Preservation Actions; where we started, and where do we go from here?

Jay Gattuso

National Library of New Zealand Te Puna Mātauranga o Aotearoa

Background

The National Library of New Zealand (NLNZ) has been in the business of digital preservation for well over a decade now. Our journey with digital collections in earnest began in 2003 (with changes to foundational legislation), but like many memory institutions we had informally acquired a variety of digital content long before our more formalised digital collection strategies were in place. With digital collecting the need to look after digital items follows at a close distance, and accordingly by 2004 we had established a project that would result in a purpose built digital preservation repository, designed to offer the insight and collection care required of digital heritage materials.

In 2010 the project was completed, and the National Digital Heritage Archive (NDHA) went from project into production, where it currently exists conceptually adjacent to 65[1] kilometres of stacks that comprise our physical collections.

Within the digital collection sits a growing estate of old and new digital items. We have tens of terabytes of web harvests, through to megabytes of text files created in early 1980's applications, and all shades of files types in between.

During the transition from a fledgling project, to a fully featured digital preservation programme the NDHA team needed to explore what it meant to "do" digital preservation as a day-to-day activity inside a national library. The products of this exploration can be found in much of the core functionality found inside Rosetta[2], the digital preservation software built in conjunction with Ex Libris through the NDHA project.

Much of this exploratory thinking was deeply speculative while also leveraging heavily on prior work, with direction, suggestions and hints coming from works that preceded the project[3,4,5].

---

[1] National Library of New Zealand (2016) Strategic directions to 2030. Available from: https://natlib.govt.nz/about-us/strategy-and-policy/strategic-directions (accessed 22 June 2017).

[2] ExLibris (n.d.) Rosetta - Digital Management and Preservation. Available from: http://www.exlibrisgroup.com/category/RosettaOverview (accessed 22 June 2017).

[3] Consultative Committee for Space Data Systems. 2002, Reference Model for an Open Archival Information System (OAIS), CCSDS 650.0-B-1: Blue Book. http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf. (accessed 22 June 2017).

[4] CEDARS Project. 2002 http://www.leeds.ac.uk/cedars/ [No longer accessible. Archived http://web.archive.org/web/20041011141405/http://www.curl.ac.uk/projects/cedars.html. (accessed 22 June 2017)]

[5] Garrett, John and Donald Waters, co-chairs, et al. Preserving Digital Information: Report of the Task Force on Archiving of Digital Information, Commission of Preservation and Access and Research Libraries Group, May 1, 1996

As the project team started to focus on the requirements of a preservation "system", there was very little established knowledge to draw from. This led the project team to speculate on what the future would look like, addressing such thorny issues as "what will we know about formats, and their associated risks?", "how will we know that we have content with an associated technical risk?", "how will we undertake risk mitigation activities on file like objects?", "how will we process items and workflow tasks", "who is the intended audience of this work?" and so on. Some of these questions resulted in design decisions, and with our ten year vantage we now have an excellent opportunity to revisit some of that early speculation and apply some new understanding drawn from our experience.

We can skip forward in time, and find ourselves in the fortunate position of starting to address these questions from a more practical angle, leaning less on theoretical constructs and increasingly focusing on the tangible outputs of experimentation and practical experience.

The first release of the Rosetta product was moved into a production state for us in October 2008, and most of the intensive and complicated thinking that drove the early system design phase has found its way into the tool in various stages of completeness. We are now able to leverage 10 years of collecting technical data about the file objects in our care to start to apply some of that speculation to practice, and revisit some of those questions we asked ourselves when we set out to build our digital preservation repository.

A key proposition that made its way into the heart of the system design was based around *how* we might undertake the technical processes we have envisaged. From an architectural perspective, we speculated that the system would be somehow omnipotent and monolithic. Complex preservation operations on content would be undertaken from within the systems, and informed by internalised quality assurance processes that would be constructed to support decision making, analysis and other tasks. However, our experience to date indicates that this might not be such a viable proposition as we find ourselves increasingly undertaking processes outside of the system for technical, philosophical and efficiency reasons. This eventuality demarks a clear deviation from our originally speculated approach.

In the early system design phases we envisaged the preservation action processing to be a largely linear mechanism. Content would be collected, ingested, and any technical risk to renderability would have some preservation activity applied. Typically we speculated on a migration process, resulting in a new version of the artefact being created. This was envisaged as a very controlled linear process. Version 1 would become Version 2. Version 2 would become Version 3 and so on. A loose approximation of this model looks like this:

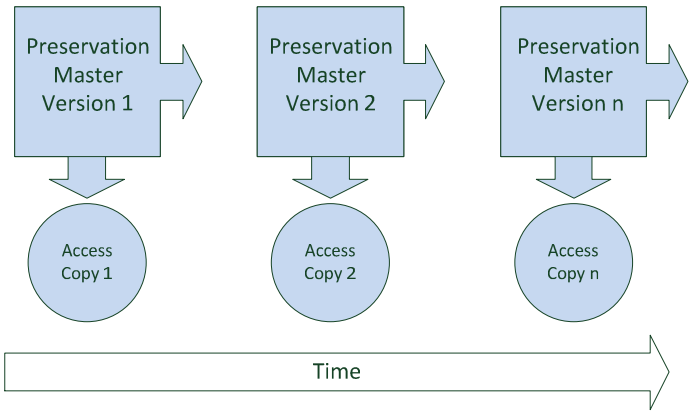

Figure 1 – Approximate Migration Workflow

The access copies may not be required, but conceptually and systemically they are a valid object that might comprise part of the original objects' expression.

We are increasingly interested in maturing this view point, and this conversation will be expanded on later in this paper.

