Then and Now: The Evolution of Digital Preservation and Collecting Requirements Over a Decade

Leigh Rosin
National Library of New Zealand
P O Box 12349
Wellington 6001
+64 4 474 3077
Leigh.Rosin@dia.govt.nz

Kirsty Smith
National Library of New Zealand
P O Box 12349
Wellington 6001
+64 4 474 3077
Kirsty.Smith@dia.govt.nz

ABSTRACT

This paper reflects on a decade of digital collecting and digital preservation development at the National Library of New Zealand. It will examine the workflows, policies and tools that have been developed in the decade since the funding for the National Digital Heritage Archive was received. The paper will look closely at the requirements that were identified for the initial development of the digital preservation system and compare them to the status of the current preservation programme and requirements roadmap.

General Terms

Preservation strategies and workflows, Case studies and best practice

Keywords

Digital preservation, Requirements, Digital policy, Ingest

1. INTRODUCTION

The National Library of New Zealand (the Library) has been actively collecting born digital heritage collections since the mid-1990s. During these first years, processes were still being developed and there were very few organisational policies governing the management and preservation of digital collections. As an organisation we were experimenting, learning and trying to figure out how to deal with these new kinds of collections.

In 2003, the governing legislation was revised, providing the Library with the legislative mandate to collect and preserve digital content under legal deposit.

The following year, government funding was secured for the National Digital Heritage Archive (NDHA) Programme. The goal for this programme was the establishment of a digital archive that would enable the Library to meet its mandate to collect, make accessible and preserve in perpetuity New Zealand's digital heritage.

The NDHA programme spent some 18 months gathering extensive business and functional requirements. In 2006, the Library formed a development partnership with Ex Libris to build a digital archive and preservation management system. The resulting Rosetta system was launched in October 2008, and for the Library the ingest and preservation of digital material became a 'business as usual' activity. To date, the archive holds approximately 5.5 million objects, spanning across 137 different formats and consisting of approximately 50TB.

Throughout the requirements and development phases of the project, the Library continued to create and acquire digital collections. Workflows, guidelines and policies developed and evolved and continue to do so, even today.

This paper will reflect on the past decade of digital collecting and digital preservation development at the Library. We will examine the workflows, policies and tools that have been developed in the decade since the funding for the NDHA was received. We will look closely at the requirements that were identified for the initial development of Rosetta and compare them to the status of the current preservation programme and requirements roadmap.

Four key functional areas will be used to drive this comparison: Ingest and acquisition of digital collections; Content maintenance; Format library; Preservation planning and execution.

Simply, we will ask the question: If we knew then what we know now, how different would our requirements and processes be?

2. INGEST OF DIGITAL COLLECTIONS

2.1 Depositing Methods

When the Library first put together its requirements for the ingest of digitally born material, they were based on a theoretical workflow model. The assumption was that content producers would 'push' digital content to us, and therefore we created an area of the system through which we could manage individual producers' details and their deposit arrangements. These arrangements would outline what they intended to provide, how they preferred to send the files and when. These arrangements would allow us to personalize the depositing experience for the content producer. Our deposit tools would be geared towards supporting this external depositing experience, and would be set up to allow producers an easy, web-based interface by which they could provide us with files and metadata.

The reality since go live has strongly tested this theoretical workflow assumption. Content producers to date have by and large preferred to make content available for legal deposit via their existing communication and/or distribution channels - websites, email subscription lists etc. They prefer to email the files or to let us know where a copy is available for download and we have had minimal uptake of the web deposit functionality. Therefore we have a "pull" rather than a "push" workflow, whereby Library staff do the bulk of the depositing for digitally born content, which continues to be acquired by the usual distribution channels.

As a result, we developed a web deposit tool as well as an area of the system where content producers could manage their submission arrangements and personalize their depositing experience, that is largely being unused by external depositors. Since library staff are doing the bulk of the depositing, we were creating and maintaining rich personalised producer accounts within our preservation system, that don't support the ingest of material and duplicate data held in our acquisitions system. Given the volume of material legal deposit staff are processing we are now moving to streamline the staff mediated workflow by associating deposits to a generic library producer, thereby avoiding the need to create and maintain individual producer records.

