PREMIS Editorial Committee Conference Call Notes

17 January 2008

In attendance: Rebecca Guenther, Steve Bordwell, Olaf Brandt, Angela Dappert, Brian Lavoie, Zhiwu Xie, Gerard Clifton (notes).

Apologies: Priscilla Caplan, Bill Leonard, Markus Enders

1. Draft Data Dictionary

A draft of the Data Dictionary (v2.0) is available for review at:
http://pec.lib/uchicago.edu:8888/pec/uploads/1/premis-dd-rev.doc
The draft currently excludes revisions for format and additional objectCharacteristics.
EC members should check the areas for which they were responsible. Steve volunteered to check the whole DD – others may also volunteer. Review comments should be submitted by 30 January.
Brian is revising the introduction and special topics sections and a draft is expected in the near future. As many people as possible should read and review the entire combined document, to find any inconsistencies between parts of the document.
ACTION: All EC members should review the changes for which they were responsible.
[**ACTION:** Rebecca Guenther suggested that she would send out a list of changes and who was responsible for what – perhaps this is unnecessary?]
ACTION: Steve Bordwell volunteered to review the whole DD. Other members may also volunteer. Comments, corrections etc, should be received by 30 January (preferably earlier).
ACTION: Brian Lavoie to circulate draft introduction and special topics when available. As many members as possible should review the entire document.

2. [Change #18:] formatDesignation and formatRegistry

Issue: Two further issues related to the format container have been raised recently:

1. How to record multiple format designations (e.g. where a tool returns more than one identification – Tiff 3.0, Tiff 4.0 etc.)
2. How to indicate that different identified formats for the same object are all correct (conjunction) [e.g. BWF, WAV] or are in conflict (disjunction).

Proposal: Steve Bordwell previously proposed removing the formatDesignation and formatRegistry containers in favour of a simpler, single repeatable format container containing formatName (mandatory), formatVersion, formatRegistryName, formatRegistryKey and formatRegistryRole. (see
http://pec.lib.uchicago.edu:8888/pec/uploads/1/18_formatDesignation_and_formatRegistry.doc).

This allows multiple formatNames to be recorded, can provide an explicit association between a formatName and formatRegistryKey, which may better accommodate automated identification tool output, and can make it clear, in the case of conflicting registry keys, which keys and names may be in doubt. This proposal appears to solve the first new issue.

Discussion: Discussion covered a number of points:

- Conflicting format names or key values should perhaps be resolved through applying policies or rules about which registry may have greater authority or precedence, or by scoring, etc. and this might normally be done at a later time during the validation process. However, you would currently be forced out of PREMIS until multiple formatName values had been resolved.
- You might want to record a status for identification – e.g. candidate, disjunct, etc. – in a new semantic unit to qualify the identification. The new unit could contain a free text note, a reference or a controlled list. It can be optional, and repeatable to allow more than one entry. This could resolve issue 2.
- Test cases for accommodating combinations of format names and keys as the result of manual identification or one or more tool outputs seem OK for the single container proposal. (see http://pec.lib.uchicago.edu:8888/pec/40)
- Making the original format container repeatable, but retaining formatDesignation and formatRegistry containers would also work.
- Making formatRegistry non-repeatable within the original format container would provide an explicit association between format name and key when both are used. To express multiple key values, repeat the whole format container.
- The single container proposal, by necessity, must make some identifier (name or key) mandatory, as did the previous structure, but it may remove user choice about which one to use. The proposal makes formatName mandatory, as this would seem to be a reasonable minimum to know, but some implementations may use a registryKey for recording format. For unidentified formats, formatName could be recorded as “unknown”.
- formatName, registryKey and registryName (the context for the key) could all be made mandatory within a single container, with “unknown” values allowable, but this may cause some confusion.

The two (new) proposed containers are:

A: format **(R)** (M) • formatDesignation (NR) (O) o formatName (NR) (M) o formatVersion (NR) (O) • formatRegistry **(NR)** (O) o formatRegistryName (NR) (M) o formatRegistryKey (NR) (M) o formatRegistryRole (NR) (O) • **formatNote (R) (O)**	B: format **(R)** (M) • formatName (NR) **(M)** • formatVersion (NR) (O) • formatRegistryName (NR) (O) **[M?]** • formatRegistryKey (NR) (O) **[M?]** • formatRegistryRole (NR) (O) • **formatNote (R) (O)**
• Has additional units – more complex • Can associate name / key • Allow combinations, repeats, etc. • Can allow multiple names (by repeat) • Still allows user choice on which mandatory part to record (name, key, both) • Better for X-path queries	• Simpler • Associates name / key • Allows combinations, repeats, etc. • Issues about what should be mandatory (one, other, both?) • Schema rules may be harder to construct or apply.

Decision: There was general agreement to go with option A – the original format container structure, now repeatable, with a non-repeatable formatRegistry container, and new repeatable, optional formatNote semantic unit for qualifying the format identification. The formatNote semantic unit can contain free text, a reference pointer, or a value from a controlled list.

3. Controlled vocabularies

It was suggested that members think about whether to keep the suggested value lists for controlled vocabularies in the Data Dictionary. Another possibility is that the Library of Congress could host the list sets.

It was decided that we should leave the sample value lists in the DD and we can add more later if we need to. They could also be amplified with machine readable lists from EC / LC at a later date. Leaving the lists in the DD is also helpful to those starting out with PREMIS.

4. Open-ended Dates

Some further revision on use of 'OPEN' to express open-ended dates may be done with regard to recommendations on format – e.g. could be either structured dates (like ISO 8601) or not, etc.

[Notetaker's note: I may have missed the gist and outcomes of this quick discussion at the end of the call.]

5. Next call

Thursday, **24 January**

9:00am Eastern US, Ottawa

7:00am Mountain time US

14:00 UK

15:00 Europe/Germany

1:00am (next day) Canberra AU