We recognise that there will invariably be capability offered by alternative preservation strategies, for example emulation, that will complement our capability as the tools and technology matures. As this is a maturing technology in the digital heritage space, we can't really sensibly speculate on how this looks through generations of use. For comparison, and convenience we can conceptualise the emulation based workflow in simple terms as looking approximately like this:

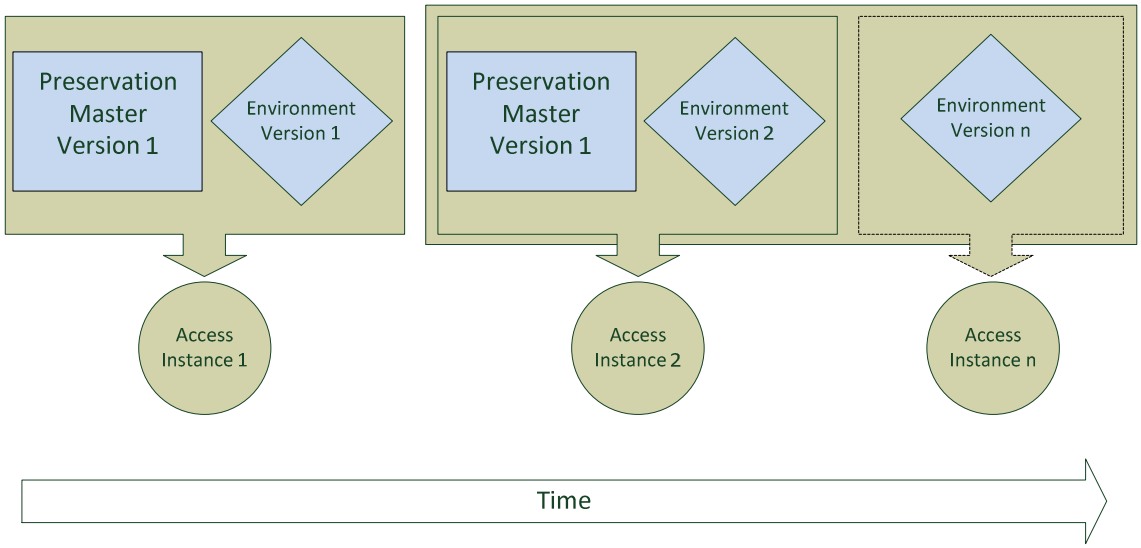

Figure 2 – Approximate Emulation Workflow

In this approximation there is no access copy, the rendered instance of the preservation master via the emulated environment is the authoritative rendered instance of the object. We also anticipate the future states of this approach are likely to require nested emulation environments, with each layer addressing technological step changes that are found at large. This might be a fresh instance of an emulation stack that contains all the required prior machinery, or nested instances of related environments that help traverse technological strata.

We are increasingly comfortable that references to a "digital-preservation-system" refer to at least two levels of systemic capability. At times we refer to the Rosetta platform as our preservation system and at other times the preservation system is Rosetta, all of our attendant storage and pre-processing servers, and the people and their evolving toolsets that undertake the ongoing maintenance of the technical infrastructure and the digital collections.

We are (still) obsessional in our commitment to understanding file formats, and how they need to be understood, recorded, or otherwise meaningfully addressed as platonic ideals, sets of technical instructions and instances of digital content.

We still care deeply about risk detection, but have learnt that we do not yet have the tooling and insight to progress tool based / automatic risk detection, let alone mitigation. It is an area we expect to keep pushing into, however we are comfortable that at this time we need to see a maturing of tools and thinking to properly discharge this type of processing.[6]

We are slowly starting to unpick what it means to undertake technical risk mitigation, and are working in a few directions addressing both quality and preservation concerns, as well as embedded "risks" that come with the territory of building systems that leverage a large variety of open source and homebrewed applications (for example JHOVE[7], DROID[8], FFmpeg[9], Safemover[10] and many other tools and scripts that are routinely used to deliver operational activities).

Those learnings, and perhaps shifts in position from our early thinking to where we find ourselves today are the focus of this paper. We intend to walk through some of the larger propositions and constructs we formed ten years ago, and reflect on what we make of them today, with experience of meaningfully working towards long-term safekeeping of digital collections. These learnings will be framed in three key areas: controlling ingest; controlling future state; and, linking technical operations to the intellectual endeavours of the organisation. The sum of these areas broadly represents our current position on the process of technical risk mitigation, and what that means for us going forward towards perpetual access to digital materials.

---

[6] Our initial considerations on risk analysis leaned heavily on institutional rendering capabilities (see DeVorsey, K. & McKinney, P., 2010. 'Digital Preservation in Capable Hands: taking control of risk assessment at the National Library of New Zealand'. *Information Standards Quarterly*, 22(2), pp.41–44.; DeVorsey, K. & McKinney, P., 2009a. One Man's Obsoleteness is Another Man's Innovation: A Risk Analysis Methodology for Digital Collections, paper presented at *IS&T Archiving*. pp. 101–106. Washington D.C., USA. That is, what can the Library render 'correctly'. This is a practical view of risk that can be based on local testing, rather than the more abstract notion of sustainability factors. However, while there is system functionality to give such a risk view of our collections, we have not systematically updated the information required to support the functionality. Our current preservation programme is still in its relatively infancy. For the moment, this is taking precedence over revisiting our view on risk analysis; the preservation programme should give us a far more complex and robust view of what future risk analysis routines will be.

[7] JHOVE - JSTOR/Harvard Object Validation Environment (n.d.) Available from: http://jhove.sourceforge.net/ (accessed 22 June 2017).

[8] The National Archives (UK) (n.d.) File profiling tool (DROID). The National Archives, Available from: http://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/ (accessed 22 June 2017).

[9] FFmpeg (n.d.) Available from: https://ffmpeg.org/ (accessed 22 June 2017).

[10] SafeMover (n.d.) Available from: https://github.com/NLNZDigitalPreservation/Safe_mover (accessed 22 June 2017).

