Compliance Conundrums: Implementing PREMIS at two National Libraries

Haliza Jailani; National Library Board; Singapore.
Peter McKinney; National Library of New Zealand Te Puna Mātauranga o Aotearoa; Wellington, New Zealand.

Abstract

The purpose of this paper is to examine compliance with PREMIS at National Library Board Singapore and the National Library of New Zealand. It will look in detail at how the development process, variation in content types, existing embedded technologies, and current knowledge all play a role in influencing the shape of the preservation metadata that is created, stored and used in a digital preservation system.

Introduction

PREMIS (Preservation Metadata: Implementation Strategies) is the de facto standard for digital preservation metadata. With a clearly defined scope, it details a large part of the metadata required to manage digital objects across time. This paper uses the experience at National Library Board, Singapore and the National Library of New Zealand to discuss what it means to be compliant with the PREMIS Data Dictionary.

The Data Dictionary is well-written and conveys complex issues in a relatively concise and simple manner. However, this does not necessarily protect against misunderstandings and it certainly does not guarantee that it will not be deliberately or otherwise reinterpreted.

Standards

Standards serve multifarious purposes. For digital preservation, where well-documented, consistent actions, undertaken on fully described and identified content are the cornerstones of success, standards are critical. Briefly they offer:

- Consistency: standards allow implementers to use homogenous metadata to manage their content.
- Consensus: standards are created by experts in the field. Implementers benefit from agreement on best practices.
- Sharing: crucial in the worst-case scenarios, where another organisation must take custody of the content. Sharing content can also aid in mitigating risks.
- History (or perhaps better expressed as 'Memory'): Standards should document why they are being used and the meaning behind their use. This allows future users to understand what was being done and how they can interpret it.

The digital preservation community consistently refers to a number of touchstones. While a variety of standards and frameworks are often invoked (e.g., Trusted Digital Repository, METS) the two main metaphorical pieces of jasper used to test value are in the shape of the OAIS model and PREMIS.

This paper is concerned with conformance to the PREMIS data dictionary.

Compliance

Complying with standards in the heritage sector is a matter of institutional rigour driven primarily by perceived benefit, rather than audited necessity. Which is to say; there are no fiscal ramifications for erroneously asserting compliance (of course, there may be other ramifications, e.g. reputational risk). The benefits of compliance must therefore be sufficiently strong. The task then for implementers (potential and definite) is to judge the benefits against any barriers to conformance.

PREMIS conformance

The PREMIS Committee's statement on conformance[1] suggests that the levels are "lightweight, and considerable scope for flexibility and choice is reserved for implementing repositories". [1] There is perhaps some dissonance here though with a statement in the Dictionary that says it lists '"implementable metadata": rigorously defined' [2]. With such rigour a more stringent conformance level would be expected.

In terms of benefits that should drive conformance, the Committee's list includes inter-repository data exchange, certification, shared registries, automation, and vendor support [3]. This list is, unsurprisingly, mostly in accord with the benefits listed above.

Institutional Background

Both organisations use the Rosetta digital preservation system, which was developed by Ex Libris in conjunction with the National Library of New Zealand (NLNZ). While care will be taken to clarify where implementation choices have been made by the institutions and where decisions have been made by the vendor, the development role played by NLNZ means that this boundary is not fully demarcated.

The National Library of New Zealand has a legal mandate to collect and preserve documentary heritage and taonga (treasured items) for all people of New Zealand [4]. New Zealand legislation explicitly includes digital content as falling under this mandate. In practice, the Library collects and receives content in variegated formats.

National Library Board, Singapore (NLB) is mandated the function of preserving the published heritage of the nation through legal deposit. By law, every Singapore publisher must deposit copies of every publication published in the Republic with the NLB [5]. The Board has undertaken the task of preserving this collection as part of its responsibility. A review conducted by the

[1] For good or ill, we use 'compliance' and 'conformance' interchangeably in this paper.

