Digital preservation, provenance and Rosetta

(presentation based on a paper by Sean Mosely, Digital Preservation Technical Specialist at the National Library of New Zealand)

Rosetta Advisory Group meeting

New York, June 6-8, 2016

Steve.knight@dia.govt.nz

INTERNAL AFFAIRS
Te Tari Taiwhenua

Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY
OF NEW ZEALAND

Today

A discussion of the effectiveness of Rosetta, as a digital preservation system, in recording and maintaining the provenance of digital objects in the National Library of New Zealand's digital preservation programme.

Are the right events maintained as provenance events?

Are provenance events sufficiently transparent and clearly demarcated from other events in Rosetta?

Is the method of describing provenance events sufficiently granular and structured, particularly in terms of the carrying digital objects into the future?

Is there appetite for a project to re-look at provenance within Rosetta?

What do we mean by provenance (1)

Historically, provenance has been a record keeping term referring to the owners of and contributors to a record over its lifecycle, e.g. this from the National Archives of Australia:

The office of origin of records, (i.e. the agency, office or person that created, received or accumulated and used the records in the conduct of business). In archival theory, the principle of provenance requires that the archives of an agency or person not be mixed with the archives of another.

http://www.naa.gov.au/records-management/publications/glossary.aspx#p

What do we mean by provenance (2)

This presentation uses the definition provided by the PREMIS Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group, v1.0 published in May 2005.

Documentation of processes in a Digital Object’s life cycle. Digital Provenance typically describes Agents responsible for the custody and stewardship of Digital Objects, key Events that occur over the Digital Object’s life cycle, and other information associated with the Digital Object’s creation, management, and preservation.

This is also the definition used by the National Library of New Zealand in its National Digital Heritage Archive project and the development of Rosetta.

What do we mean by provenance (3)

Note that this is a much broader definition than that from the NAA including as it does:

- key events
- metadata related to an object’s creation, management and preservation.

This extension of the scope of the notion of provenance is entirely valid in the digital space as it:

- acknowledges the need to identify the characteristics of digital objects to ensure that future renderings are faithful to the original
- incorporates the need for ongoing technical validation of digital objects, through initial deposit processes and migration to new storage media
- allows us to track and validate digital objects as they are migrated/emulated to different formats and are subject to other preservation processes as they arise over time

The DNX Metadata Model

Digital Normalised XML (DNX) is the Rosetta metadata model for recording and tracking the preservation metadata of digital records.

It is built into the Rosetta database and represented within a METS XML file along with associated Dublin Core descriptive metadata.

The different components of the DNX are displayed in subsections of the XML, specifically within the METS Administrative Metadata section.

The various components of the DNX are then distributed amongst the appropriate subsections of the METS Administrative Data (techMD, rightsMD, sourceMD and digiprovMD).

DNX and PREMIS

At its core the DNX is an implementation of PREMIS (PREservation Metadata: Implementation Strategies).

DNX extends PREMIS with additional sections and semantic units including details on:

- Content Management Systems
- file validation
- virus scanning
- objects' producers and producer agents.

One of the major differences between the two models is in the way that they treat Intellectual Entities (IEs).

However, PREMIS 3 (released in late 2015) has become more aligned with the Rosetta DNX through the ability to treat an Intellectual Entity (IE) in the same way is it previously allowed for representations, files and bitstreams.

How do we go about ensuring that Rosetta continues to be PREMIS conformant?

Provenance in Rosetta

Provenance metadata in Rosetta is split between the descriptive metadata and the administrative metadata sections of the database and METS structures.

This includes:

- identifiers (internal to Rosetta and relating to external systems)
- creation/modifications dates, details about objects' producers and the agents that deposit the content (if they are different)

Provenance events include:

- changes to IE level metadata (descriptive metadata access rights, structure map or any other part of the DNX)
- adding new representations to the IE;
- validation checks (although an event will only be added if the fixity of an object changes);
- enrichment, such as the generation of a persistent identifier

- Provenance events are hard coded in Rosetta although they can be enabled or disabled by admins (**and this is a good thing**).

So, what should be a provenance event?

There are currently 55 live provenance events in Rosetta (out of a total 270 events).

Event 31 – The Technical Analyst requests the Producer to fix problems in file and resubmit – is a provenance event.

However, event 34 – Technical Analyst requests the Producer to fix problems in file(s) and to resubmit the entire deposit - is not a provenance event.

Why is that?

Other non-provenance events in Rosetta

- event 35 - Technical Analyst approved the SIP for further processing
- event 367 - Technical Analyst skipped problematic files in block {1}
- event 368 – Technical Analyst decided to abort the execution of the block
- event 369 - Technical Analyst decided to re-run the conversion for block {1}
- event 372 - Technical Analyst manually determined single format id for File {5}
- event 378 - Technical Analyst manually ignored an error for File {5}
- event 382 - Permanent failed processing work:{1} version:{2}
- event 383 - Permanent checksum failed on file:{1) in:{2} version:{3}
- event 384 – Resubmit deposit activity.

Assuming these events are still valid and actually able to be executed (e.g. the block events – there is some uncertainty here), why is this? Is it time to address the full set of events, particularly the provenance events?

Optimising provenance

In 2012, National Library of New Zealand submitted a proposal for expanding PREMIS to include elements for tracking changes to digital objects before they are ingested in a more systematic fashion.

The proposal suggested a ‘ProvenanceNote’ semantic unit within the PREMIS data model allowing for a structured description of the changes undertaken to a file.

This was enacted in PREMIS 3 as eventDetailExtension

The extension allows for descriptions of changes in a consistent format, both human-readable and machine-readable.

Structured descriptions could provide the base logic for automated re-construction of files in their original state.

Unstructured changes would provide a location for describing steps taken, so that future researchers can be satisfied that the changes undertaken were necessary and sufficient to ensure continued render-ability of the record, in a manner that is faithful to the original.

DNX and PREMIS 3.0 conformance could include increased machine-readability and action-ability of events comprising common manipulations of files, e.g.:

- adding or replacing file extensions
- renaming files to remove illegal or reserved characters
- discrete changes to a part of a file (such as changing a header tag from 0x2A to 0x2B)
- extensive changes to a file (such as replacing a bitstream)

These are all primary events that involve changing some component of a digital item’s content, such as a file name, a component of a file or the file itself which could easily be defined as provenance events.

Pursuing such changes in Rosetta would allow for continued alignment between Rosetta and PREMIS noted in slide 7 above.

Next steps - a project to review …

1. all current provenance events and their ongoing suitability

2. all current non-provenance events and the potential for some to be re-labelled as provenance events

3. documentation related to provenance events in Rosetta

4. other related questions, e.g. whether actions involving declining or rejecting a SIP constitute provenance actions if the object never enters the digital preservation system (i.e. is this a provenance question – where are the boundaries?)

5. DNX and PREMIS 3.0 conformance – enhancing provenance

On the whole …

Rosetta system is capable of tracking and maintaining the provenance of digital objects to a high degree.

Nevertheless, some gaps remain in the system's capacity.

The Rosetta user community is in a strong position to take the initiative in developing a more detailed semantic structure for Events, due to the breadth of experience and the diversity of the various Rosetta customer institutions.

There is also the opportunity for sharing the structure with members of the wider digital preservation community and generating an agreed standard for the representation of provenance information in digital preservation systems/programmes.

Recommendation: Provenance is a fundamental plank of digital preservation. Let's set up a combined User Group/Ex Libris working group to address the issues raised herein.

And so our journey continues …

Steve.knight@dia.govt.nz