Controlling ingest

Arguably, one of the most interesting shifts in position for us lies in the way we are attempting to head off some technical risks before content is committed to the permanent repository. Early considerations required us to address what can be loosely described as adhering to the "cult of the artefact". That is to say, that our mechanised workflow, direction-setting policy and operational activities pivoted around the notion that the bits we get are the bits we keep. Forever**.** We expect to ingest an original item following a well explored process that supports an integrious record of the original item. We then expect to operate on at risk originals to generate 2nd (and ultimately nth) generation versions that become the defacto original items. The originals are never removed from the collections; they are simply "versioned" away from typical consumption, accessible if needed.

It didn't take long for us to find some rough edges in this position, and the first sight of a problem arrived with the labels that came with those bits. These problems arose first with filenames and file extensions. Next came issues with file dates; what do we do when a given file purports to predate computing, or come from the future?

In a related series of concerns, we frequently encounter files that are technically malformed in their construction. These cases include various examples like extraneous data added to the beginning or ending of a file, missing or damaged technical markers expected in a given file format (e.g. a damaged or missing "End Of Image" marker in JFIF JPEG file).

All these cases are currently being addressed via development of a "preconditioning policy"[11] which describes what treatments[12] we are permitted to undertake on an incoming file, with a view to stabilising it before we commit it to the permanent file store. The policy describes use-cases that are viable for this kind of treatment, what processes this treatment is allowed to include, and most importantly how we record the detail of the treatment in a consistent and predictable way.

This development has become one of our most useful tools, both intellectually and technologically. By allowing ourselves to purposefully, carefully and sensitively make reversible and audited changes to files (and file metadata) we are able to operate our bigger less forgiving machinery faster and more accurately. Prior to having the policy, and tools to support its use if we had attempted to ingest a file containing a caret ("^") we would have expected at least one of our processing tools to have struggled with the presence of this reserved character[13], failing its task in any number of predictable ways. Having established an authorising workflow that allows this caret to be removed or replaced as needed prior to the item reaching the heavy machinery means we can keep the loading chutes clear, and expect cleaner ingest routines.

---

[11] National Library of New Zealand (2012) Digital Preservation Policy Manual. Available from: https://digitalpreservation.natlib.govt.nz/assets/NDHA/About-Us/Strategic-Partnerships/Digital-Preservation-Policy-Manual.pdf (accessed 22 June 2017)

[12] The word "treatments" is pointedly used in this context. We are increasingly interested to note that there are vast and compelling overlaps between the work we expect to undertake on digital items, and broadly accepted traditional physical conservation methods and concepts.

[13] Naming Files, Paths, and Namespaces (Windows) (n.d.) Available from: https://msdn.microsoft.com/en-us/library/aa365247.aspx#basic_naming_conventions%22 target=%22_new%22 (accessed 22 June 2017).

We identify problems early, address where possible within the rules of the policy and ingest the item with record of our actions.

We are incrementally learning about the impact of having this policy as we become more comfortable with its early-practice driven roots, and we expect to make further progress in this area in the future as our tools, thinking and processes mature. This approach is currently relatively experimental, finding a place somewhere between file format normalisation[14], and what we would regard as an intrusive preservation action[15].

By controlling the ingest process in this way we are attempting to reduce the complexity / variability of ingest issues, resulting in the successful ingest of binary items that have an accurate technical provenance record, and are *less* likely to cause us technical concerns in the future when we expect to operate at scale on similarly constructed binary items. It's not a done deal – it is extremely unlikely that we will neutralise all threats related to file format malformation, but by preconditioning our digital objects to the best of our current abilities, we give ourselves the best possible opportunity for implementing effective preservation actions in the future.

Controlling future state

A large part of digital preservation thinking revolves around the future consumption of digital content in a meaningful and accurate way. Digital preservation is the organisational defence mechanism that expects to protect the information found inside a binary object from the relentless march of technological advances. The digital preservation community has a growing list of concerning experiences that documents how information has been lost through the wane in popularity/usage of hardware, software and other related technologies [16,17].

There are a number of documented strategies[18] that aim to tackle this problem, and there is no doubt that as we continue to make technological advances new theories, methods and capabilities will emerge. The digital preservation community as a whole has been working on the problem for a few decades, garnering insight, concerns and experience along the way.

---

[14] The word "normalisation" hides a myriad of sins. We're employing the term here to describe a rationalisation of a file format's feature set and removal of idiosyncrasies, as opposed to the preservation methodology known as "normalisation". The first definition would include the example of a correction of a Word document to a more valid (or maintainable, at least) Word document of the same era, versus the proactive migration of a Word document to, say, ODF or PDF-A.

[15] Moran J and Gattuso J (2015) Beyond the Binary : Pre-Ingest Preservation of Metadata. Proceedings of the 12th International Conference on Digital Preservation, 137–143, Available from: https://phaidra.univie.ac.at/view/o:429524 http://ndha-wiki.natlib.govt.nz/assets/NDHA/Publications/2015-16/Beyond-the-Binary.pdf.

[16] The Atlas of Digital Damages (n.d.) Available from: https://www.flickr.com/groups/2121762@N23/ (accessed 22 June 2017)

[17] Cramer T (2014.) PASIG Digital Preservation Boot camp Digital Preservation in Theory and Practice PASIG Digital Preservation Boot camp. Available from: http://web.stanford.edu/group/dlss/pasig/PASIG_September2014/20140916_Presentations/20140916_02_Introductions_Tom_Cramer.pdf.

[18] PrePARe project literature review (2011) Available from: http://www.lib.cam.ac.uk/preservation/prepare/litreview.html (accessed 22 June 2017).

In our experience these threads are commonly not led by the organisation as coherent strategic direction. They are typically predicated on digital preservation expertise talking to each other, or to their parent organisations. We find that these digital preservation based conversations struggle to add context and colour to our conversations with curators and non-specialists who make the final decisions on what we can do to digital content (i.e. offer final organisational sign-off on preservation actions).

The paper proposes an organisational framing of these digital preservation discussions that makes sense at the practical, organisational level. Such a framing allows us, as digital preservation practitioners embedded within our own organisations, to successfully splice our techniques into the other organisational yarns in order to construct the strong rope supporting organisational outcomes.[19]

The underpinning archival strategies that inform the operation of a collecting entity can be loosely folded into one of two different approaches; **artefactual** or **informational**.

