A next generation technical registry: moving practice forward

Peter McKinney
National Library of New Zealand
Cnr Molesworth & Aitken St
Wellington
New Zealand
peter.mckinney@dia.govt.nz

Steve Knight
National Library of New Zealand
Cnr Molesworth & Aitken St
Wellington
New Zealand
steve.knight@dia.govt.nz

Jay Gattuso
National Library of New Zealand
Cnr Molesworth & Aitken St
Wellington
New Zealand
jay.gattuso@dia.govt.nz

David Pearson
National Library of Australia
Parkes Place,
Canberra ACT 2600
Australia
dapearso@nla.gov.au

Libor Coufal
National Library of Australia
Parkes Place,
Canberra ACT 2600
Australia
lcoufal@nla.gov.au

Kevin DeVorsey
National Archives and Records
Administration
#1 Bowling Green
Room 450
New York, NY 10004
United States of America
Kevin.DeVorsey@nara.gov

David Anderson
University of Portsmouth:
Future Proof Computing Group
Eldon Building
Winston Churchill Avenue
Portsmouth PO1 2DJ
david.anderson@port.ac.uk

Janet Delve
University of Portsmouth:
Future Proof Computing Group
Eldon Building
Winston Churchill Avenue
Portsmouth PO1 2DJ
janet.delve@port.ac.uk

Ross Spencer
Archives New Zealand
10 Mulgrave Street, Thorndon
Wellington 6011
New Zealand
ross.spencer@dia.govt.nz

Jan Hutař
Archives New Zealand
10 Mulgrave Street, Thorndon
Wellington 6011
New Zealand
jan.hutar@dia.govt.nz

ABSTRACT

In this paper we introduce the work of the National and State Libraries Australasia Digital Preservation Technical Registry project.

Digital preservation practitioners must be able to assume technical and intellectual control of content they are charged with preserving. Our experiences tell us that the information and services used to underpin this control are insufficient. Enterprise-class digital preservation services require something better. We believe the solution outlined here is well placed to deliver information required to preserve digital content. Ultimately, this means that the practitioner can say with a strong degree of certainty that they do indeed have control of the content they are charged with preserving.

General Terms

Infrastructure, communities, strategic environment, preservation strategies and workflows, specialist content types, digital preservation marketplace.

Keywords

Technical registry, formats, hardware, carrier media, operating systems, community, NSLA.

1. INTRODUCTION

The digital preservation practitioner, working within the constraints of their institution's mandate has to be able to assume physical and intellectual control of digital objects and maintain that control for the long-term. Physical control requires them to be able to store the file and protect it from harm and further, understand any risks that may relate to its encoding. The nature of that storage and protection is dependent on the mandate and preferences stated at the national, professional, institutional and personal level.

Practitioners are not immediately (if at all) concerned with the actual content of the file or the context of its creation: who the author is, the purpose the record was created, or story told in the book, or the historical importance of the audio. They are fundamentally concerned though with intellectual control through a technical understanding of the file. Principle questions to be answered as they undertake their work include:

- Can I retrieve this file from the medium it is on?
- What format is this in?
- Can I render this file?
- What are the key details of this format that might impact rendering?
- How long will I be able to render it for?
- Should I consider a undertaking a preservation action?
- What can I use to undertake preservation actions on this content?
- What are other practitioners' experiences?

Our experiences tell us that answering these questions with the current tools and services available, while not impossible, requires that results be gathered from many unconnected sources, which can be questionable in terms of their veracity. In general, these results are pitched at a level that is acceptable only for a high-level technical understanding of a file or format.

Missing from this current landscape of tools and information resources is a holistic view of all strands of technical information required to preserve digital content. In addition, where information is available it is often sporadic and incomplete.

Enterprise-class digital preservation services require something better.

