Reimagining the Format Model.

Introducing the work of the NSLA Digital Preservation Technical Registry

Peter McKinney, Steve Knight, Jay Gattuso, David Pearson, Libor Coufal,

David Anderson, Janet Delve, Kevin De Vorsey, Ross Spencer, Jan Hutař

Acknowledgements

We acknowledge the funding received from NSLA via the Digital Preservation Working Group. We also acknowledge the in-kind funding from the non-NSLA project team members and their respective organisations. Bill Ross and Dilip Cheerala from Equinox IT Limited provided expert assistance in data modelling and requirements capture. Stephen Abrams, University of California, has commented on various iterations of this paper and we are indebted to him for sharing his knowledge and time so freely.

Abstract

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

Any technical registry model must allow digital preservation analysts to understand the technical form of the content they are tasked with preserving, understand capabilities they have in relation to that content and reflect on the community position in relation to those capabilities. We believe the solution outlined here is well placed to deliver the information required to answer these questions, and in a manner that makes it easy to understand, reference and augment.

The primary focus of this paper is to describe the format model, which is the most radical part of the Digital Preservation Technical Registry. The flexibility the model provides delivers on all of the requirements outlined by the NSLA partners and project team members; this includes the ability to reference many layers constituting a format, including relationships between specifications and implementations of real-world formats. We seek input from members of the community on the model and suggestions for use cases and requirements that we have not envisaged.

.

Introduction

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

The project aims to develop a technical registry that will be a repository of core technical and relationship information for the file formats (which this paper focuses on), computer applications, hardware and media that have been used to encode (and can be used to decode for human consumption) the digital objects that are contained in 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.

During the first phase of the project, the team worked to develop a vision, use cases and a logical data model that support such a registry. Understanding the file formats used to encode and contain the digital objects present in our repositories is a central

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

focus. While the project worked on all aspects of a technical registry (including software, carrier media, hardware), the primary focus of this paper will be the proposed modelling of file formats that is central to current thinking in digital preservation. We outline a departure from more traditional models for one more dynamic and encompassing a wider range of use cases.

Methodology

Towards the end of 2012, the five institutions came together to resolve what we define as a failure in the digital preservation community: the lack of a centralised resource containing trustworthy representation information that can be used in the daily transactions of institutions charged with the preservation of digital material.

The five institutions were self selecting. This self-selection was predicated on research and experience that the project members had already undertaken. Each one recognised the problem space and had begun work on creating a solution for a part of the problem (as noted below in “High-level data model” below).

The main tasks of the project were to:

- Understand the problem space and the current situation.
- Validate the existing work against requirements and use cases.
- Unify that work into one model.
- Test the work within the digital preservation community.

The foundational work for this project has varying degrees of published output. Some of these are based on academic research others on the day-to-day experience of running preservation processes across nationally and internationally significant digital

content[2]. This work was reassessed against a large number of requirements and use cases that the project generated and most importantly, it was assessed by a representative sample of potential users of the proposed Technical Registry. As is discussed below, the basis for the work (that is, our statement of the failure in this space) has been accepted by all the institutions, interest groups, individuals and system developers that we have presented it to.

Through unifying the existing models a validation process was undertaken. Could they joined successfully? Is there redundancy? Did anything better exist? Where were the gaps identified by the requirements? Is this the most efficient model? This validation is also an ongoing process and current work has included the format section of the model undergoing detailed internal and external testing.

Background

Project background

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: "file formats are a crucial layer, indeed a hinge between the bits in storage and their meaningful interpretation" (ERPANET 2004)[3]. This type of information is

[2] This spread in published output is evidenced through such papers as: Delve and Anderson 2013; Anderson, Delve and Pinchbeck 2010; Anderson, Delve, Pinchbeck, Alemu and Ciuffreda 2009; Gattuso *Evaluating* 2012; Hutchins 2012; De Vorsey and McKinney 2010.

[3] For more on this see for example, Harvey 2005, p.139, "In order to access and display digital content it is necessary to decipher the bit-stream, to learn what the information in that bit-stream represents. That is the role that file formats play"; Brown 2013, p. 137, "Identifying the format of a file is the key to establishing a means to read the content, and therefore a fundamental requirement for preservation"; and Garrett and Waters 1996, p. 12, "In the digital environment, as we have seen, ideas are typically

known more commonly classed as being part of ‘representation information’, as defined in the Reference Model for an Open Archival Information System (OAIS) (CCSDS 2012). Representation information is the network of information required to understand the message that bitstreams contain. The relationship between the content and the representation information is “a distinguishing feature of digital information preservation” (CCSDS 2012, p. 2-4).

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 (McKinney, *et al* 2012). There is an increasingly urgent need to move to an “industrial”

---

embedded in particular formats and structures that are dependent on hardware and software technologies subject to rapid change”.

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.[4]

The concerns can be split into two groups. The first set 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 information for every format record, but PRONOM has the capability to record a large amount of information).
- Sources are incomplete and many records are only partially filled (Pearson and Webb 2008, p.99).
- There is little (accessible) historical view of technical information. Is Format A still Format A as I understood it five years ago? (Gattuso *Evaluating,* 2012).

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 straw men 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

[4] Examples of tools would be DROID, JHOVE, FITS. Information sources would include: PRONOM, UDFR, COPTR. These are only examples demonstrating that where ongoing support exists, it is from individuals or individual organisations. This is, to us, proof that the digital preservation community is not yet developing enterprise-class solutions: there is no understanding of, or willingness to accept the costs of creating and supporting these solutions.

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.[5] 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.

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[6], the Unified Digital Format Registry (UDFR)[7], the Planets Core Registry[8], and the current expressions of technical information used in the Rosetta[9] and Safety Deposit Box[10] systems, which are 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.[11]

2. As the digital preservation market matures it is undesirable that there should be a multitude of incomplete technical registries globally.[12]

---

[5] Participating organisations included National Libraries, large collecting institutions and organisations with funding and national strategy mandates.
[6] http://www.nationalarchives.gov.uk/PRONOM/Default.aspx
[7] http://udfr.cdlib.org/
[8] http://www.openplanetsfoundation.org/planets-core-registry
[9] http://www.exlibrisgroup.com/category/RosettaOverview
[10] http://www.digital-preservation.com/
[11] A number of sections in CCSDS 2011 can be supported through the use of a technical registry, for example 4.1.5, 4.2.5, 4.3.2, 4.3.3.

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.

These assumptions are part aspirational statements, part boundary markers. They are statements developed by the project team to frame the project work. They are based on the experiences of the team[13] and reflect our understanding of not only the current state of representation information in the digital preservation community, but the community itself.

Vision

The vision of the technical registry is to provide a comprehensive, consolidated, accurate information resource that can be used in conjunction with any digital preservation repository. This repository of key technical information and relationships will support the digital preservation community in understanding, characterising, validating, risk identification, and preservation of digital objects. It should also stand as a resource for organisations and individuals becoming involved in, or learning about,

[12] This is the current situation which, we contend, does not satisfy the needs of the community.
[13] The National Library of New Zealand, Archives New Zealand, NARA and the National Library of Australia have been using representation information to help preserve their digital content for over a decade. The University of Portsmouth have been investigating this information across a number of projects.

digital preservation.

High-level data 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[14], Mediapedia for carrier mediums,[15] and the internal work of NLA, NARA, Archives NZ and NLNZ 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.

[14] See http://www.keep-totem.co.uk/.
[15] See http://www.nla.gov.au/mediapedia.

- 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).