The **artefactual** approach considers the archived item to be the primary unit of interest, and endeavours to protect the item and access to it. This model is strongly aligned to emulation or emulation-like preservation approaches. That is to say that the focus is on maintaining the original file and to couple that item with a viable rendering method that allows the file to be consumed in a modern computing environment and in a way that is faithful to the original experience.

A useful reference for this approach is the NAA's Performance model[20]. In this model, the traditional experience of the physical/paper record ← researcher is mapped onto a new model that accounts for the binary object, technology, rendering equipment, and finally the reader (Data File → File Software and Hardware → Rendering on Screen ← Researcher), with the essence of the record itself being maintained somewhere between the delivery of the item to its rendering mechanism and the reader themselves[21].

The **informational** approach considers "information" found inside the archived item to be the primary unit of interest, and endeavours to protect the "information" and access to it. This model is strongly aligned to migration or migration like preservation approaches. That is to say that the focus is on maintaining meaningful access to the informational domains, and treating the file object to be a temporary (but important) mechanism for propelling the informational aspects of the file into the future. Information exists in a perennial abstraction, temporarily bound to a carrier technology that allows it to be consumed.

---

[19] Web C, Pearson D and Koerben P (2013) 'Oh, you wanted us to preserve that?!' Statements of Preservation Intent for the National Library of Australia's Digital Collections. D-Lib Magazine, 19(1/2), Available from: http://www.dlib.org/dlib/january13/webb/01webb.html (accessed 22 June 2017).

[20] Heslop H and Wilson A (2002) An Approach to the Preservation of Digital Records. Archives, (December).

[21] It is worth noting that the Performance Model, while privileging an artefactual approach to preservation (insofar as it focuses on look-and-feel rather than the mobility of content), does not enforce an emulation-based strategy. Rather, the model aims to establish a mandate for the normalisation of formats, especially closed formats or ones where the original intended rendering application may have copyright limitations.

Whilst examples of this approach being used are relatively uncommon in digital preservation, it can be seen commonly in modern computing as any file format change, for example, creating JFIF .jpeg from TIFF .tif files, OOXML .docx from MS WORD OLE2 .doc files, OOXML .docx to PDF. The software is left to its own devices, and the consumer of the migrated file generally (and mostly rightfully) assumes the migration from format A to format B was undertaken accurately, moving the important parts of the original item (e.g. words, layout, fonts, images etc) and leaving less useful parts out of the migration (e.g. the specific structural / technical make-up of the original file). Our specific interest and experience in this approach focuses on what happens if you don't implicitly trust a software migration tool, or you can't find one that works with your files[22].

Both approaches have an inherent interest in the information inside the file. We do not keep content simply for the sake of it. Either approach supports the ongoing access to the information bound inside the file, one through the file itself, the other by extracting the information from the file.

In an artefactual approach, the informational aspects are considered authentic if the file is consumed in/by suitable processes true to the creation epoch of the file. Any changes to the bits that comprise the original item to provide a convenient copy for contemporary consumers are considered to be of limited value, representing only a current technology best-effort attempt to consume the item in question. The relationship between the file object and its consuming environment are of paramount importance, and it is only through the delivery of the file object to the authenticated compute environment that the informational aspects of the file object can be appropriately/authoritatively/reliably consumed.

In digital preservation circles, this argument has rumbled since digital preservation as a requirement was established in the 1990's[23,24,25,26,27,28]. It can be found in digital preservation conferences and journals right up to the present day as discussions around the relative merits of migration versus emulation for undertaking preservation actions.

We are arguing here for a lifting the conversation to a more deliberate organisational vantage point. There can be little argument that either of the two approaches offers insight and technical risk mitigation. For us, a useful contribution to the discussion is to accept that both approaches are viable. The value in the discussion, when we work on our internal preservation actions, comes from increased understanding as to why we have these apparently competing paradigms, and ultimately in the organisation having a fundamental understanding of the possible approaches. We aim for the organisation to have a controlling stake in the technology being used, rather than technology driving methodological choices.

---

[22] Gattuso J and McKinney P (2014) Converting WordStar to HTML4. iPres.

[23] Rothenberg J (1995) Ensuring the Longevity of Digital Documents. SCIENTIFIC AMERICAN, January 1995, Volume 272, Issue 5.

[24] Rothenberg J (1999) Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation. A Report to the Council on Library and Information Resources (1999), Available from: http://linkinghub.elsevier.com/retrieve/pii/S1464905500002141.

[25] Hedstrom M and Lampe C (2001) Emulation vs. Migration: Do Users Care? ResearchGate, (January 2001).

[26] Library N (2003) GUIDELINES FOR THE PRESERVATION Prepared by the National Library of Australia. Organization, (March), 1–9, Available from: http://unesdoc.unesco.org/images/0013/001300/130071e.pdf.

[27] Rothenberg J (2012) Digital Preservation in Perspective How far have we come and what's next. Available from: https://www.youtube.com/watch?v=2Idbur1qR8I (accessed 22 June 2017).

[28] Johnston L (2014) Considering Emulation for Digital Preservation | The Signal. Available from: https://blogs.loc.gov/thesignal/2014/02/considering-emulation-for-digital-preservation/ (accessed 22 June 2017).

In an attempt to move the conversation on, we offer the artefactual vs informational vantage as a possible perspective through which to carry on the conversation within our organisations. We do not offer a preference of one approach over another. The value in the model is a lifting of the discussion from one in which technical positioning predicates the workspace to one where the collecting organisation has the controlling and well-sighted say in what happens to collections under its care. This paper seeks to explain and exemplify this position.

Our organisational imperatives form part of the technical positioning, as well as broad appreciation that perhaps we need to find a way of working within both of the contested paradigms in concert, thereby distilling the value and benefits of both modes of work to offer enhanced, insightful and trusted access to digital content into the future. This argument will be explored further in this paper through the establishment of a new "network of representations" model (see fig.3).

