Modelling File Formats and Technical Environments using the NSLA Digital Preservation Technical Registry (DPTR)

Jay Gattuso, Peter McKinney, Steve Knight
National Library of New Zealand
Cnr Molesworth & Aitken St
Wellington
New Zealand
Jay.Gattuso@dia.govt.nz

Jan Hutař, Ross Spencer
Archives New Zealand
10 Mulgrave St
Wellington
New Zealand
Jan.Hutar@dia.govt.nz

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

Kevin DeVorsey
National Archives and Records Administration
One Bowling Green, 3rd Floor
New York, NY 10004
USA
Kevin.DeVorsey@nara.gov

ABSTRACT
This workshop introduces the work of the National and State Libraries funded work on a Digital Preservation Technical Registry. In particular it will allow participants to gain an understanding of the new model for modelling formats. They will be tasked with working through exercises designed not only to give participants an understanding of the model, but to test and critique it.

General Terms
infrastructure, communities, strategic environment, specialist content types,

Keywords
Technical Registry, Models, File Formats, Hardware, Software, Community, Collaboration.

1. PROPOSAL
The heart of contemporary digital preservation is multi-faceted. Where once the technical registry sat at its core, we now seek 'war stories' from the community that detail experiences with legacy digital information. We seek high quality information about file formats, carrier mediums, software, and the complete picture of the technical environments that we're dealing with. With archival principle at the centre of our work we also look for detailed provenance, validation and verifiability of that information.

Along with that, the technical registry still holds a key role in the community's multi-faceted approach. The technical registry, through enabling file format identification and validation, can aid the filtering and routing of content at the pre-deposit, and pre-ingest stages of the digital preservation lifecycle.

The technical registry is still the core information source for the migration and maintenance of content as part of any technology-watch capability or preservation action.

We believe that the technical registry should provide information that is accessible to all involved in digital preservation at all levels of skills and knowledge. The information also needs to be actionable, that is, machine readable information that can be accessed by the tools in the digital preservation toolkit. Registries should use relationships to describe more complete technical environments – the links between specific instances of software, hardware, carrier mediums and file formats. Registry information also needs to be augmented with user-level, 'community text' describing war-stories, domain expert knowledge, and the institutional relevance of specific registry entries. Our aim is to create a registry flexible enough to contain and identify real life file format instances with all their specialities and varieties.

Striking that balance, we'll be introducing our work on the NSLA Digital Preservation Technical Registry, providing an overview of its core features before introducing our more radical changes in thinking. The biggest advancement we'll introduce is the overhaul of the format model traditionally used in digital preservation - presenting three interpretations of file format that we believe encompass the many different ways we talk about the subject in the community. Workshop participants will be given a more thorough introduction to this side of the work package, learning about these three components and the building blocks used to create them – format Aspects.

Using Aspects and the knowledge of how to create our three format objects, participants will engage in a modelling activity to help us challenge the work we've completed this far and help us to reinforce a step-change in thinking about the requirements of a modern, comprehensive technical registry.

2. DRAFT AGENDA
0900-1000: Introduction to the NSLA Digital Preservation Technical Registry
1000-1100: The new format paradigm
1100-1115: Coffee and refreshments
1115-1230: Breaking down a format specification
1230-1330: Lunch
1330-1500: Building an Implementation

1500-1515: Working across the format domain
1515-1645: Use case and the wider Registry environment
1645-1700: Wrap up

3. WHO SHOULD ATTEND?

Digital archivists, digital preservation analysts and developers.
Digital preservation service providers and organisations.
Repository managers and organisational digital preservation, information technology leadership.

4. WORKSHOP OUTCOMES

The primary purpose of the workshop was to introduce the format model that the NSLA Digital Preservation Technical Registry team has been developing. The workshop was attended by a wide-spread of people who were eager to hear about the Technical Registry work. The beginning of the session was side-tracked slightly away from the format model specifically, but this, it runs out was necessary to explain the context of the format work. One learning from this was the need for the team to spend more time setting up the reasons for the Registry and the use cases that it fulfills.

The outcomes from the workshop are shaped by the scope of the event. The workshop was a 'transmit' event. That is, it was the first time that the (majority) of participants had been introduced to the format work that had been developing over the past three years. The goal was therefore to successfully describe the format model in order that the participants could not only give initial thoughts but more importantly, spend time considering the work and engage in deeper discussions at a later date.

Essentially, the team believe that the participants felt that the work was going down the right track. It will be hard to achieve and there are still a number of areas unresolved and questions unanswered, but these were points that the team were aware of (and in many cases, there was not time to satisfy all questions, particularly those that were about the broader project rather than the format model).

.