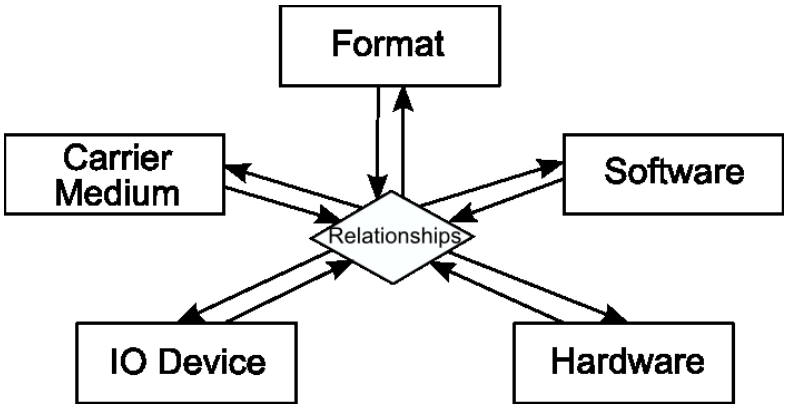


Figure 1. High Level Conceptual Model

These entities show that the key to the Registry is not one entity but rather the relationships between them, generating a fuller picture required for digital preservation activities. The model offers ten bi-directional relationships between the primary entities. These relationships allow practitioners to understand formats, environments of creation and rendering, and characteristics of physical media that objects may be on. It should be of no surprise, that at this level of granularity, the model can be seen to reflect the PRONOM version 4 model which contains hardware components, software, storage media and file formats (see Figure 2).[16] The hardware, software and carrier medium entities reflect models that have already been developed by other preservation centres and are available for the community to use. In the proposed Registry the format model reflects, and builds on, experience of using existing sources (for example, see Gattuso *National Library* 2012; Hutchins 2012). In addition, through addressing the

[16] Brown, Adrian. 2005. "PRONOM 4 Information Model" (January): p. 4.

requirements of the NSLA partners and the project team, it introduces new terminology and allows a new method of expressing format. This new model moves format away from the PRONOM model and other expressions of format.

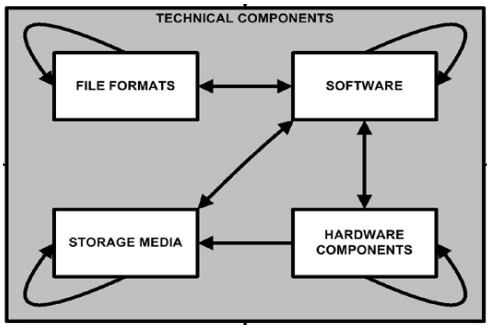


Figure 2. PRONOM v4 Information (cropped)

Format background

Format identification is needed in order to begin to technically understand content that is to be preserved. As UDFR stated in 2012: "[a] deep understanding of digital formats is necessary to support the long-term preservation of digital assets, as it facilitates the preservation of the *information content* of those assets, rather than just their *bit stream representations*"[17]. Information about formats plays a key role in identification, validation, characterisation, risk analysis, and preservation planning and execution.

"Format" is defined by the Oxford English Dictionary as 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".[18] For the digital preservation community, the most meaningful and directly relevant come from UDFR and PRONOM.

[17] Unified Digital Format Registry (UDFR), *Final Report*, California Digital Library, 2012, p. 3. http://udfr.org/project/UDFR-final-report.pdf
[18] "format, n." *OED Online*. Oxford University Press, June 2014. Web. 13 August 2014.

UDFR: "the set of syntactic and semantic rules that govern the mapping between information and the bits that represent that information." (UC Curation Centre, 2012)

PRONOM: "A file format is an encoding of a file type that can be rendered or interpreted in a consistent, expected and meaningful way, through the intervention of a particular piece of software or hardware which has been designed to handle that format. File formats may be defined by a specification, or by a reference software system. Many file formats exist in forms with minor variations, and many also in more than one version. Typing of file formats should be interpreted generously rather than strictly, but sufficiently precisely to distinguish versions where such distinctions have significant interpretive consequences."[19]

While the definitions are different, they do not contradict one another. UDFR makes note of information, whereas PRONOM focuses on the relationship between file types (as opposed to information) and the computing platform. Both give a solid foundation upon which we have based our format work. Rather than redefine 'format', the project accepted the definitions from UDFR and PRONOM and worked towards a model that encompasses both and allows for the requisite variation.

Format registries already exist. These include PRONOM[20], UDFR[21], TRiD[22], and the Library of Congress Digital Formats[23] as well as local instantiations of format libraries that exist.[24] Inter- and intra-registry information varies in both breadth and depth:

- TRiD has over 5,000 file extension records, but very light information for each of those entries.

[19] http://test.linkeddatapronom.nationalarchives.gov.uk/vocabulary/pronom-vocabulary.htm
[20] http://www.nationalarchives.gov.uk/pronom/
[21] http://udfr.org/
[22] http://mark0.net/soft-trid-e.html
[23] http://www.digitalpreservation.gov/formats/index.shtml
[24] For example at NLNZ, NARA and NLA.

- PRONOM has information on around 1,000 formats, but the vast majority of these do not have any detail beyond a name and identifier.

- The Library of Congress in comparison has detailed information on around 350 formats.

The community could benefit greatly from centralising these sources. This would allow users to begin to take advantage of the benefits brought from different sources without having to traverse numerous websites (and different terminologies).

Our experience tells us that current sources of information describing formats are generally based on differing data models.

- Library of Congress is concerned mainly with sustainability factors and uses around 125 attributes to do this.

- PRONOM contains around 79 attributes and can store information on, among other things, related formats, technical properties and identification strings.

There is some overlap in the types of information that registries hold and this leads to the situation where registries may convey similar, exactly the same or indeed, conflicting information about the same format. As there is no global arbiter, registries do not always agree what a format is and how format families can be described; if at all.

Comparing JPEG2000 across the Library of Congress and PRONOM is instructive.

Library of Congress has three categories for holding records on JPEG2000. These are: ‘JPEG 2000 encodings’; ‘JPEG 2000 file formats’; and, ‘JPEG 2000 File format with encoded bitstreams’. There are 21 entries that come under these headings (as shown in Table 1). In comparison, PRONOM holds only three.