2.2 Ingest Tools

Since we were working under the assumption that depositing would largely be undertaken by external producers, the requirements for our ingest tools for staff did not initially include bulk, automated functionality for uploading born digital collections. Take for example our ingest tool INDIGO[1]. Our initial requirements for INDIGO were largely focused on supporting our internal digitisation programmes, who we imagined would be the primary users of the tool. While it allowed for the uploading of born digital collections as well, its main functions were tailored for allowing simple, homogenous objects (such as high resolution Tiff files), to be sent to Rosetta with minimum metadata requirements.

Several factors caused us to re-evaluate our requirements for our ingest tools over the past several years. First, as has been previously mentioned, was the "pull" nature of collecting, which was putting pressure on the Library to create easy, automated workflows for staff to use when uploading collections.

In addition, there was also a factor relating to staff confidence. In the early days of the NDHA, when the electronic deposit workflow was new, the manual nature of deposit tools was less of an issue for staff. New systems and new types of content meant there was a high degree of caution on the part of staff, and there was a desire to manually check everything that was being deposited. Therefore the need for an ingest tool like INDIGO to easily support automated workflows for depositing born digital content was minimal. However, as staff have become more experienced and increased their technical knowledge, their needs and requirements began to change. They became more confident both in their abilities and in the preservation system and they began to shift their requirements. They became interested in exploring tools and workflows that would result in the largest amount of files being uploaded with the minimum amount of manual intervention from staff.

Finally, the more staff worked with the tools they had and expanded their knowledge about digital collections, the more they were able to imagine and articulate how such ingest tools could be enhanced. Working with INDIGO for several years allowed staff to evaluate the areas of the tool that worked well, and the areas where the tool could be improved to create more automation and efficiencies. Staff were able to see more clearly how the tool could be changed to allow for more complex objects to be loaded; how functions could be altered to allow for more varied metadata inputting. They were also beginning to see how staff time could be saved by engineering the tool to do the bulk of the work. These are ideas and requirements that grew out of staff experience and have resulted in five new iterations of INDIGO being developed since 2009.

2.3 Ingest Activities and Documentation

The ingest and technical analysis activities performed during the acquisition of born digital collections, as well as how we document those activities, is another area where our policies and processes have changed greatly over time.

In the early days of acquiring digital collections, there was very little use of tools or digital forensic technologies. Our main goal was to migrate files off original media and get them onto a secure server. However, the methods used were pretty basic. For example, when files were copied there were no checksums generated before and after copying, therefore making it difficult to ascertain whether changes to the files had occurred during the transfer process. This is an area where the Library's processes are being re-developed, to ensure the authenticity and integrity of the files can be maintained throughout the transfer process.

Several tools[2] are being trialed by staff in an effort to make improvements to our processes. A fixity policy for the library, which will govern the handling of digital collections, has also had an impact. Although at the time of writing this article the policy has not yet been signed off, staff are preparing for it already. We need to prepare our processes so that they can meet the main fixity policy goal, which is "[t]o ensure that all content under the control of the Library or Archives can be, and is monitored for corruption and unauthorised change.". [7]

The documentation of ingest and technical analysis activities is another area where many changes have occurred. Staff have always described their activities, but ingest reports and file listings were sometimes missing key details about hardware/software used, actions taken or methods trialed. There was a lack of consistency and it was often difficult for staff who would later work on these collections to know what tools had been tried, what actions had been taken and why.

This continues to be a challenging area, where considerable improvement is still necessary. Although reporting is more consistent now (templates are used by all staff) and key details and actions are better documented, the process continues to be extremely labour-intensive and manual. Staff continue to try and improve the way they document their activities to make the process more efficient.

3. CONTENT MAINTENANCE

One of the Library's initial business rules underpinning system requirements was that objects would not be 'touched' prior to ingest. Any maintenance actions that needed to be undertaken would only be done within the confines of the preservation system, where they would be auditable. The initial data migration quickly highlighted the constraints of such a business rule. Our data was far less 'clean' than we imagined. During ingest, files

[1] INDIGO is an internal submission tool, developed by the Library, 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

[2] Shotput Pro, FreeCommander and TeraCopy are several examples

are run through a validation stack, where a series of technical checks are run (virus checks, fixity, format identification and validation). Where we had imagined ‘unclean’ files would be the exception, it was immediately apparent that a large percentage of our data triggered errors in the format identification (DROID) and validation (Jhove) tools. While system requirements quickly evolved to enable rules to be set to ignore certain tool errors, not all errors were ones that we wanted to ignore.