In July 2012, the Chief Executives of the National and State Libraries of Australasia (NSLA) approved funding to investigate developing a Digital Preservation Technical Registry (DPTR). This work is undertaken under the auspices of the Digital Preservation Working Group of NSLA.[1] In order to ensure the project captured the best available thinking in the Registry space the NSLA led project team was assembled with a mix of NSLA and international expertise. The project team comprised: the National Library of New Zealand Te Puna Mātauranga o Aotearoa (NLNZ), National Library of Australia (NLA), the National Archives and Records Administration (NARA) in the United States, the University of Portsmouth (UoP) and Archives New Zealand Te Rua Mahara o te Kāwanatanga (ANZ).[2]

The aim is to develop and sustain a Technical Registry (the Registry) that will be a repository of core technical and relationship information for the file formats, computer applications, hardware and media that have been used to encode (and can be used to decode for human consumption) the digital objects that make up digital collections around the world. This comprehensive, consolidated information resource will be able to be used in conjunction with any digital preservation repository in order to support institutions in their efforts to preserve the digital objects in their care.

2. Problem space

In an effort to extend the traditional concepts of physical and intellectual control to digital collections, digital preservation programmes strive to understand how the digital objects in their collection are encoded. They should know what file format each object is encoded in, as well as the format's technical characteristics, dependencies and requirements. Formats evolve through time and as a result often change dramatically, while their names and external identifiers (for example a PRONOM PUID) often remain unchanged across versions. Additionally, application developers often misinterpret specifications or intentionally vary from their instructions, resulting in digital objects that may require special attention. A registry must endure as a resource of reliable, accurate and comprehensive information capable of describing the variations that are known. This information may be stored locally by individual institutions but, due to the complexity and scope of this domain, we are convinced that it will be more efficient to store this data in a collaboratively designed, developed and maintained registry. It will include descriptions of technical environments and the perceived risks to each whether individually or in combination. That is; file formats, software applications, media, hardware, operating systems and input/output devices.

Over the last few decades there has been activity in the form of collaborative discussion (via wikis, other on-line fora, formal conferences, hackathons, and other workshops) and research to identify information, define and validate models, tools, methods, and other mechanisms that are needed for long-term preservation of digital content. To date, much of this work fits the profile associated with "hobbyist" and "artisan" epochs [5]. There is an increasingly urgent need to move to an "industrial" model capable of supporting enterprise-class digital preservation programmes.

We do not believe that previous or current efforts fully meet the needs of a robust, scalable, enterprise-class digital preservation programme. Consequently, there is a lack of a global, consolidated, open, flexible, authoritative, and trustworthy registry of technical information. There are various impacts on the digital preservation community including the time and effort required to find, interpret and match the necessary information from dispersed sources and the potential to undertake work based on insufficient, erroneous or out-dated information.

This project is intended to extend previous work (whether local or global) including PRONOM[3], the Unified Digital Format Registry (UDFR)[4], Mediapedia[5], TOTEM[6], the Planets Core Registry[7], Just Solve It[8], and the current expressions of technical information used in the Rosetta[9] and Safety Deposit Box[10] systems, which are

[1] http://www.nsla.org.au/projects/digital-preservation

[2] http://natlib.govt.nz/, http://www.nla.gov.au/, http://www.archives.gov/, http://www.port.ac.uk/, http://www.archives.govt.nz.

[3] http://www.nationalarchives.gov.uk/PRONOM/Default.aspx.

[4] http://udfr.cdlib.org/.

[5] https://www.nla.gov.au/mediapedia.

[6] http://keep-totem.co.uk/.

[7] http://www.openplanetsfoundation.org/planets-core-registry.

[8] http://fileformats.archiveteam.org/wiki/Main_Page.

[9] http://www.exlibrisgroup.com/category/RosettaOverview.

based on the PRONOM model. Work began in November 2012 to create a vision and logical data model for the proposed Registry in line with the following assumptions.