Table 1: Library of Congress records for JPEG 2000

Library of Congress ID	LoC Category	LoC name
fdd000138	JPEG 2000 Encodings	J2K_C, JPEG 2000 Part 1, Core Coding System
fdd000139	JPEG 2000 Encodings	J2K_C_LL, JPEG 2000 Part 1, Core Coding, Lossless Compression
fdd000140	JPEG 2000 Encodings	J2K_C_LSY, JPEG 2000 Part 1, Core Coding, Lossy Compression
fdd000194	JPEG 2000 Encodings	J2K_C_Profile_0, JPEG 2000 Part 1, Core Coding, Profile 0
fdd000196	JPEG 2000 Encodings	J2K_C_Profile_1, JPEG 2000 Part 1, Core Coding, Profile 1
fdd000211	JPEG 2000 Encodings	J2K_C_Profile_3, JPEG 2000 Part 1, Core Coding, Profile 3
fdd000213	JPEG 2000 Encodings	J2K_C_Profile_4, JPEG 2000 Part 1, Core Coding, Profile 4
fdd000170	JPEG 2000 Encodings	J2K_C_BIIF_01_00, JPEG 2000 Part 1, Core Coding, BIIF Profile (v. 01.00)
fdd000192	JPEG 2000 Encodings	J2K_C_NDNP, JPEG 2000 Part 1, Core Coding, NDNP Profile
fdd000141	JPEG 2000 Encodings	J2K_EXT, JPEG 2000 Part 2, Coding Extensions
fdd000143	JPEG 2000 File Formats	JP2_FF, JPEG 2000 Part 1 (Core) jp2 File Format
fdd000154	JPEG 2000 File Formats	JPX_FF, JPEG 2000 Part 2 (Extensions) jpf File Format
fdd000144	JPEG 2000 File Formats	JPM_FF, JPEG 2000 Part 6 (Compound) jpm File Format
fdd000167	JPEG 2000 File Formats with Encoded Bitstreams	JP2_J2K_C_LL, JP2 File Format with JPEG 2000 Core Coding, Lossless
fdd000168	JPEG 2000 File Formats with Encoded Bitstreams	JP2_J2K_C_LSY, JP2 File Format with JPEG 2000 Core Coding, Lossy
fdd000195	JPEG 2000 File Formats with Encoded Bitstreams	JP2_J2K_C_Profile_0, JP2 File Format with JPEG 2000 Core Coding, Profile 0
fdd000197	JPEG 2000 File Formats with Encoded Bitstreams	JP2_J2K_C_Profile_1, JP2 File Format with JPEG 2000 Core Coding, Profile 1
fdd000212	JPEG 2000 File Formats with Encoded Bitstreams	JP2_J2K_C_Profile_3, JP2 File Format with JPEG 2000 Core Coding, Profile 3
fdd000214	JPEG 2000 File Formats with Encoded Bitstreams	JP2_J2K_C_Profile_4, JP2 File Format with JPEG 2000 Core Coding, Profile 4
fdd000169	JPEG 2000 File Formats with Encoded Bitstreams	JP2_J2K_C_BIFF_01_00, JP2 File Format with JPEG 2000 Core Coding, BIIF Profile (v.01.00)
fdd000193	JPEG 2000 File Formats with Encoded Bitstreams	JP2_J2K_C_NDNP, JP2 File Format with JPEG 2000 Core Coding, NDNP Profile

Table 2 maps the JPEG2000 records in PRONOM to the related Library of Congress records. The difference exists because of the decisions that the Library of Congress have made around the way that they understand a format and the aim of their registry. It is recognised that the site has a very broad interpretation of format and contains information on file formats, bitstream encodings, wrappers and bundling formats and classes of related formats.[25] PRONOM, in comparison, does not have such a broad range.

Table 2: PRONOM records for JPEG2000 matched to Library of Congress records

PRONOM ID	PRONOM name	LoC ID
x-fmt/392	JP2 (JPEG 2000 part 1)	fdd000143
fmt/463	JPM (JPEG 2000 part 6)	fdd000144
fmt/151	JPX (JPEG 2000 part 2)	fdd000154

Are these differences important? At worst they can cause a decision that affects the integrity of the content being preserved. At best, they inflict upon users and their institutions large inefficiencies. A key consideration for a digital preservation specialist is how they can make sense of these different sources and marshal the information in such a way as to make it useable and auditable. Integrity of digital content is a fragile thing: slight changes can affect it dramatically. If that integrity is broken, then the entire framework upon which heritage institutions rests is put at risk: what is the value of a national archive if it cannot prove that the content it contains is integrious? The following experiences of the project team's institutions give further context to the requirements and solution outlined below.

[25] See 'Formats, Evaluation Factors and Relationships' page. http://www.digitalpreservation.gov/formats/intro/format_eval_rel.shtml.

Format Requirements: General

The research undertaken during Stage 1 of the project included developing a set of requirements and use cases that would be representative of why and how users would want utilise the Technical Registry.[26] In this paper, we present those that are pertinent for the format model.

The project members have very clear high-level requirements for format information. Such information drives many digital preservation activities (risk analysis, migration, emulation and access). The requirements are therefore to:

1. Understand the technical form of the content to be preserved.
2. Understand capabilities in relation to that content; e.g., does the institution have the resources required to render or migrate?
3. Reflect on the community position in relation to those capabilities.

Our requirements take these top level statements and develop them in order to create capabilities for users of the Technical Registry. Users will be able to:

- Understand and reference a format as it is defined in formal statements describing the format;

  This is a key component, for example of the guidance work that NARA and Archives New Zealand undertake; a trustworthy, absolute reference to standards and specifications of format.

- Understand, reference and identify a format as it exists as an Implementation of a Specification;

[26] These are available upon request.

For organisations that manage and preserve content, there is a strong requirement to understand and identify formats that exist, as opposed to those that are specified.

- Understand, reference and identify features not mentioned in a Specification but are contained in an Implementation;

  It is known that within the collections housed in New Zealand, Australia and the United States that content is encoded in formats that do not conform to specifications. The Registry must be able to support instutions in identifying them and understanding what that means for their capabilities.[27]

- Have a resource that contains all sources of format information in one centralised place;

  Currently, they are scattered geographically and intellectually (as has been shown above). There are great efficiency benefits to be gained from such centralisation.

- View levels of format information from detailed features through to high-level format 'families';

  Institutions do not deal with content at the same level. The National Library of New Zealand currently require very detailed technical views of their content. This can be contrasted with the requirement from NLA's creation of a register of

[27] The 'Managing Government Records Directive' states that "NARA will complete, and make available, revised guidance, including metadata requirements, for transferring permanent electronic records, to include additional sustainable formats commonly used to meet agency business needs" (Executive Office of the President 2012). This advice is based in part upon the Library of Congress sustainability factors but leans heavily on the tension between published specifications and files that agencies actually generate: the GIF dissection outlined under, 'Using the Model', is based on a current NARA discussion point.

relationships that must retain a high-level view of format due in part to issues with tools.[28]

- Have a clear insight into, and potential to participate in, the decision-making of the format definition and identification;

  Given the above requirements to be able to identify sometimes unique content, there is a necessity to be able to influence what records are added to the Registry and the shape that those records take.

- View a full history of format identification as it exists across external sources and the Registry;

  It is clear from previous research that format description and identification is a moving target (Gattuso Evaluating *2012). For the sake of audit, it is crucial that there is a historical view across the Registry.*

- Have access to persistent identification for the formats that their content is in;

  As has been argued above, format is the linchpin for understanding the content. Persistent idenfiers for format identification are therefore a key piece in maintaining the integrity of the content that is being preserved.

- Have access to identifiers used by other sources;

  The Registry will store a number of "external" sources of information (e.g. PRONOM, Library of Congress format information). There are a few scenarios

[28] On tools, see Hutchins 2012. Among other tasks, based on earlier thinking about format obsolescence (Pearson and Webb 2008), the digital preservation team has been working on developing a detailed list of relationships between formats and software and building a corpus of test files (this will demonstrated at iPRES 2014 by NLA). This allows the team to develop a capability register that can then support the wider Library through the notion of preservation intent and level of support (Del Pozo *et al 2010;* Pearson 2012; Pearson 2013; Webb *et al* 2013). The experience of generating this capability register is that format is a varied and wonderful thing as described by written material accompanying software. Some applications' accompanying literature describe formats solely by file extensions, some are more defined with versions of formats, certainly though, not one mentions a registry identifier. The NLA team will have to deal with format descriptions at both the very high level and more detailed levels.

that demonstrate the need to be able to continue to use the identifiers (or indeed, entire structure) of that external source.

- Use qualified links between format, software, carrier media and hardware;

  To take advantage of the format information, there must be meaningful and flexible linking not only between formats, but also to software and all other parts of the Registry. This is the ecosystem through which content is actively preserved.[29]

The model proposed below attempts to resolve all of these requirements.

Proposed Solution for Format

The Digital Preservation Technical Registry will encapsulate information on formats, hardware, IO device, software and carrier media. In relation to formats, the DPTR will do six key things:

- Bring together various format information sources (known as 'external sources').
- Store every version of the sources collected.
- Build a central core format registry that will:
  - describe the format world as defined in specifications;
  - describe the world of formats as found amongst digital objects "in the wild";
  - allow for varying levels of understanding a format;

[29] See Delve and Anderson 2013, and Mediapedia, http://mediapedia.nla.gov.au/home.php.