What then is our informational "thing" that we collect and preserve?

This is the million dollar question that remains at the heart of any approach to digital preservation. There have been a number of attempts to describe or explain the quantum unit that matters to memory institutions. At times the definitions are broad and sweeping (e.g. National Library of New Zealand's collections policy[29]) and at other times it is a specific and sophisticatedly defined collection of expressions of data (e.g. Archives New Zealand's general disposal authorities[30]).

It would be extremely difficult to generate a comprehensive single descriptor for all the definitions of information that can be found inside memory institutions. Different types of organisations have different remits and different organisations of the same type have different operational philosophies. We need to find another way of bounding or describing this complicated problem.

Perhaps it is useful to consider how the digital preservation paradigm fits within the organisational context that drives its collection and preservation efforts. By that we mean to understand what the unit of preservable "thing" is. Is it a file? A record? A data-point? The answer to this question is directly related to the *how* of digital preservation.

If we have an artefactual framing, we're probably interested in maintaining the original bits in context, with some notion of an "authentic" expression of those bits in an electronic performance or rendering that offers the consumer an "as was" experience. This might also be described as museological or archival frame.

If we have an informational framing, we're probably interested in maintaining the abstraction of those bits into a form that can easily deliver the information to the consumer in as pliant an experience as we can muster.

---

[29] National Library of New Zealand (2015) Collections policy, Strategy and policy, National Library of New Zealand. Available from: https://natlib.govt.nz/about-us/strategy-and-policy/collections-policy (accessed 22 June 2017).

[30] National Archives New Zealand (2016) General Disposal Authority. Available from: http://records.archives.govt.nz/resources-and-guides/general-disposal-authorities (accessed 22 June 2017).

In the former case, we want to bind the user to the original experience to infer authenticity. In the latter, we want to free the user from the tyranny of technological constraints without comprising the integrity or accuracy of the information we are supplying. The common starting point is a file or file-like object that houses the informational unit of interest.

The digital preservation layer in the organisation needs to somehow understand this origin if it is going to successfully impel any digital source in any of its desirable guises through time.

Pausing for a moment to reflect on this conjecture, we invite the question "does this really matter?"

Are we really bound to these two world views? And are they mutually exclusive? Looking backwards we can see much of the foundational thinking for digital preservation allows for a world of blended outcomes. The PREMIS data model supports multiple representations[31] inviting us to generate multiple modes of engaging with content under our care. OAIS offers language and a framework that supports the satiation of the expectations of varied and differing consuming parties[32].

What then happens when we have user communities that straddle both these worlds? Or content that sits neatly inside both artefactual **and** informational paradigms?

Our current mechanisms struggle to respond to the anticipated needs of both artefactual and information centric consumption. We are not able to elegantly construct multiple representations of an intellectual entity that service different user requirements (and as such different designated communities) and allow the user to decide what version to consume.

We are increasingly conscious that a polarised view of migration vs emulation as a default methodology of choice is a red herring, and have started to explore what a long term lifecycle looks like for objects that fall into either (and both) models.

Allowing ourselves a little speculation, we want to explore what happens after a few preservation interventions. Fig 1 and Fig 2 allow us to imagine what a clean and well-bounded series of preservation actions that result in new versions of preservation masters being created or exposed, might look like in an organisation that utilises one methodology. We have completed at most three generational hops of content.

But what about a less sterile and predicable future state where our capabilities are influenced directly by innovation and research that is not fixed to one single preservation methodology?

By speculating on what it might mean to preserve digital materials through different technologies and via different methodologies we can start to imagine what tools and frameworks we might need to ready ourselves for these potential future states.

Figure 3 is offered as an attempt at modelling some anticipated future states as a series of interventions that have been undertaken on a single digital "thing" throughout its life within a preservation system.

---

[31] PREMIS Data Dictionary for Preservation Metadata, Version 3.0. (2015) Available from: https://www.loc.gov/standards/premis/v3/ (accessed 22 June 2017).

[32] McDonough J (2008) OAIS, Designated Communities & Metadata. Available from: http://www.digitalpreservation.gov/meetings/documents/ndiipp08/session11_mcdonough.ppt (accessed 22 June 2017)