1. A technical registry supporting preservation risk management, planning and action is central to an ongoing active digital preservation programme.
2. It is undesirable that there should be a multitude of incomplete technical registries globally.
3. A successful registry will have a clearly defined and understandable data model that will enhance user understanding of the data it holds and allow them to make informed decisions.
4. A successful technical registry should be able to provide data to digital preservation repository systems (e.g. Rosetta, SDB, FEDORA, DuraSpace, Archivematica, RODA etc.).
5. A successful technical registry should be more effective than individual products or services that would be required to maintain an active digital preservation programme, e.g., NLNZ Metadata Extractor, JHOVE, DROID and FITS.

2.1 Current Situation

2.1.1 International strategic imperatives

The international digital preservation community is now at a stage of maturity that is a step beyond the advocacy and awareness raising that was a feature of activities at the beginning of the century. National bodies exist, organisations have experience in operating some level of preservation systems as business-as-usual and first-generation tools and services have been developed. This maturity has allowed the community to begin to assess the status quo and lay down some priorities and strategic markers for movement to the next stage of digital preservation activity.

The National Digital Stewardship Alliance (NDSA) in the United States brings together over 160 organisations who wish to advance the practices of preserving digital resources. The NDSA has recently launched an Agenda to highlight gaps and areas requiring development in digital preservation within the United States. The *National Agenda for Digital Stewardship* [9] contains a number of priorities that the Registry would help support. These include "File Format Action Plan Development", "Integration of Digital Forensics Tools" and "Preservation at Scale". The Registry will provide information and services that will directly support these three priorities.

In Britain, the Digital Preservation Coalition (DPC) works from its *DPC Strategic Plan 2012-2015* [10]. As primarily an advocacy body, the DPC does not directly undertake preservation work, but it has objectives to facilitate "knowledge exchange" and "partnership and sustainability" [10, p1]. The Registry, as a community resource and hub will support the DPC members requirements around digital preservation and the DPC itself could play an important role in the sustainable model of the Registry.

The DPC also commissioned the *Mind the Gap* report. This states that "All organisations need to encourage an international 'market' for digital preservation tools by linking up with other projects around the world and engaging with software vendors. This would deliver economies of scale and reduce risk for individual institutions" [11, p7]. In addition, "[o]rganisations should consider the long-term preservation characteristics of the formats they use." [11, p7] The Registry should be the key resource for both of these activities. The registry will ultimately be home to tools used by the digital preservation community; the centrality of the Registry benefitting their ongoing development and fitness for purpose. It will also be the central resource for risk analysis information about formats and actions to mitigate those risks.

UNESCO convened a meeting of experts in 2011 and developed a declaration on digitisation and preservation [12]. This declaration argues that "digital preservation should be a development priority, and investments in infrastructure are essential to ensure trustworthiness of preserved digital records as well as their long-term accessibility and usability" [12, p2]. It also calls on the UNESCO Secretariat to: "establish a multi-stakeholder forum for the discussion of standardization in digitization and digital preservation practices, including the establishment of digital format registries"[12, p2].

It is clear that there is strong alignment of this proposal for a Digital Preservation Technical Registry to NSLA, National and International priorities and strategic directions. Through:

- supporting the preservation and access of content for the benefit of all citizens;
- the supply of trusted information for digital preservation programmes that will engender trust in their activities and the content they preserve;
- supporting a community that will promote collaboration, develop best practices and peer review Registry information.

Two of the strongest imperatives running through the strategies, policies and agendas mentioned are those of trust and collaboration. The Registry supports both of these goals. Through the supply of comprehensive high-quality, peer-reviewed information, organisations can demonstrate that the actions taken are based on best practice thus reinforce or otherwise improve the trust placed in its custodianship of digital materials. At the heart of the Registry will be a community of practitioners and organisations committed to the long-term preservation of digital content. This community will co-create new information, review existing information and help develop tools to take advantage of the information in the Registry. This community will also share their experiences and allow the collaborative creation of best practice. We also hope that the development of the Registry will be a collaborative exercise with various partners including digital preservation organisations and private sector vendors.

2.1.2 Current technical information

As has been stated above, the five member organisations of the project team posit that the current state of technical information for digital preservation is insufficient.