  - offer information for identifying these descriptions;
  - be linked to collected associated sources;
  - be linked to software, hardware and carrier media.

- Allow institutions to build profiles of formats and relationships specific to them.
- Provide a space for community discussion and interaction.
- Enable the annotation of records to add institutional or community specific value.

Terminology

The proposed format model for the Registry introduces four new concepts that extend the current way of talking about formats to allow a greater range of expressiveness at both the macro and the micro level. Three of them are ways to represent how a format should be understood: **Specification**, **Implementation** and **Composition**. The fourth concept is the building block for these three types and is called **Aspect**. These are modelled below in Figure 3.

Format is an abstract entity in the model below. Specifications, Implementations and Compositions are different forms of Format. The following model and rules only look at Format and its forms, as they exist within a record in the Registry.

The entities of Specification, Implementation and Composition record different ways in which 'format', the method of arrangement or what may also be classed as the encoding scheme, is presented or interpreted. A Specification entry sets out the 'official'[30] encoding scheme, as defined for example in a published standard. An Implementation entry is a description of an actual example of the scheme being used, as

[30] This notion is discussed further below.

instantiated in an object in a collection. A Composition entry aggregates Specifications and Implementations. It is an entity that allows various levels of format identification to be utilised. All of the concepts are defined and discussed further below.

Definitions for Entities

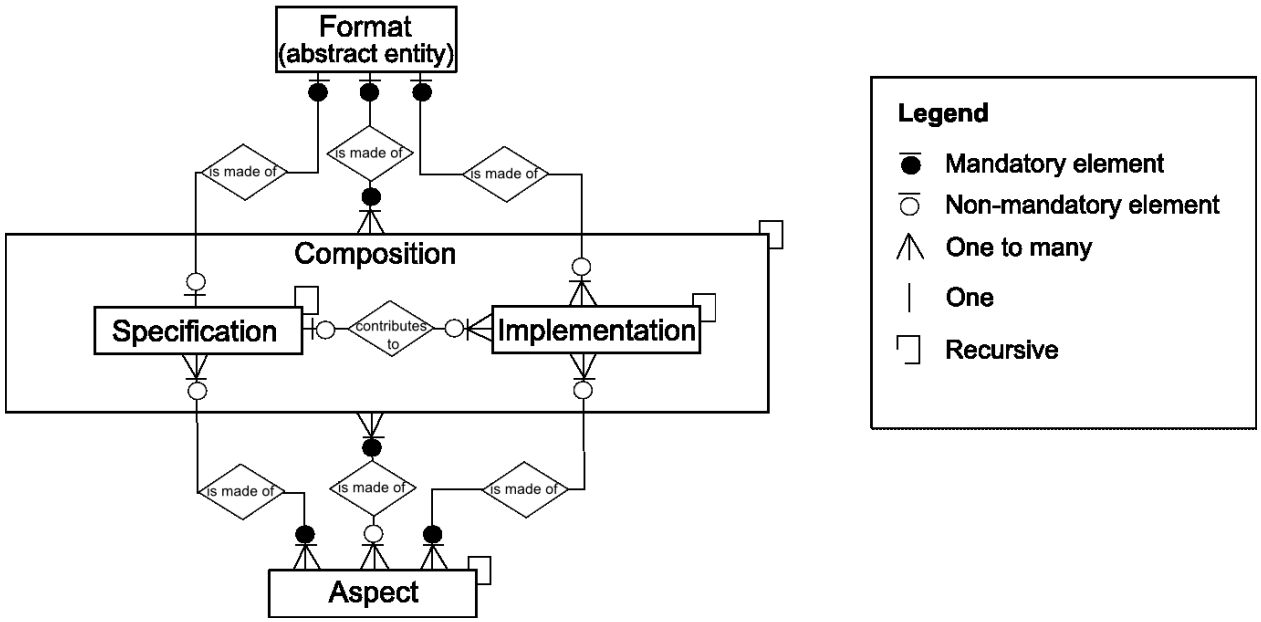
Format entities

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).
Specification	A formal statement of the precise features and characteristics (Aspects) by which a format may be identified, i.e. a formal statement of the precise requirements which a format must satisfy.
Implementation	An actualisation of a specification.
Composition	All of the Aspects associated with a format as manifested by differing Implementations of a Specification.

Building Block

Aspect:	A discrete feature / characteristic of a format.

Figure 3: Model of format types and building blocks



Model Explanation

The following rules draw out some more details about the new concepts. The following rules only look at Format and its forms, as they exist within a record in the Registry.

A Format may (but does not have to) be defined by a Specification. A specification always defines one Format only. A Specification may be composed of many Aspects and must have at least one. An Aspect may belong to many Specifications, however, it does not have to belong to any Specification.

A Format record in the Registry may have many Implementations (but does not have to have any). An Implementation always represents one and only one Format. An Implementation may be composed of many Aspects but must have at least one. An Aspect may belong to many Implementations, however, it does not have to belong to any Implementation.

An Aspect must belong to a Composition. It may also belong to one or many Specifications and Implementations (it could also belong to both). An Aspect may occur in many different Specifications and Implementations pertaining to different Formats.

A Specification may (but does not have to) contribute to many Implementations; An Implementation does not have to be derived from a Specification but when it is, it can be derived from exactly one only. (Note: this does not exclude a Specification referring to another Specification, however this is modelled via the recursiveness of the Specification entity).

Related Implementations along with the Specification they are derived from (where it exists) are encompassed in a Composition. A Composition can contain other Compositions (Note: this is modelled via its recursiveness). A Format must have at least one Composition (i.e. it must have either a Specification or at least one Implementation or both) and may have many of them.

Detailed Descriptions

The following offers further detail on the model.

Specification

In our experience, a format cannot exist without some descriptive documentation (whether formal or not). Specifications can be de jure standards, de facto standards or simple notes generated by the creator of the format. The Specification entity may only reflect information contained in the published standard it is representing. For example, a Specification will be developed in the Registry for Tiff version 6. It will be directly equivalent to the Adobe TIFF Revision 6.0, Final – June 3, 1992 specification[31],

[31] See http://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf.

reflecting the features and characteristics that may exhibited by files conforming to that format.

Implementation

An Implementation is an instantiation of a format. An Implementation describes real world examples of formats that are found amongst collections. An Implementation can either conform exactly or partially to a Specification, or it can contain features or characteristics that are not described in the Specification, thus making it non-conformant.

To exemplify: the National Library of New Zealand permanent repository contains digital objects in GIF format containing the Lempel-Ziv-Welch (LZW) compression and ones that employ no compression. The GIF89a specification does not allow for files to be generated using anything other than LZW compression. In order to identify such files, an Implementation will need to be created that explicitly notes the use of no compression. This Implementation will contain a signature for identifying this variant. NLNZ will then be able to use this signature to identify all GIFs in their collections encoded with no compression.

Composition

A Composition is the macrocosm of the Aspects that are used across all related Specifications and Implementations. In addition, it may contain other Aspects that are not part of a Specification or an Implementation. For example, the published standard for TIFF v6 does not include any notion of mime-type for the format. An Aspect of mime-type cannot therefore be contained within the Specification: Specification records in the Registry will reflect only the Aspects that formally describe the format. As it is relevant for all Implementations, the Aspect will be placed in the Composition.

Aspect

Aspects are the set of features or characteristics of formats. They are the building blocks for Specifications, Implementations and Compositions. These features can range from compression schemas, to colourspaces, to bitrates, to codecs. Aspects may also be used to describe other information that may not be classed as a direct property of the file or format that it is encoded in, but rather describe things about the file. Such information may include mime-type attribution, identifiers (e.g. PUID), descriptions and references. Aspects can be used to describe different levels of properties.

