time lost as a new staff member was brought up to speed on the collection and its complexities.

For all these reasons, it took a total of five years to appraise, ingest, describe and upload this collection.

2.8 CASE STUDY – LESSONS LEARNED

In addition to contributing to the NLNZ's understanding regarding its policies, tools and processes, this case study has also been a learning opportunity for staff. Inevitably, one gets to the end of the a complex project such as this and wishes they had made different choices in some respects, and are gratified with their choices in other areas. The following are several examples of lessons learned throughout the process of this case study.

First, when the collection was initially received, a copy of the files was made, in which they were removed from their original folders. This was to aid the initial technical assessment and get a sense of the kinds of issues within the file set. Files without extensions were divided up according to potential format (word processing / spreadsheets / images). This was strictly a working copy, meant to facilitate an easier analysis of the files. However, when the files were transferred to the Digital Preservation Analyst for preconditioning activities, it was this copy that was provided, rather than the original set of files, within their original folder context.

The implications of this decision became clear when the files were finished being “cleaned” and were being prepared for upload. With the pre-conditioned files now in a flat list, the “cleaned” files had to be re-integrated into their original folder structure prior to upload, which was a time-consuming exercise. This small error resulted in many extra hours of work.

Second, the arrangement & description of the collection was done long before the technical analysis and preconditioning actions took place, which proved to be somewhat problematic. In this particular case, some files were described but, following an in-depth technical analysis, the decision was made not to retain them for technical reasons. Thus time was spent working on items that actually were not retained. Fortunately, this only occurred for a small amount of files.

Third, taking a disk inventory immediately upon receiving the media and capturing a detailed listing of how the files were arranged, including original file paths and dates was extremely valuable. In this way, staff were able to confirm (or re-create) the original structure, as files were accidentally moved around. Also, at each step of the ingest process, actions and decisions were carefully documented in an ingest report. This report proved very important in explaining and contextualising actions and decisions to the new staff members.

3. CONCLUSION

As this case study illustrates, implementing archival and digital preservation policies to complex digital heritage collections can be difficult and labour-intensive.

We have observed how archival and digital preservation policies developed and implemented within institutions can have intended and unintended consequences on business processes and workflows. Concepts and rationales that are perfectly clear and logical in the abstract may be difficult to implement in reality. Principles that are well-established and work well for traditional, non-digital collections may be challenging to apply to digital collections. Tools and technology that work well for homogenous, modern digital collections might not cope well with varied, complex legacy collections.

The relationship between policy and implementation is described very aptly in *Born Digital: Guidance for Donors, Dealers, and Archival Repositories* (2013): "The unexpected will continue to challenge and surprise repositories acquiring and managing born-digital materials, despite reasonable efforts at creating clear and actionable policies" (p.32).

Heritage institutions should give careful consideration as to how (or whether) traditional archival policies apply to digital collections and how these policies can be applied in bulk. The collection illustrated here is a relatively small one but the amount of effort required to process it is alarming. As digital collections become larger and more complex, we must identify a way that such archival policies can be applied in a way that does not compromise efficiency.

While ensuring policies are regularly reviewed, endeavouring to enhance tools to make workflows as efficient as possible and speaking to donors of digital collections in advance of donations may all contribute to more effective and stream-lined processes, one must always be prepared to acknowledge that "[t]he stewardship of born-digital archival collections promises nothing if not routine encounters with the unexpected"(Born digital, 2013, p.32).

REFERENCES

Born digital: Guidance for donors, dealers, and archival repositories (2013). MediaCommons Press: Retrieved April 22, 2013 from http://mediacommons.futureofthebook.org/mcpress/borndigital/

Guercio, Maria, *Principles, methods, and instruments for the creation, preservation, and use of archival records in the digital environment*. American Archivist 64:2 (Fall/Winter 2001), p. 238–269.

Joint Operations Group, Policy. Department of Internal Affairs (2012). *Digital content preconditioning policy* (p. 1-2).

National Archives of Australia. *Glossary*. Retrieved 1 May, 2013 from http://www.naa.gov.au/records-management/publications/glossary.aspx

National Library of New Zealand (Te Puna Mātauranga o Aotearoa) Act 2003. Retrieved April 22, 2013 from http://legislation.govt.nz/act/public/2003/0019/latest/DLM191962.html

Society of American Archivists. *Glossary*. Retrieved 1 May, 2013 from http://www2.archivists.org/glossary/terms/p/provenance

NOTES

[1] The Validation Stack includes file format identification (by DROID); file validation (by JHOVE); metadata extraction (NLNZ metadata extract tool and JHOVE); checksum (CRC32, SHA-1, MD5); virus check (CLAM AV)

[2] Several different machines and operating systems were used during the media migration. These included: Mac Performa 580CD (Mac OS 7.5.1); iMac (OS 9.2 / OS 10.1.2); HP Compaq (Windows XP); Toshiba 3.5 inch floppy disk drive (USB FDD kit); Iomega zip disk drive

[3] The following tools were used to identify the files: DROID, TriD, Fido, Hex editor (Neo Hex), Reference Objects (trusted samples of particular file types for comparison), Reference Standards (formal file format release papers that describe how a specific file type is made / implemented)

[4] One such character, the "/", had to be removed during the media migration process and prior to even storing the files on our secure server due to the fact that Windows does not support it as a file name character.

[5] A variety of methods were employed to ensure the intellectual content was preserved following pre-conditioning activities. These included running checksums and performing rendering checks (ensuring the file can be opened in the target application and behaves as expected).

[6] A variety of tools were used to rename files. These included: Bulk Renaming Tool and Python (for generating scripts).

[7] INDIGO is an internal submission tool, developed by the NLNZ, to integrate with the digital preservation system. It is a desktop application used by various business units to create deposits of files and metadata and upload them to the Rosetta digital preservation system.

[8] This original order was also retained and used during the arrangement and description process. Files were described according to their original order, as evidenced by their folder structure.

[9] A business rule was established identifying a syntax for provenance notes so that all notes are phrased in the same way. Each note begins with a standardised prefix, so that

preservation staff will be able to easily report on (and retrieve) objects on which the same types of provenance activities were performed.

[10] PREMIS. Preservation Metadata Impementation Standard. http://www.loc.gov/standards/premis/]