While files with missing or wrong file extensions can be ‘ignored’ and ingested into the preservation system, the Library felt it was preferable to load files in a relatively clean and stable state. Once files are in the preservation system, there are limits to the actions you can perform and the tools you can use. The Library has found that for collections with large numbers of files requiring a series of fixes to be applied, it is easier and more efficient to perform these actions prior to upload. As a result the Library adopted a pre-conditioning policy; this sets the limits of change that can be introduced to digital content from the time it is brought within the control of the Library to its acceptance into the preservation system. Three key operating rules underpinning the policy are:

- Changes cannot be made on the intellectual message of the object,
- All changes must be reversible 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. [6]

Throughout its implementation of the preconditioning policy, the Library has been rethinking its requirements in respect of provenance data. Metadata in Rosetta largely conforms to the PREMIS model, and thus Rosetta’s current data model supports provenance event data. However the information we require in the provenance note to fully satisfy our preconditioning policy is at a level that is more detailed than is allowed by existing metadata elements. As a result, the Library made a recommendation to the PREMIS Editorial Committee for the inclusion of a more granular Provenance metadata element. The PREMIS Editorial Committee added a new semantic unit as a place where such information could be stored in whatever structure an institution requires.

Thus, over the past decade, the Library has eased its stance on performing activities on files prior to their ingest into the preservation system and created policies and workflows to support this position. This has also resulted in updates to certain metadata elements.

4. FORMAT LIBRARY/REGISTRY

One area where the Library’s requirements have changed significantly is the Format Library. When our requirements were first compiled, the Library had a fairly simple understanding of what the role and scope of a Format Library should be. The main requirement was for a library that would document formats, and link that information with supporting applications in order to identify preservation risks. Requirements were based on the assumption that most of the detailed format information would be drawn from existing registries, principally the National Archives UK’s PRONOM database.

Over the last five years, we have gained a great deal of experience through interaction with collection items and use of external information resources such as PRONOM. [1]

These experiences are often mediated through tools developed by the community, thus giving us further insight into gaps or failings. All of these experiences have highlighted the “need to be able to more accurately define formats, relate them to relevant specifications, define their supported characteristics, and combine these things to form profiles that can be linked to software applications.”[8] A greater understanding has led to a much more complex set of requirements.

A priority for the Library this past year has been the development of a set of requirements for a Digital Preservation Technical Registry; one that would exist as a community resource able to be used in conjunction with any digital preservation repository. This work has been undertaken under the aegis of the National and State Libraries of Australasia and has a project team comprised of the Library, the National Library of Australia, the National Records and Archives Administration (US), Archives New Zealand and the University of Portsmouth.

The Technical Registry will bring together technical information sources that currently are separate. This includes descriptions of file formats, the software applications used to create or render them, the hardware and operating systems that support the applications and files, and the perceived risks they face. It is planned that this will become the defacto hub for the rich and complex technical information and tools required to undertake digital preservation as professional activity. The Registry will benefit members of the digital preservation community through offering efficient information retrieval from one central resource, supplying trusted information and finally, supporting a community that will promote collaboration, develop best practices and peer review Registry information.

While the requirements for the Technical Registry are far more complex than those developed for a format library six years ago, a move towards a global technical registry would see a corresponding simplification at the local format library level. While a local library would have a dynamic relation to the technical registry, it is our expectation that only a relatively small subset of data would be copied to the local level. The digital preservation system retaining just enough information to support identification and reporting.

5. PRESERVATION

While preservation functionality is central to the Library’s preservation system, it is the one part of the system that we have used the least. We therefore feel we are a long way from being able to specify what a fully featured preservation workflow would be like.

We are currently in the process of planning and testing preservation actions for two quite different sets of data. The first candidate, a set of Word Star files, and the second, all of the Library’s web harvest Arc files. The two sets of data sit at opposite ends of the preservation spectrum: The first is a small set of files that requires boutique level preservation while the second is a large set of homogonous files requiring bulk conversion. For different reasons, both sets of data have challenged some of our original requirements/assumptions.

The Library had based requirements on the idea that the preservation system should be the primary collection tool for information required against which a preservation decision is made. However as we have worked with our curatorial staff on the Word Star content to identify the 'intellectual' aspects of the content that we need to ensure are preserved, it has become apparent that the level (both quantity and detail) of information collected as apart of a preservation plan is far greater than envisioned, and is better generated outside of a preservation system.