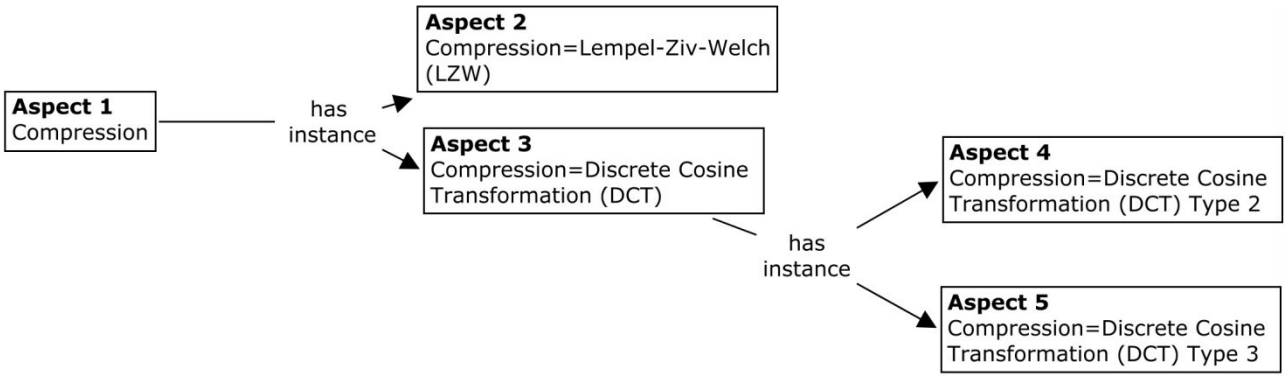


Figure 4. Aspect relationship

Aspects are composed of names and values. Figure 4 shows compression as an Aspect defined in an image Specification (TIFF 6 for example). This has a related aspect that describes a specific type of compresssion (Lempel-Ziv-Welch (LZW), Aspect 2). A further Aspect can be added that shows another sub-Aspect related to Compression. Aspect 3 describes Discrete Cosine Transformation compression. Aspect 3 can in turn be broken down into further Aspects (in this case, two extra Aspects 4 & 5).

Aspects are the atoms that build Specifications, Implementations and Compositions. While not explicitly shown above, Aspects can be specific or generic and may be used

across many format types or only once. This atomisation of information brings many benefits including flexibility and efficiency. Users may also search across and reference high-level or low-level information, depending on their specific needs. While the high-level requirement for Aspects are understood, further refinement is required to model it fully. We understand for example that there will be different types of aspects. This typing has not yet been finalised and can be viewed across many different planes: hierarchically (those that are high-level concepts "compression" and those that are very detailed values "DCT Type 3"); functionally (Implementation-aspects, composition-aspects); or indeed by type of value (single, multiple, range).

These different methods of typing are not incompatibile. Indeed, we can imagine the Registry allowing for multiple typings in order to aid users in their navigation and use of these building blocks. This work is ongoing as the format model continues to be tested and developed.

Using the Model

Use of different format types

Digital preservation practitioners are actors undertaking specific roles within a digital preservation programme including format specialists, analysts, archivists, collectors, and curators. These specialists, either by virtue of the role they play, or by virtue of institutional policy, will be interested in different levels of format identification. Some may only wish to utilize the higher level information commonly referred to currently as 'format' (effectively the Composition layer in the DPTR model), while others may wish to understand their collections at the most detailed level, and therefore deal with Aspects within Implementations. Some organisations will wish to

pay particular attention to Specifications and deviations from them within Implementations.[32]

Figure shows a possible set of relationships between Specifications, Compositions and Implementations in a simple format (in this case, MP3).

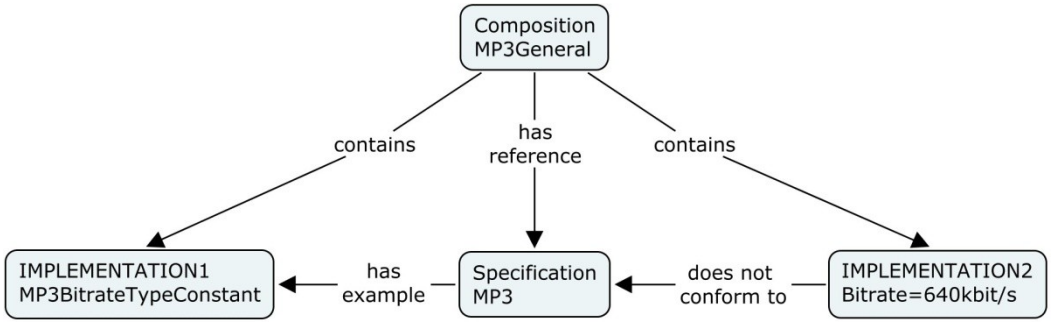


Figure 5. Relationships between a Composition, Specification and Implementations

The registry allows for full functionality of identification, reporting, and linking at all three levels. There is no loss of capability to any user. This example is deliberately simplified. The complexities of formats are many. One format description will often reference multiple specifications that are used to create the format (for example Open Office, video formats, or anything that employs wrapping and brings together various formats). This will be reflected in the Registry records for the Specifications. The case study below begins to draw a view of how this simple three-type approach can cover formats and format families in a more complex fashion.

[32] These are three examples that we know in this area. NLA has relationships built between software and format at the Composition level. NLNZ wish to understand their collection items in small, discrete piles, and therefore require greater detail. Institutions such as the Danish National Archives which mandates that "the creators of the archives must migrate digital records to a few, well-defined standard formats identified by the Danish National Archives for the purpose of long-term preservation" (Statens Arkiver 2013), *may* want to utilise Specifications far more than other organisations.

Gif87a and GIF89a Case Study

The following case study looks at how two specifications could be modelled in the DPTR. Specifically, it looks at GIF87a and GIF89a standards.

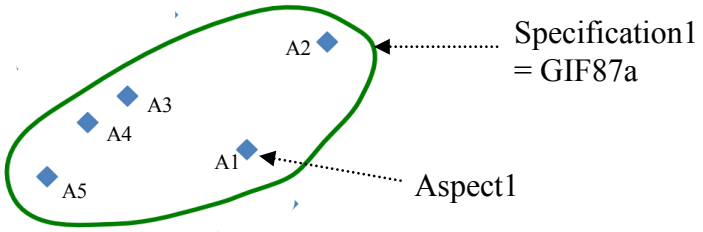
In all the images below:

Blue diamonds = aspects

Green lines = specifications

Blue lines = implementations

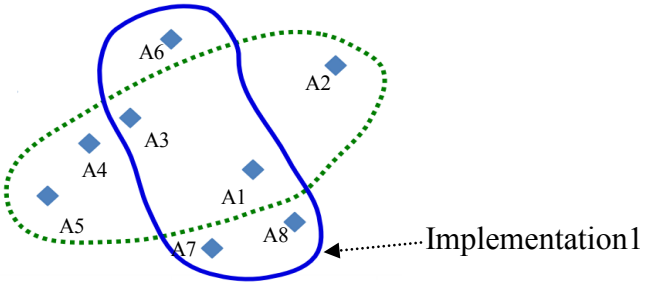
Black lines = compositions



Specification 1
A1
A2
A3
A4
A5

Figure 6. Single specification

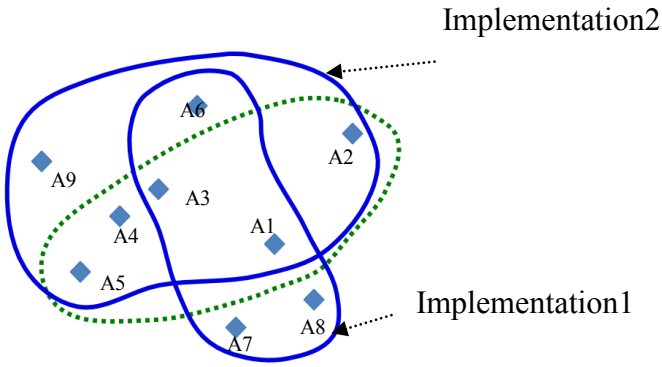
Figure shows five aspects contained within a Specification. These aspects represent the information contained in the GIF 87a specification. The Specification is shown as the green line bounding all the aspects.