The concerns can be split into two groups. The first set of cover issues with separate information sources. From the format world alone:

- sources vary in terms of the breadth of information they contain (PRONOM holds records on over 1,000 formats, but the Library of Congress around 350);
- sources vary in terms of the depth of information they contain (TRiD contains a very small amount of

[10] http://www.digital-preservation.com/.

information for every format record, but PRONOM has the capability to record a large amount of information);

- there is little (accessible) historical view of technical information. Is Format A still Format A as I understood it five years ago? [4].

The second set cover issues with the entire information space.

- Information sources rarely reference each other.
- Information sources do not agree on how to describe the world (what *is* a format?)
- There is no central community resource that links technical information with community discussion.

These are not strawmen created for the purposes of supporting this project. These concerns impact the partners' directly as they undertake their business-as-usual practices to preserve the records and/or documentary heritage of Australia, the United States and New Zealand. They have also been borne out by the results of a community dialogue exercise. We have presented our work, including our view of the problem space to a number of organisations either undertaking digital preservation research or actively pursuing a digital preservation programme.[11] Every organisation agreed that the current information landscape is not fit for purpose and limits preservation capabilities. Not one organisation said that the status quo was acceptable.

3. The Proposed Solution

The Digital Preservation Technical Registry (the registry henceforth) will do five key things:

1. bring together technical information sources into a central resource;
2. generate new content and relationships that cover a large percentage (i.e. 80-90%) of content existing in collections;
3. allow users to create new content;
4. allow users to build relationships across all information contained in the Registry;
5. allow the community to comment, discuss and share findings on or related to information contained in the Registry.

In order to make these capabilities, the underpinning data model had to take into account existing information sources and offer a change in direction for some aspects of technical information.

3.1 Model

Each of the project team's institutions had existing data models and/or requirements that formed the basis of the logical data model developed. The model is based therefore on TOTEM for hardware and software[12], Mediapedia for carrier mediums[13] and the internal work of NLA, NARA, ANZ and NLNZ [2, 3, 4] in the format area.

The logical data model developed contains five key entities (as shown in Figure 1).

- Hardware
  Information about the mother board, RAM, CPU and Storage. It also includes devices which support the functioning of a computer like data ports, a computer mouse and removable storage devices.
- IO Device
  Information about auxiliary devices such as a keyboard or hard drive that connects to and works with the computer in some way. Other examples of IO Devices are expansion cards, graphic cards, microphones.
- Software
  Information about applications, operating systems and libraries that can be used to create, edit, render, migrate or emulate files.
- Carrier Medium
  Information about the type of medium upon which data may reside.
- Format
  A "particular arrangement of data or characters in a record, instruction, word, etc., in a form that can be processed or stored by a computer" (Oxford University Press, 1989).

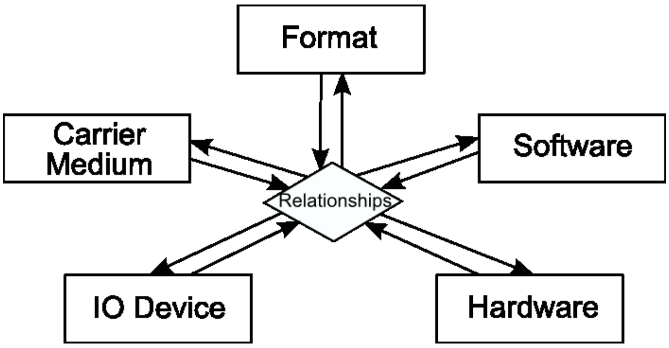

Figure 1: High-level Conceptual Model

While the carrier, software, IO and Hardware aspects of the model are based on existing data models, the format model has been totally re-imagined. It uses three classes of format: Specification, Implementation and Composition. These model the ways in which digital preservation practitioners interact with formats and content

[11] Participating organisations included National Libraries, large collecting institutions and organisations with funding and national strategy mandates.