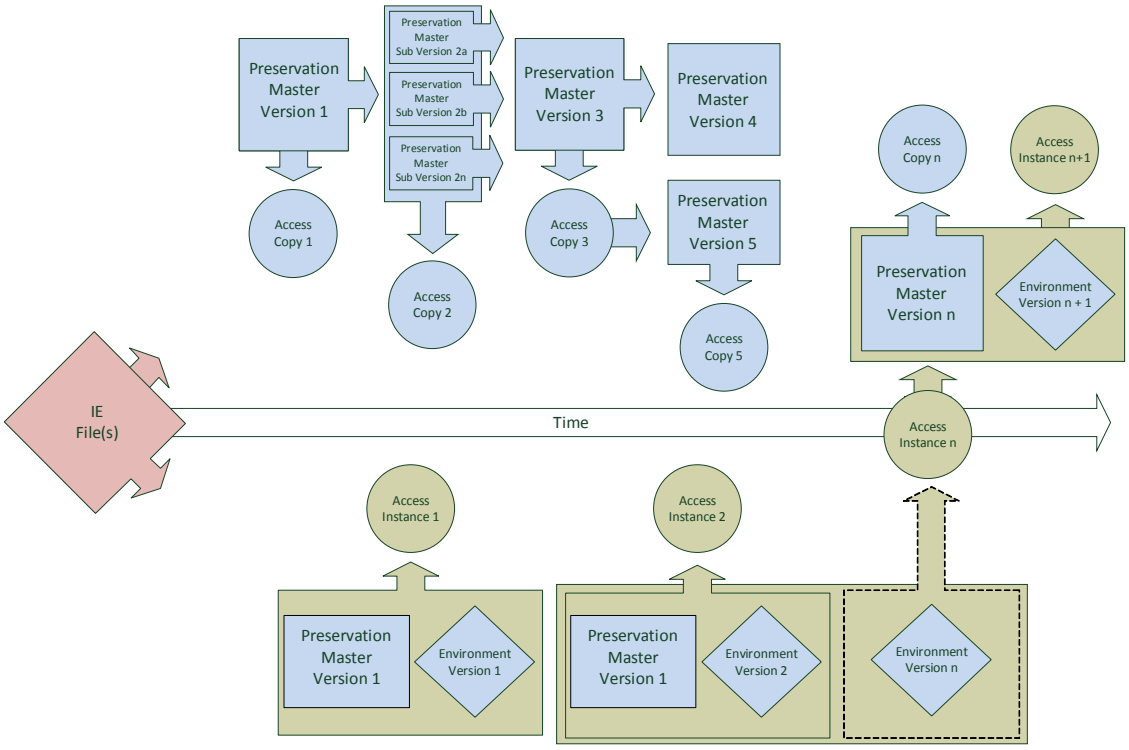

Figure 3 - Network of Representations

This speculated process diagram hides a number of useful constructs that we anticipate leveraging into the future. The diagram is by no means a complete representation of what we expect to do with files, more to indicate the types of activities that we might find useful, including (but not limited to):

Promotion of access copies to preservation master status (see AC3 to PM5)

Figure 4 – AC to PM

- In this use case an organisational decision is made to consider items previously created as access copies as full preservation masters.
- This decision might be driven by technical, fiscal or intellectual reasons.
- Possible usage:
  - An organisation has a large number of high quality TIFF files

- During ingest high quality JFIF JPEG images are created for each item to serve as simple access copies [AC3]
- A risk is flagged against providing access to the JPEG version
- An organisational decision is made to migrate the JPEG files to a newer format, without returning to the TIFF masters [PM5]

Distilling a given preservation master into various informationally incomplete preservation masters that offer a partial but accurate/authentic view into the preceding preservation master (see PM1 to PM2a, 2b, 2n)

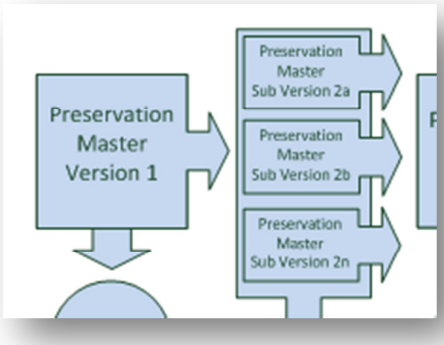

Figure 5 – PM to PMs

- In this use case an organisational decision is made to extract specific informational facets from an original object. The extraction is known to be lossy when considering the whole, but complete when considering the informational facet or aspect that is of interest.
- This decision might be driven by technical, fiscal or intellectual reasons.
- Possible usage
  - A collection of individual tweets is curated as a research collection
  - The Tweets are collected via the Twitter API as s JSON file [PM1]
  - The tweets are individually extracted from the collection as individual JSON files [PM2a]
  - A subset of the individual tweet data are flattened into a CSV file [PM2b]
  - The tweets are coerced into a fully representative HTML file [PM2c]

The coercion of a new preservation master from an access instance provided by an emulated environment (see AIn to PMn)

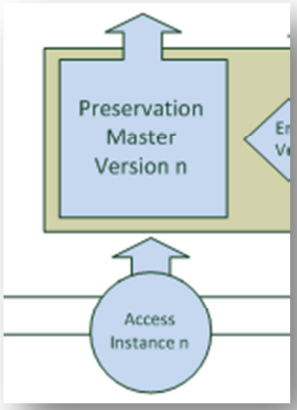

Figure 6 – AI to PM

- In this use case the initial preservation methodology is emulation. As technology and capability advances, a decision is made to extract a fixed state object from the emulated instance, and consider that a newly migrated preservation master.
- This decision might be driven by technical, fiscal or intellectual reasons.
- Possible usage
  - An emulation service is used to provide access to Word for Mac files [AIn]
  - On the same service platform, the Word for Mac files are saved as a standard MS Word .doc files as supported by the contemporary emulation service
  - These doc files are extracted from the emulation service instance and become master files of their own right [PMn]

These theorised units of work augment the previously described states of generating informationally complete migrated preservation masters, access copies, or access instances via emulation.

Linking technical operations to the intellectual endeavours of the organisation

One of the instructional threads that will help us to position ourselves both intellectually and technologically in the digital preservation space is some inspection of the route content takes on its way into the organisation. We do not collect content "just because". We collect content that fits with our collection policies, which are in turn informed by our legal mandates, professional experience, and social obligations.

By describing the desired information essences of any given file as partial representations of the file (by both defining the purpose of the item in the collection, and then by operating on at risk files with that purpose in mind), we are giving ourselves a greater chance of getting digital preservation right.

If we leave the intellectual decision making to the technical operations layer we are in danger of throwing the baby out with the bath water. If preservation actions are not pinned to specific identifiable informational domains we risk decoupling the content from the curatorial decision making.

Our early considerations of digital preservation understood this challenge and we have been working slowly towards a deeper understanding of how the organisational imperative is positioned at the heart of preservation activities. We have made inroads into our understanding of what we think preservation actions look like, specifically, for the moment, focused on migration tasks.

We have also undertaken exploratory work on a number of different text-based formats that have been identified by our internal processes as "at risk" in some way. From these experiments we have observed some useful pointers.

Wordstar/Wordstar2000 to HTML/PDF[33]

In this area we have migrated Wordstar v3.? to HTML v4.01, and we have migrated Wordstar2000 to PDF. These processes were very similar. The amount of code and knowledge we leveraged from the first set (Wordstar) when addressing the second set of files (Wordstar2000) was significant.

We found that critical information required to successfully render the files was actually never in the binary files themselves. This was not an error, but a part of the format specification. The font information was primarily maintained by the application at that time. Specifically we needed to make an educated guess for the font used, basing our assumptions on printed items contemporary to the digital objects.