Spec 1	Implementation 1
A1	A1
A2	A3
A3	A6
A4	A7
A5	A8

Figure 7. Single specification, single implementation

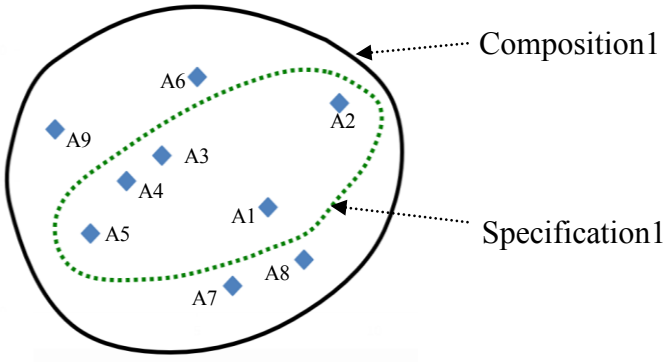
Figure introduces an implementation based on GIF87a. This implementation adds three new aspects that are not contained in the specification (for example, a non-conformant compression type).



Spec 1	Imp 1	Imp2
A1	A1	A1
A2	A3	A2
A3	A6	A3
A4	A7	A4
A5	A8	A5
		A6
		A9

Figure 8. Single specification, two implementations

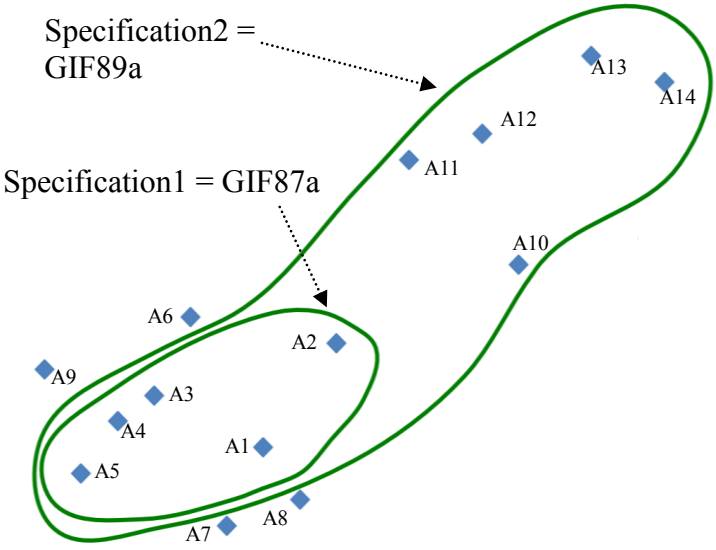
Figure adds a second implementation. It also adds one further aspect. These two Implementations reflect variations from the Specification as found in actual collections.



Spec 1	Composition 1
A1	A1
A2	A3
A3	A6
A4	A7
A5	A8

Figure 9. Single specification and single composition

Figure describes the same set of aspects as utilised in the Implementations and Specification. This time it includes a Composition that encapsulates all of the aspects. This composition is a superset of the aspects related to the GIF87a Specification and the Implementations. The value of the Composition in this case is that it allows a digital preservation specialist to look at all the possible Aspects used across related format types. Note that the Implementations have been removed for the sake of clarity.

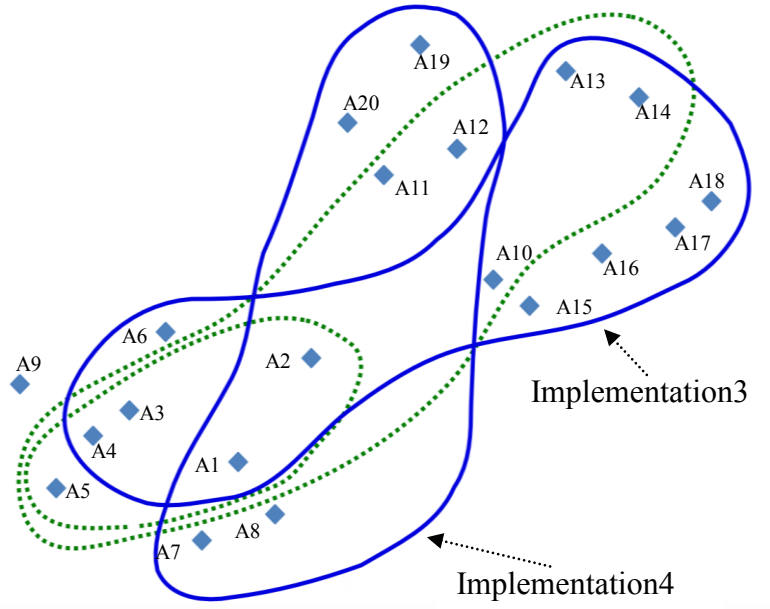


Spec 1	Spec 2
A1	A1
A2	A2
A3	A3
A4	A4
A5	A5
	A10
	A11
	A12
	A13

A14

Figure 3. Two specifications

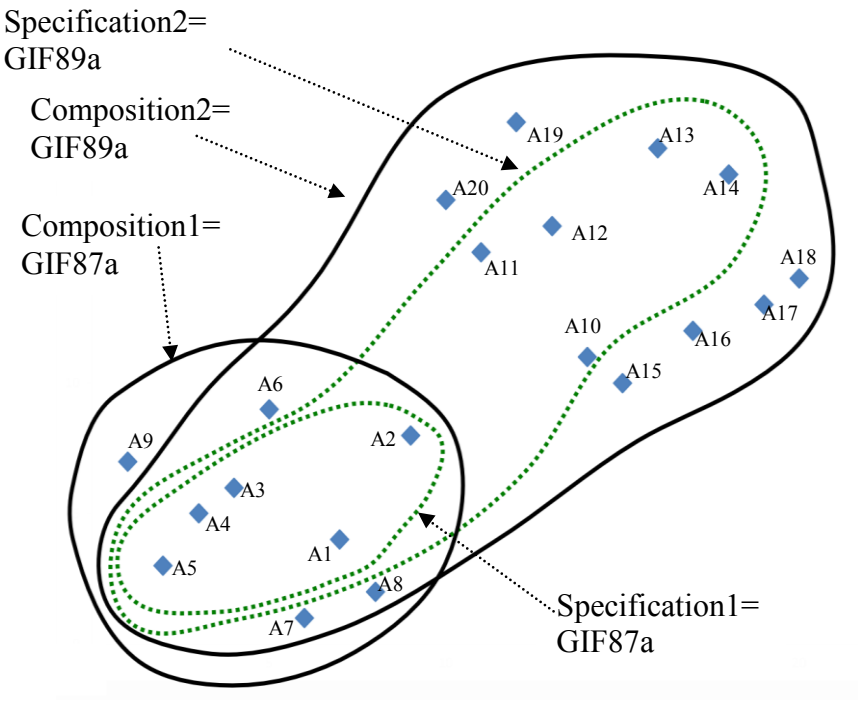
Figure 3 shows the relationship between GIF87a and 89a specifications. The GIF89a specification is an extension of the 87a specification, so it contains all of the 87a Specification Aspects and adds five new Aspects. This new Specification is bounded by a second green line.