In the case of the large scale preservation action, converting Arc files to Warc, this has highlighted that the amount of technical data we are able to generate on the conversion process and file characteristics is far more than initially envisioned, and a lot more than can be interpreted by inbuilt technical evaluation criteria. Our original requirements for technical evaluation criteria were based on the assumption that they would be limited to only data that metadata extractors could pull out.

Another of our original assumptions that we are revisiting is the idea that, wherever possible, preservation actions should run within the preservation system. The sets of data we are currently working with have, for different reasons, caused us to rethink this idea. The Word Star preservation action is one that essentially involves handcrafting a small number of files, and it is only practical to do that external to the preservation system. In the case of the bulk conversion of Arc files to Warc, the conversion tool could be added as a simple plug-in tool and run within the Rosetta framework. However, the Library does not run a stand alone preservation system, but one that supports the day-to day collection work of the Library and the delivery of content. The Library's current system architecture and hardware infrastructure is not at the point where it can support large scale preservation processing, without impacting the performance of other system areas such as ingest and delivery.

When we first started articulating our requirements we thought we knew how we would perform preservation actions. Now that we have started to plan and test preservation actions on real content it has become apparent that we cannot as yet run transformation processes that will follow a set pattern. We do not know enough about the content in terms of all its idiosyncrasies, what acceptable change is for all types of content (or conversely what some would call the significant properties). We do not fully understand the processing requirements for each transformation and the method of undertaking it that least impacts the other library processes that depend on the system. In short, if we were preparing requirements for preservation functionality now, we expect they would be a lot simpler and less prescriptive than our original ones.

6. CONCLUSION

This has been a brief view of some of the changes in requirements that we have had since the inception of our preservation work ten years ago. Clearly, we have not included everything, nor even hinted at the scope of the changes across this decade (an entire article could be devoted for example on hardware requirements). It is clear though, that the experiences of creating initial requirements, developing a preservation system and processes and working with them as business-as-usual for five years has afforded us a different (if not better) view of what our real requirements are. We started with a theoretical, almost academic view of what we wanted our world to be. Our requirements are now shaped by business need and, as such, are focused on practical, efficient, effective processes, thus making our requirements more pragmatic.

7. REFERENCES

[1] Gattuso, J. 2012. *Throughput efficiencies and misidentification risks in DROID*. http://ndha-wiki.natlib.govt.nz/ndha/attach/ReadingResources/MSB%2BDROID%20v1_05.pdf

[2] Gattuso, J. (2012a).*National Library of New Zealand-DROID, PRONOM Developments at the National Library of New Zealand.* Paper presented at Preservation and Archiving Special Interest Group (PASIG), Dublin. http://lib.stanford.edu/files/pasig-oct2012/04-Gattuso_PASIG_presentation_2012.pdf

[3] Gattuso, J. (2012c*). Full results for DROID version 6 "Max Byte Scan" test*. http://ndha-wiki.natlib.govt.nz/ndha/attach/ReadingResources/MBSResults.pdf

[4] Gattuso, J. (2012d). *Evaluating the historical persistence of DROID asserted PUIDs*. http://ndha-wiki.natlib.govt.nz/ndha/attach/ReadingResources/Historical%20View%20of%20format%20via%20DROIDv4_2.pdf

[5] Gattuso, J. (2012e). *Main results of the DROID version tests*. http://ndha-wiki.natlib.govt.nz/ndha/attach/ReadingResources/historical%20view%20droid_results.pdf

[6] Joint Operations Group, Policy (JOGP). Department of Internal Affairs. 2012. *Digital Content Preconditioning Policy*. Pp 1-2. [This is an internal policy document, however it will be made available upon request. Contact Peter.McKinney@dia.govt.nz to make a request.]

[7] Joint Operations Group, Policy (JOGP). Department of Internal Affairs. 2014. *NDHA Draft Policies – Fixity*. P1 [This is an internal policy document, however it will be made available upon request. Contact Peter.McKinney@dia.govt.nz to make a request.]

[8] National and State Libraries of Australasia (NASLA). 2013. *Digital Preservation Technical Registry System Vision Document.* [This is available upon request. Contact Steve.Knight@dia.govt.nz to make a request.]

[9] Rosin, L. 2013. *Applying theoretical archival principles and policies to actual born digital collections*. http://ndha-wiki.natlib.govt.nz/ndha/attach/ReadingResources/Applying%20theoretical%20archival%20principles%20to%20actual%20collections.pdf