[12] See http://www.keep-totem.co.uk/.

[13] See http://www.nla.gov.au/mediapedia.

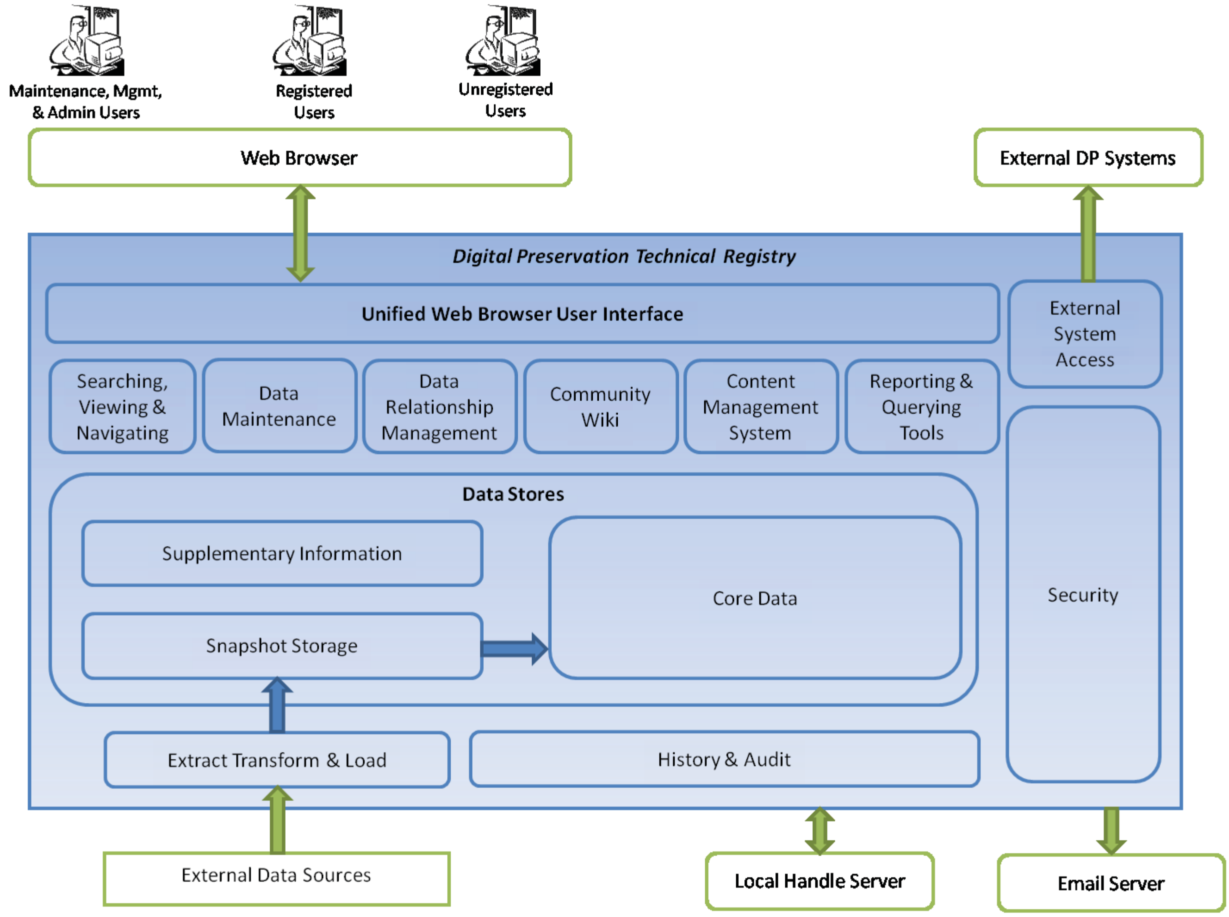

Figure 2: Functional Composition of the Registry

that is represented in those formats.[14] A critical component of the new format model is the concept of an "Aspect". These are the properties that comprise the format types, they are the discrete features and characteristics that are used to build varieties of formats.