When exploring different modes of migrating text-based content we observed that there is an intellectual or perceptual difference in how soft line breaks (software added line breaks) and hard line breaks (user added line breaks) are regarded by colleagues. In some cases it is of paramount importance to maintain line perfect migration, in other cases soft line breaks have a lower weighted value.

Understanding what operating rules might be needed appears to include knowing some details of the intent of the document creator, the purpose of the original document, and the perception of weighted value of the typography in the document.

When considering the various transformation steps that have been discussed we might consider this to be a classic example of the linear progression outlined in Figure 1 – Approximate Migration Workflow. The organisational vantage is decidedly informational, depreciating the WordStar object, pulling out the informational essence and placing it in a suitably constructed contemporary format.

---

[33] Gattuso J and McKinney P (2014)

MCW to PDF

In this test we migrated a set of word-for-mac (.mcw) files into PDF version 1.6. This process was relatively straight forward from a technical perspective, leveraging the de jure Microsoft Word products, and its current .mcw codec, bootstrapped by some light Python code to wrangle the test corpus.

We discovered that the migration process in this mode was extremely simple to deliver. We caught misshaped files that failed the migration process with ease, and in the main undertook the relatively successful migration of the 6,000 file set with minimal effort.

Some interesting challenges that will require more thought became apparent including; the way current versions of MS Word interprets bullet points in .mcw files, how we as a community have used umlauts to infer macronised characters (important when rendering te reo Māori language texts) in the past due to technical constraints at the time (and accordingly how we ought to address that problem when we encounter it twenty years later), and how PDF renderers themselves can adjust and scale the bounds of a page of content on a physical printed page.

In addition, we observed simple macro ‘codes’ inside some documents that supports the rendered presentation of the current date in some documents. The future state of this type of information is a conundrum that we as a digital preservation and information management communities alike will need to carefully address as we explore the realities of any preservation actions.

When considering the various transformation steps that have been discussed we might consider this to be a classic example of the linear progression outlined in Figure 1 – Approximate Migration Workflow. The organisational vantage is decidedly informational, depreciating the MCW object, pulling out the informational essence and placing it in a suitably constructed pdf file.

Data-tables

In our collections we have some older database table files in various shapes. These files were collected some years ago, and the relationship between the actual database files and any applications that consumed these records is long forgotten.

We have undertaken a naïve migration from the database record file into a flat record structure, using either a fixed field size constraint or observed field delimiters to coerce the database table files into a csv or similar delimited file type.

We have noted the caution with which we need to communicate these types of technical operations to non-technical colleagues, and have identified a role loosely described as a technical custodian to sit between the file objects and their intellectual expression as understood by curatorial functions within the library. The purpose of this role is to understand the technical minutiae of the file objects and translate that technical capability into a form that is easily understandable by non-technical roles, helping to shape the long term planning for the information bound inside these files.

We observed the delicate and at times overlapping space occupied by a derived digital entity we might describe as an access copy, and one that we might define as a new (derived) preservation master. In our experience we set ourselves a significantly higher bar for any work that results in a new preservation master. There is something instructional in the space that we are yet to conclude but we do know that we need to spend some time understanding how to operate swiftly, at scale, and to document our process steps in a way that ensures consumers of any derived file have a clear understanding of the provenance of the specific digital item they are consuming.

We are also increasingly aware that an access copy that might be created to offer simplified access to a piece of content, which might not have the associated rigours of a preservation led migration of that content, *becomes* the surrogate of the original object regardless of the provenance of the item in question. This has some obvious implications around quality assurance processes that inform and permit the creation of access versions of content that may not be fulsome expressions of the original content.

When considering the various transformation steps offered by the speculative network of representations in Figure 3, we might consider this to be a partial implementation of the step described as figure 5 – PM to PMs. In this case we are asserting an informational vantage, and describing just one (the csv) of any number of informational facets we are able to establish from within the original object. Over time, and with alternate set of requirements we may establish new partial representations that offer insight into the original object. Unknown formats

We have a very small number of digital files in our collection that have defeated our usual technical processes used to ascertain the file format of the items in question. This in turn causes us a problem when we then want to give the item to a researcher. What is our obligation for giving the researcher insightful access to digital items? Are we "allowed" to simply return the bits to the consumer and invite them to undertake any technical analysis of the bits to be able to infer an authentic rendered experience of those bits, or are we obliged to only provide an "authenticated" rendered experience.

In an attempt to explore this question we have undertaken some lossy migrations of some content in this class that contains coherent UTF-8 character encoded text. In this case we were able to understand enough about the structure of the binary item to extract the text contents and dump it into a simple text file. This process has resulted in the coherent text information being made available to a researcher, removing the need for a new bespoke content viewer (for a format we don't understand) at the cost of sacrificing approximately 20% of the binary contents of the file, which (we suppose) contains various formatting and structural components.

The question is: is this known lossy extraction simply a low quality access version of the original file? A high accuracy version of part of the original file? Or is it something else that we need to establish the bounds of.

We are comfortable that this process has resulted in the creation of a highly consumable piece of digital information that can supplant the existing digital object that we would struggle to provide meaningful access to.

This experiment pushes us towards a future state that accepts a nonlinear, non-one-to-one progression of preservation actions for digital items in our care, and hints towards a future where facets of information can be extracted and presented as a less than complete but equally valuable instance of the information bound inside a digital file.

When considering the various transformation steps offered by the speculative network of representations in Figure 3, we might consider this to be a partial implementation of the step described as figure 5 – PM to PMs. In this case we are extracting the only informational essence we can find, and encapsulating it in a contemporary format. Over time we might revisit the original files and create more fulsome or complete or otherwise different instance of the original bits.

Executable wrapped content

In our collection we have two executable files that have been donated for ingest. The decision to collect was made over ten years ago in both cases.

