PREMIS Editorial Committee Call Notes
February 16, 2012

PRESENT: Rebecca Guenther (chair) , Yair Brama, Karin Bredenberg, Priscilla Caplan (recorder), Angela Dappert, Angela Di Orio, Markus Enders, Tracy Meehleib, , Sébastien Peyrard, Robert Sharpe, Kate Zwaard

GUESTS: Ray Denenberg

REGRETS: Karsten Huth, Rob Wolfe, Sally Vermaaten

1. Getting 2.1 records for testing 2.2
We need good 2.1 records for testing to make sure they validate under the new 2.2 schema. Ray pointed out that records contributed to the Registry are not validating under their current versions, so they won't be useful for testing 2.2. Most are missing the schema declaration, some miss required fields. Some that reference MODS 3.1 won't validate but would if they referenced MODS 3.4. Ray will keep on plugging, but if EC members have good PREMIS records they should send them to him. The more 2.1 records we have to test with, the higher our comfort level will be.

This led to a discussion of whether we should review contributed records. Ray suggested we could validate records and reject those that do not validate, but Kate was wary of discouraging folks from contributing records. Tracy suggested we could put a disclaimer on the registry indicating the records have not been checked for quality. Priscilla suggested offering validation and correction of flawed records as a service. Ray pointed out if fields are missing, we won't know what value should have been in them. We could possibly offer to work with the submitters. This issue was tabled for now, as the immediate problem is to get PREMIS version 2.2 out.

2. Issue of open dates
The Rights entity defines two elements, startDate and endDate, to define a date range for the termOfGrant. The problem here is what to do when there is no end date. The EDTF format defines an interval as date/date, allowing "unknown" as a start date and "open" or "unknown" as an end date. This is not compatible with the two elements in PREMIS. There are two options for what to do in PREMIS: 1) put the entire interval in to startDate; this is semantically unsatisfying; 2) use "open" for the value of endDate; this is not legal in EDTF but could be valid in PREMIS, and "open" will validate as an ETDF string.

There was consensus that the second option is preferred, and there is no point in trying to be compatible with EDTF for 2.2 but we should move in that direction. The PIG did not chime in with an opinion. Karin will send a message to the EC to review before forwarding on to the PIG.

3. Two- or three-character country codes
The Rights entity defines copyrightJurisdiction as the country whose laws apply. The Data Dictionary says to use ISO 3166 but does not specify whether to use 2- or 3-character codes. Currently the schema just restricts the value to a string; do we want to restrict it to 2 or 3 letters? Right now the examples are all 2-character codes, but that was not deliberate.

Ray thought that the schema should not be restrictive, but a comment should be added saying the value should be an ISO country code. Karin will draft a comment for Ray. We will also expand the examples in the Data Dictionary to include a 3-character example.

Marcus thought the Data Dictionary should be more explicit, since not everyone uses the PREMIS Schema. He asked if anyone was using MARC country codes, which are not identical to ISO 3166. He thought that people would likely use the same codes they used in other parts of the system, so if they were using bibliographic records with MARC county codes they might carry this over into PREMIS.

Priscilla said that for interoperability, we either have to be more restrictive, or we have to require implementers to specify which code set they are using. Ray pointed out that URIs will be self-

defining, but there are no URIs for the ISO codes yet. He suggested that LC could mint URIs for a controlled vocabulary corresponding to the ISO codes. Meanwhile an optional attribute could be added to the schema for the vocabulary.

Rebecca decided we should table this issue for version 3. Only one element in the Data Dictionary is affected and nobody has complained so far.

4. Regular date for PREMIS calls
Rebecca will put out a Doodle poll for people to indicate the date/times they like.

5. Peter van Garderen's email on backwards compatibility & interoperability
http://listserv.loc.gov/cgi-bin/wa?A2=ind1112&L=pig&T=0&X=6538023B606A57CE5E&P=836
Rebecca will draft a reply and run it by the EC before sending to Peter.