The heart of the Registry is the relationships between the entities. It allows all the separate types of information to come alive and become meaningful.

3.2 Functional view

Figure 2 takes a functional composition view of the Registry.

The Registry will give the digital preservation community the following capabilities.

- Ability to import information from current and potential future source registries.
- Ability to store past versions of the external source registry records.
- Ability to support internal registries and online maintenance of the internal registries.
- Ability to flexibly link records within and across external source and internal registries.
- Ability to define the valid link types that can exist between records.
- A web-based user interface.
- Ability to configure what a user, role, or institution can view by allowing information to be filtered based on these attributes.
- Support for creating and running reports across external source and internal registries.
- An API available for external system data export.
- An architecture that supports a decommissioned external source registry becoming an internal supported registry.

[14] The format work is described in more detail in a forthcoming paper.

4. What does this mean?

For the digital preservation practitioner, it means that a whole cosmos of information is available to them and that it resides in one place. It will offer them a breadth and depth of information that is currently unavailable.

Clearly, as can be inferred from the above, the Registry will contain large volumes of information. One way of visualising the information in the Registry and how users will be able to comprehend all the information can be to use the analogy of the night sky. Every piece of hardware, software and media information, every aspect of every format are stars, planets, moons, comets and asteroids.

A wide variety of people 'interact' with the night sky. The more experienced the night-sky-watcher, the more detailed their knowledge and more depth they engage with. Large objects are easily identifiable to anyone: a child can see and identify the moon and milky way. As experience of the sky watcher grows, constellations (relationships enforced upon the sky by man) can be identified and used as tools.

At the far end of the scale of experience, the professional astronomer uses high-powered telescopes based on earth or in space to grapple with the universe. These experts use different modes of retrieving information (x-ray, ultraviolet and broad-spectrum views) to understand space from different angles and analyse things that cannot be 'seen'.

The experience and requirements of the digital preservation practitioner will impact on the level they interact with the information in the Registry. They can stay at the highest level of description and identification ("this is a TIFF") or can delve through the layers of information and begin to grapple with this cosmos of technical information. They can break down that TIFF file into a version, reflect on the properties (aspects) that comprise it, understand how they impact rendering or preservation activities and converse with other experts on those properties.

Likewise they can understand that they have just a 3M-Scotch magnetic tape. Or they can go deeper and understand that it was created under product code 139, rather than product code 140.[15]

The deeper the interaction with the information, the more meaningful the information. Once the practitioner has knowledge of the exact type of magnetic tape they have, they can understand the impacts of having content stored on that exact variety. Once they know the exact type of TIFF they have (and the exact properties) they can ensure that they are making rendering or preservation decisions based on the best information available. This depth also makes community interactions more meaningful. The question "why won't this PDF validate in JHOVE" suddenly becomes "why won't this PDF with encryption and key-length of 128 (Registry ID=xxx) validate in JHOVE 10.2b (Registry ID=yyy)?"

The power of this depth of information is clear. The Registry allows for persistent identifiers to be assigned to such levels of understanding. Users can therefore identify the content they have and bind their relationships and community conversations to that level. It should be noted, that systems or institutions that use existing resources (such as PRONOM) will still be able to use and reference those sources. The Registry will allow for full referencing of those sources and also have the added benefit of allowing users to have historical views of those sources (something that is currently not possible).

Ultimately, this means that the practitioner can say with a strong degree of certainty that they do indeed have intellectual control of the content they are charged with preserving.

At a higher-level, the Registry has the potential to bring a number of benefits to the digital preservation community.

- Trustworthy, high quality information
- More granular understanding of digital collections
- Supporting collection management
- Increased trust in activities
- Efficiency gains
- Economies of scale
- Shared experiences and knowledge
- DP tools utilise Registry

A technical registry is a fundamental component of digital preservation. By moving the current state of the art forward the entire practice of digital preservation benefits.