Spec 1	Spec 2	Impl 3	Impl 4
A1	A1	A1	A1
A2	A2	A2	A2
A3	A3	A3	A7
A4	A4	A4	A8
A5	A5	A6	A11
	A10	A10	A12
	A11	A13	A19
	A12	A14	A20
	A13	A15	
	A14	A16	
		A17	
		A18	

Figure 4. Two specifications, two implementations

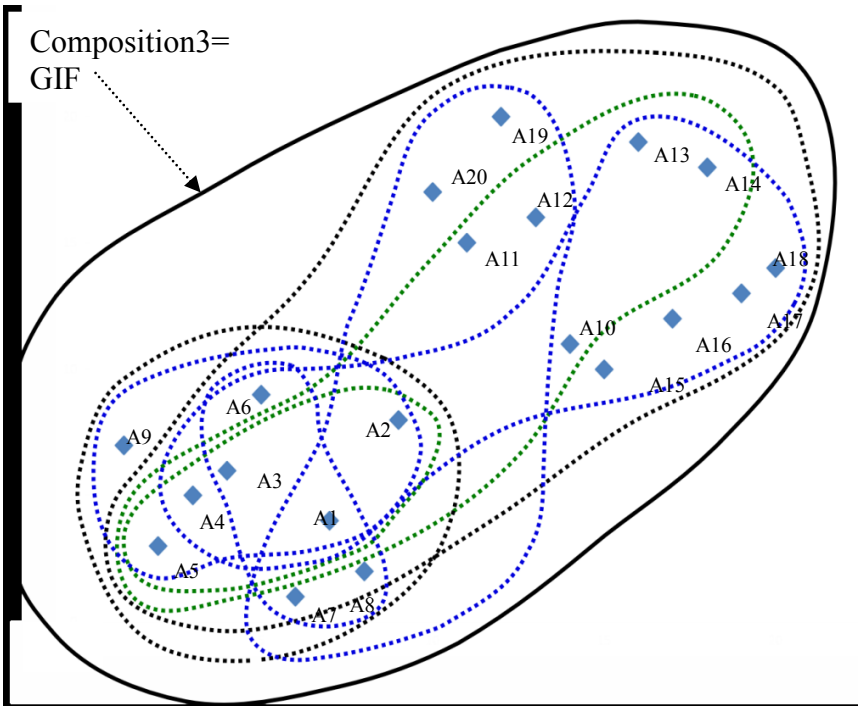
Figure 4 shows two implementations of the GIF89a Specification. As above, these implementations reflect content as found in actual collections.



Spec 1	Spec 2	Comp1	Comp2
A1	A1	A1	A1
A2	A2	A2	A2
A3	A3	A3	A3
A4	A4	A4	A4
A5	A5	A5	A5
	A10	A6	A6
	A11	A7	A10
	A12	A8	A11
	A13	A9	A12
	A14		A13
			A14
			A15
			A16
			A17
			A18
			A19
			A20

Figure 5. Two specifications, two compositions

Figure 5 displays the two specifications and adds two compositions. Specification1 and Composition1 relate to GIF87a. Specification2 and Composition2 relate to GIF89a.



Comp3
A1
A2
A3
A4
A5
A6
A7
A8
A9
A10
A11
A12
A13
A14
A15
A16
A17
A18
A19
A20

Figure 6. Two specifications, three compositions

Finally, Figure 6 introduces a third composition level. This shows how a composition can encapsulate both Specifications. This is used here to denote the “GIF” family, under which the 87a and 89a standards sit.

This example reflects a relatively simple format and how the various elements are used to model it. It highlights how Aspects are used as atomic elements that can be used to create the universe of features and characteristics which comprise the more expressive format world of Specifications, Implementations and Compositions. It also shows how aspects can be shared.

The format world however, is often far more complex than the GIF image. Many formats are amalgamations of more than one specification, such as Open Office XML, EPUB, or any sort of video format. Figure 7 below uses the same form as the use case below to model a more complex format. Implementations are not drawn in this image for the sake of clarity but are assumed through the adding of aspects to the composition that are not contained in the specification.

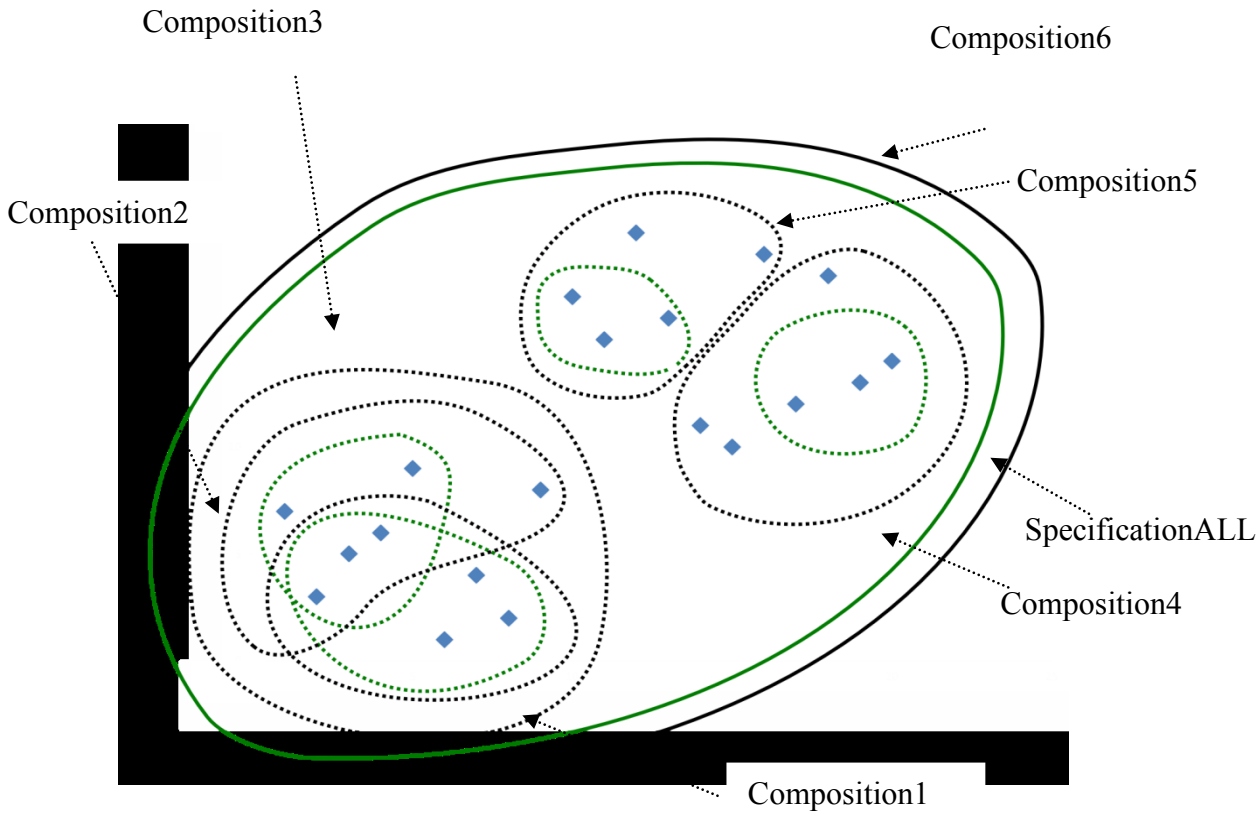


Figure 7. Multiple specifications and multiple compositions

In this example, there are six compositions and five specifications. The two specifications and three compositions on the bottom left corner (Composition1, 2 & 3) essentially mirror the relationships described in the GIF images above. In the top right corner are two more specifications with their own compositions (Composition4 & 5). Finally, a specification brings together all the other specifications (SpecificationALL) and it has its own composition (Composition6). This could perhaps reflect a format such as Open Office XML, which references many other specifications (such as pkzip, xml, and others).