6. Environment Working Group update (Sébastien)
The Environment group has had three meetings since the last PREMIS call. Right now they are diagramming use cases. The following update is taken from Sébastien's notes with minor edits (thank you Sébastien!):

Use cases:

- web archives: we crawled a web page containing an ePub file, document the capture environment (crawler, user agent, server) ; we associate it with the rendering environment in the reading room configured for Firefox 2, and the ePUBreader plugin is not compatible with this version. So there is an update of the environment.
- TIFF to JPEG2000 migration environment: have to record the environment, that is, the migration software, the call parameters used when invoking this migration tool, and the supporting documentation (notably the test bed)
- AVI normalization environment: we take an AVI file and normalise the audio and video streams. Document the service used (software and libraries used)
- Emulation environment: document the original environment of a software game and the rendering environment using an emulator running on an OS and an hardware architecture etc.
- 3 use cases from the TIMBUS project:
    - MRI and Xray scans: their rendering and migration environments
    - Scientific CERN data: documenting the rendering environment where the data can be viewed, and the processing environment where the "facts", the raw data from the experiment can be analysed and processed to have interpretable results. The idea is to be able to store the original data and document the processing environment so that the processing of the data can be re-enacted in the future for peer-review and validation
    - Documenting the business processes of a design multinational enterprise, where the objects to be preserved consist of virtual machines with software / hardware dependencies that allow to match a part of a process to the software that does the actions, so that the full automated part of the workflow can be preserved and re-enacted over the long term.

We came up with 3 possible description styles in PREMIS:

- 1. **Full outsourcing to external registries**. You only make a reference to an environment supporting a particular object, or agent, by referring to an external description of the environment in a registry, e.g. TOTEM.
    - Here you just need a reference from an object, or an agent, to an environment designation / registry. The mechanism would be quite similar to the one we have with formats, with additional units in Agent.
- 2. **Description of an environment as a whole**. The repository decides to be more precise, and describes his whole environment as an instance of PREMIS:Environment. He lists all the components of that environment here, outsourcing more up-to-date / accurate descriptions to external registries whenever deemed relevant.

    - Here you have Environment as a standalone entity that can be linked to from an agent or an object (with a particular role if necessary, which is the environmentPurpose), containing roughly what we already have in PREMIS with additional reference to external registries
    - Maybe move away from software / hardware but have a generic "component" repeatable semantic container which can have a specific type (software, hardware architecture, peripheral, software application, operating system, software library, virtual machine…)
- 3. **Atomic description of the environment as a network of components**. The repository decides to describe each component (application, hardware, operating system, libraries etc.), as a particular environment, then document the dependency links between these different pieces. Each environment description can be outsourced to external registries whenever deemed relevant.

We still have to check if all these description styles are relevant.

Remaining dark areas to solve:

How do we refer to supporting documentation? E.g., SourceForge for a particular piece of open source software, etc. Use linkingObject, or reuse of the documentationIdentifier / role mechanism provided in PREMIS Rights 2.2? Still to be investigated.

How do we record the parameters used when invoking a particular environment? We could relate an environment to an event, and record the parameters in event Detail. But we're keeping in mind the idea that one does not want to add links from everywhere to everywhere in the model.

Next steps:

1. The group decided to **iterate on the use cases** for the **next call on February, 23th**: the idea is that everyone tries to describe its own use case according to the 3 aforementioned description styles. Next call will be centered around these issues, and the "remaining dark areas" mentioned above. We will also check what should be in PREMIS as core preservation metadata and what needs to be outsourced to other schemas (KEEP data model)
2. Once we have a big picture we feel comfortable with, Angela and Sébastien will draft a rough Data Dictionary update with all the new semantic units on environments
3. Then, as a final step, the group will try to describe the use cases using the data dictionary. This should be a way to make sure everything proposed in consistent and useable.
4. [**Edit**: Maybe some implementation guidelines for environment would feed on step 3., wherever they will end up being declared — as separate guidelines or in the Data Dictionary the special topics and/or as an appendix—.]

Deadline problem: this is a very difficult task at hand and we need to do it right for PREMIS 3.0. March doesn't seem a realistic deadline (we can't do it quicker than we do: 3 meetings since last EC call!), but we're making definite progress. It is hard to evaluate when we will have come up with a solution which seems satisfying, but April seems reasonable for this.