5. Next steps

Our current work is focused on generating enough collaborative interest in order to build the Registry. A business case has been developed. This proposes a preferred option of international collaboration supporting the build of the Registry and the transition to business as usual. It is clear that the hardest part of the work is not the modeling or requirements capture, nor indeed the build. Rather, the most challenging part will be the transition to a business-as-usual service. The business case therefore focuses not only how to achieve the build, but the transition from completion of the build to a sustainable business.

If successful, this would be a resource built collaboratively and sustained by the community (including the vendors operating in the market). This will require that the digital preservation community consider the weaknesses of the resources currently available, determine how such services can be improved, and ultimately decide the responsibilities of community member institutions to invest in and support a registry that will be of benefit to all.

6. ACKNOWLEDGMENTS

We acknowledge the funding and support from the National and State Libraries of Australasia (NSLA). We also acknowledge the in-kind contributions from each participating organisation. Bill Ross and Dilip Cheerala from Equinox Consulting provided modeling and requirements-capture expertise.

7. REFERENCES

[1] Delve J, & Anderson, D. 2013. *The Trustworthy Online Technical Environments Metadata Database – TOTEM*. Hamburg: Verlag Dr. Kovač.

[2] Gattuso, J. 2012a. *National Library of New Zealand-DROID, PRONOM Developments at the National Library of*

[15] In this case the base material (polyester versus acetate) is different. [http://mediapedia.nla.gov.au/browserecord.php?-action=browse&-recid=110; & http://mediapedia.nla.gov.au/browserecord.php?-action=browse&-recid=111].

New Zealand. Paper presented at Preservation and Archiving Special Interest Group (PASIG), Dublin. Retrieved from http://lib.stanford.edu/files/pasig-oct2012/04-Gattuso_PASIG_presentation_2012.pdf

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

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

[5] McKinney, P., *et al.* 2012. *From Hobbyist to Industrialist. Challenging the DP Community. Paper presented at iPRES 2012, Retrieved from* http://digitalpreservationchallenges.files.wordpress.com/2012/09/mckinney.pdf

[6] Oxford University Press. 1989. *Oxford English Dictionary, 2nd ed.* 3 December 2013. Retrieved from http://www.oed.com.

[7] UC Curation Centre. 2012. *Unified Digital Format Registry (UDFR) Final Report.* Retrieved from http://udfr.org/project/UDFR-final-report.pdf

[8] Webb, C., Pearson, D., & Koerbin, 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*. January/February 2013, 19:12.

[9] National Digital Stewardship Alliance. 2013. *National Agenda for Digital Stewardship,* http://libraries.ucsd.edu/news/_files/2013/ndsa-natl-agenda-cover-2014.pdf. Accessed 9 January 2014.

[10] Digital Preservation Coalition. 2011. *Our digital memory accessible tomorrow. DPC Strategic Plan 2012-2015,* December 2011. http://www.dpconline.org/component/docman/doc_download/713-dpcstrategicplan2012-15. Accessed 9 January 2014.

[11] Digital Preservation Coalition. 2006., *Mind the gap. Assessing digital preservation in the UK,* 2006. http://www.dpconline.org/index.php?option=com_docman&task=doc_download&gid=340. Accessed 9 January 2014.

[12] UNESCO/UBC Vancouver Declaration. 2012.. *The Memory of the World in the Digital Age: Digitization and Preservation, 2012,* http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/mow/unesco_ubc_vancouver_declaration_en.pdf. Accessed 9 January

[13] Del Pozo, N., Long, A. S. and Pearson, D. 2010. ‘“Land of the lost”: A discussion of what can be preserved through digital preservation’, in *Library Hi Tech* Vol.28, No.2, pp.290-300.

[14] Pearson, D. and Webb, C. 2008. ‘Defining File Format Obsolescence: A Risky Journey’, *The International Journal of Digital Curation* (IJDC), Issue 1, Volume 3 (July 2008), pp.89-106. Retrieved from http://www.ijdc.net/ijdc/article/view/76/78.