Of course, the Specifications in this example can be used in their own right and will each have their own Implementations. So, if we continue with the theme of Open Office XML, one Specification will be a version of XML. This of course can also therefore be referenced as a format in its own right. The model is designed to allow for this building of complex formats through sharing and reuse of all levels of information.

Conclusion

In this paper, we have introduced the work of the National and State Libraries Australasia Digital Preservation Technical Registry project. The format entities (Specification, Implementation and Composition) offer the ability to understand the world as it is described in format documentation (formal or not) and how it is implemented in the real world. The trick in digital preservation is understanding reality over a theorized reality. Digital content found in a collection is not often structured in a way that matches the published Specification. Real-world examples deviate. The degree of this deviation can vary greatly, and analysts need to be able to understand, describe and reference that. We believe the use of Aspects as the building blocks for Specifications, Implementations and Compostitions are the best method for providing

the levels of granularity reqired to document these layers of variation within the format world and thereby provide the evidence for enterprise class digital preservation risk management, preservation planning and preservation actions.

Any technical registry model must allow digital preservation analysts to understand the technical form of the content they are tasked with preserving, understand capabilities they have in relation to that content and reflect on the community position in relation to those capabilities. We believe the solution outlined here is well placed to deliver the information required to answer these questions, and in a manner that makes it easy to understand, reference and augment.

The primary focus of this paper was to describe the format model, which is the most radical part of the Digital Preservation Technical Registry. The flexibility the model provides delivers on all of the requirements outlined by the NSLA partners and the project team; this includes the ability to reference many layers constituting a format, including relationships between specifications and implementations of real-world formats. We are currently actively testing this section of the model and will be seeking input from members of the community on the model and suggestions for use cases and requirements that we have not envisaged.

Phase 1 of the NSLA project is now complete. The collateral developed includes a data model and data dictionary, vision document, user stories, and system actors and use case descriptions. The next steps of the work include peer review of the collateral (of which this paper is a part) and the development of options for a business model for the Registry. Experience across the spectrum of digital preservation initiatives has taught us that the sustainability of any solution will be the toughest challenge to tackle.

References

Anderson, D., Delve, J., and Pinchbeck, D. “Towards a Workable, Emulation-based Preservation Strategy: Rationale and Technical Metadata”. *New Review of Information Networking*, (2010): 1361-4576.

Anderson, D., Delve, J., Pinchbeck, D., Alemu, G.A., & Ciuffreda, A. (2009). Preservation metadata standards for emulation access platforms. FP7 Report to the European Commission. 85pp, available to download at: http://www.keep-project.eu/ezpub2/index.php?/eng/Download/Public-deliverables (Deliverable D3.1).

Anderson, D., Delve, J., Ciuffreda, A., Pinchbeck, D., Alemu, G.A., Joguin, V., Lohman, B., Kiers, B., Michel, D. (2009). Guideline document and peripheral input/output libraries for digital preservation. FP7 Report to the European Commission. 23pp, available to download at: http://www.keep-project.eu/ezpub2/index.php?/eng/Download/Public-deliverables (Deliverable D5.1).

Brown, A. *PRONOM 4 Information Model*. The National Archives, 2005 <http://www.nationalarchives.gov.uk/aboutapps/fileformat/pdf/pronom_4_info_model.pdf>

Brown, A. *Practical Digital Preservation. A how-to guide for organizations of any size.* London: Facet Publishing, 2013.

The Consultative Committee for Space Data Systems (CCSDS), *Audit and Certification of Trustworthy Digital Repositories. Recommended Practice.* September 2011. The Consultative Committee for Space Data Systems (CCSDS), *Reference Model for an Open Archival Information System (OAIS). Recommended Practice*. June 2012.

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

Del Pozo, N., Long, A. S. and Pearson, D. '"Land of the lost": A discussion of what can be preserved through digital preservation', in Library Hi Tech Vol.28, No.2, (2010): 290-300.

De Vorsey, K. & McKinney, P. 'Digital Preservation in Capable Hands: taking control of risk assessment at the National Library of New Zealand', *Information Standards Quarterly,* Spring *22:2*, 2010: 41-44.

Managing Government Records Directive. Executive Office of the President, 2012 <http://www.whitehouse.gov/sites/default/files/omb/memoranda/2012/m-12-18.pdf>

Garrett, J. & Waters, D., *Preserving Digital Information*. Report on the Task Force on Archiving Digital Information, Commissioned by The Commission on Preservation and Access and The Research Libraries Group, 1996.

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

Gattuso, J. (2012). *Throughput efficiencies and misidentification risks in DROID*. National Library of New Zealand report. <http://ndha-wiki.natlib.govt.nz/ndha/attach/ReadingResources/MSB%2BDROID%20v1_05.pdf>

Gattuso, J. (2012). *Full results for DROID version 6 "Max Byte Scan" test.* National Library of New Zealand report. <http://ndha-wiki.natlib.govt.nz/ndha/attach/ReadingResources/MBSResults.pdf>

Gattuso, J. (2012). *Evaluating the historical persistence of DROID asserted PUIDs.* National Library of New Zealand report. <http://ndha-wiki.natlib.govt.nz/ndha/attach/ReadingResources/Historical%20View%20of%20format%20via%20DROIDv4_2.pdf>

Gattuso, J. (2012). *Main results of the DROID version tests.* National Library of New Zealand report. <http://ndha-wiki.natlib.govt.nz/ndha/attach/ReadingResources/historical%20view%20droid_results.pdf>

Giaretta, D. *Advanced Digital Preservation.* Berlin: Springer-Verlag, 2011.

Harvey, R. *Preserving Digital Materials.* Munich: KG Saur, 2005.

Hutchins, M. (2012). *Project Report: Testing Software Tools of Potential Interest for Digital Preservation Activities at the National Library of Australia.* <http://www.openplanetsfoundation.org/blogs/2012-08-12-file-characterisation-tools-report-testing-project-conducted-national-library>

McKinney, P., *et al. From Hobbyist to Industrialist. Challenging the DP Community.* Ninth International Conference on Digital Preservation, Toronto, 2012 <http://digitalpreservationchallenges.files.wordpress.com/2012/09/mckinney.pdf>

Pearson, D. 'The Adventures of Digi: Ideas, Requirements and Reality', at *Future Perfect 2012,* Museum of New Zealand Te Papa Tongarewa, Wellington (26 March 2012). <http://www.slideshare.net/FuturePerfect_/dave-pearson-the-adventures-of-digi>

Pearson, D. 'Those Mad Men from the Antipodes' at *CurateGear 2013,* University of North Carolina at Chapel Hill, North Carolina (9 January 2013). <http://www.slideshare.net/natlibraryofaustralia/u-nc-2013-v10>

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

Statens Arkiver, (2013). *Strategy for archiving digital records at the Danish National Archives*. <http://www.sa.dk/media(4826,1033)/Strategy_for_archiving_digital_records.pdf>

UC Curation Centre. (2012). *Unified Digital Format Registry (UDFR) Final Report.* <http://udfr.org/project/UDFR-final-report.pdf>

Webb, C. "Digital Preservation – A Many-Layered Thing: Experience at the National Library of Australia", *The State of Digital Preservation: An International Perspective, Documentation Abstracts, Inc. Institutes for Information Science, Washington, D.C. April 24-25, 2002 (*pp. 65—77). <http://www.clir.org/pubs/reports/pub107/reports/pub107/pub107.pdf>

Webb, C., D. Pearson, and P. Koerbin. "'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 19:12 (2013).