As a file type, executable files (.exe) are markedly different from the types of files we are more used to handling. The very nature of these files means they are a delicate blending of data and application. Our basic consideration is that application is software, and data is content. We are comfortable with the abstraction of these two types of binary objects, and much, if not all, of our operational policy and capability pivots on the notion that we do not collect software. The boundary between software and data is intellectually terse but operationally fluid, and in many ways these files exemplify this confused space between objects that generate the behaviour or process *in* a render experience, and objects that deliver the behaviour or process *of* a render experience.

Our current position is that we are not a software preservation shop, and accordingly we need to understand if this position requires some adjustment, or if the information found inside these exe files is extractable or otherwise viable to preservation within our existing mechanisms.

We observed that for both these executables we are able to operate on the original binary items, and extract what we believe to be the salient information bound within in a form that is more akin to the other ten million non-executable files in our collection.

This challenge is an extension of the previous examples, and is in turn starting to inform how we might conceptualise future states of informational objects in our care.

One of the foundational questions these specific files invite is at the heart of the artefactual / informational proposition. Are we preserving the replay mode, the look and feel of the executables, or should we focus on the extracting the core information so it can be rendered in a way homogeneous with the rest of digital collections or that sits comfortably on contemporary platforms?

In both these cases we believe we can extract the intellectual essences of the executable. In one case we can unpack the exe and attendant files, resulting in a small number of data tables, which can in turn flattened into a simple CSV file. The behavioural components of the .exe file offers the user a narrow set of data searching tools that are trivial to replicate in modern .csv rendering applications. By cleaving the data from its parent application layer we are able to circumvent basic DRM structures that have been established to lock the data to the application, removing at least one obfuscating layer of complexity, and further presenting the core information in a format that is trivial to provide meaningful access to.

In the other we can decompile the executable, resulting in a relatively clean presentation of the assets that comprise the performance of the .exe file when it is deployed. This file is a packaged shockwave flash animation, which appears to be the ordered delivery of image frames with a synced audio track. It is trivial to create a replicate video file that offers a very similar render experience, without the technical overheads of needing to deliver an environment that is capable of "playing" an executable file, and as it turns out, offering easier replay controls for the resulting audio visual experience.

When considering the various transformation steps offered by the speculative network of representations in Figure 3, we might consider this to be a partial implementation of the step described as figure 5 – PM to PMs. In both these cases the organisational vantage is informational, seeking to find ways of depreciating the original exe files, and extracting as complete as possible new instances of the informational essences found in the original object.

We can see that in all of these examples a decision has been made to treat the objects we need to work on informationally. During deliberations with curators, librarians, archivists and preservationists we encouraged ourselves to operate at this new organisational layer, inviting those invested in the collection to consider only the artefactual, informational shaped ingestible objects rather than the specific shape or form of the ingest. The process has proved to be a positive step in introducing the new language and concepts. The informational/artefactual paradigm led to more insightful and decisive treatment planning, reducing the complexity required at the digital preservation level, whilst retaining full organisational control and comprehension of the collection. Essentially, decision making was made easier for all involved once the technological aspects were separated out from the underlying institutional goals.

Summary

We started this paper by describing the early work in the NDHA's history and highlighted the key questions that drive preservation work: "what will we know about formats, and their associated risks?", "how will we know that we have content with an associated technical risk?", "how will we undertake risk mitigation activities on file like objects?", "how will we process items and workflow tasks", "who is the intended audience of this work?".

In the early days, there was very little experience. We now have a decade's worth of experience with which to re-evaluate our answers to those questions. The central pivot for us rests in a translation from technology-driven terminology to a language that is bound by organisational imperative.

We have described an ongoing conversation in the digital preservation community, the somewhat binary at times nature of which, we think, obscures our organisational mission. We have argued that emulation vs migration debates are of little value, and we have attempted to change the debate to one of organisational choices using the constructs of artefactual and informational approaches to digital preservation. As a consequence we can allow our organisational mission to inform our technology choices/strategies (not the other way round).

Through this new lens we have argued that new possibilities become apparent. By looking at the organisational life of digital content and allowing any viable methodology to play important roles in the life of that content we can imagine a raft of insightful processes and treatments such as those we have outlined in the network of representations.

We have discussed some of the immediate connotations that the network of representations offers, and indicated ways of undertaking digital preservation actions that are not currently obvious.

Finally, we have explored the premise via four different collections, and sought to explain where we believe viable preservation actions to lie, in the context of this new conversation.

In conclusion, we would like to return to an earlier statement. Quoting ourselves:

"A large part of digital preservation thinking revolves around the future consumption of digital content in a meaningful and accurate way"

How does a national library negotiate the vast digital shift in landscape, remaining relevant and authoritative in the way we have become accustomed through our traditional work? We should encourage our organisational thinking to steel around empowering concepts, concepts that entwine digital preservation thinking into the language and requirements of the organisation and open up new or well positioned possibilities through this shared language. This will allow us, in our own organisations to drive our practice with insightful direction setting. Understanding the nuance between an artefactual object and an informational one seems to be a solid starting place.

Through the act of writing this paper we forced ourselves to inspect our thinking both with, and without the organisational imperative offered by the informational/artefactual construct. In our limited experience in asserting this new vista we feel that we are able to move from the analysis paralysis borne from having ill bounded requirements into something much more deliberate. Without a driving organisational imperative, the digital preservation layer has to attempt to solve all of the problems all of time. In our experience, progress is slow and complex as the various expectations of the organisation are sought and addressed. By limiting this plurality of outcomes through meaningful statements of intent (i.e. we expect to treat this collection informationally) we have observed that we can significantly reduce the time and effort needed to make genuine progress in our mission to understand what it actually means to "do" digital preservation.

As surely as there are opportunities bound to the digital future, there are also negative outcomes, and we feel strongly that national libraries need to be invested in this digital future, consuming, curating, collecting and conserving digital content with the same dedication and expertise that we have established over hundreds of years in the physical domain.

Acknowledgements

Thanks to Peter McKinney, Sean Mosely, and the rest of the National Library of New Zealand National Digital Heritage Archive team for their input.