Board in 2005 had recommended the building of infrastructure and a centralised database for the preservation and access of Legal Deposit materials and the wider ambit of national heritage materials [6]. To this end, the Rosetta system was implemented.

PREMIS Implementation

As stated above, both organisations implement the Ex Libris Rosetta system. Their implementation of PREMIS is through Rosetta. Rosetta undertakes various processes on content as it is ingested, adding metadata to the intellectual entity. Simply, the end result is that the content files are placed in the permanent storage along with a METS file, which contains the metadata that both organisations have deemed to be required for permanent preservation of the content. This METS file contains, but not exclusively, data that is expressed in a schema called "the DNX". This in turn contains, again, not exclusively, the PREMIS Data Dictionary.

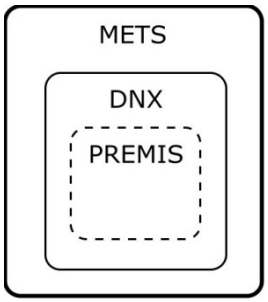

Figure 1. *Structure of Rosetta METS file*

Examining conformance

Both Libraries undertook comparisons of the data contained within their METS files against the PREMIS data dictionary. This comparison included checking:

1. The nomenclature used;
2. The semantics of this nomenclature;
3. The sub-level of object that the units are used at;
4. The obligation associated with the units; and,
5. The repeatability of the units.

The following section will attempt to highlight one or two key areas in each of the five areas of comparison.

Nomenclature and Semantics

With the exception of *objectCharacteristicsExtension* which is optional, the semantic components of PREMIS *objectCharacteristics* are present in Rosetta's DNX albeit with some name variations. They are mostly captured under a DNX element called *generalFileCharacteristics*. PREMIS format and fixity semantic components are captured separately under DNX f*ileFormat* and *fileFixity* container units respectively. These containers group the capture of granular details such as the different values for semantic components of fixity types MD5, SHA1 and CRC32, in a more user-friendly way.

Of note however is the use of the metadata element *objectCharacteristics* which does not share the definition of PREMIS *objectCharacteristics* as its sub-units relate to metadata such as *objectType*, *parentID, groupID, creationDate, createdBy, modificationDate, modifiedBy* and *owner.* The DNX *objectCharacteristics* is not used for format specific technical metadata as defined in PREMIS. In addition, this metadata element applies to the representation, file and bitstream levels while PREMIS *objectCharacteristics* is applicable to only file and bitstream. Although confusing at first, its use was adopted as there were no adverse impact on internal operations. External conformity might be an issue, although it is possible that this container unit might be excluded when extracting PREMIS-conformant information from Rosetta for another repository.

Another non-conformance is *storageMedium* specified by PREMIS to be applicable only at the file and bitstream level. Rosetta defines this semantically in p*hysicalCarrierMedia* under *generalRepCharacteristics,* a semantic container at the representation level.

In NLNZ and NLB's business-as-usual routines, the outputs of our metadata extractors are mapped directly to the *significantPropeties* section in the DNX (*significantPropertiesType, Value and Extension*). In essence, it is the dumping ground of the technical properties as they are culled from the file. This information should, in order to conform, be placed in the *objectCharacteristicsExtension* section which is meant for additional object characteristics from format-specific technical metadata schemas such as the Z39.87-2006. The *significantProperties* container should be used to store properties "determined to be important to maintain through preservation actions" [7]. It is intended to enable flexibility for implementers but has caused inconsistency instead. Notably, the inclusion of external format-specific technical metadata is more easily done in PREMIS *significantProperties* than in PREMIS *objectCharacteristicsExtension* where explicit associations will require repeating the entire semantic unit and it is recommended that information about the external metadata be provided.

There is a deeper discussion to be had however. NLNZ and NLB believe all technical properties to be important, irrespective of whether or not they should remain across an action. Some properties we may actually want to deliberately take action to remove from the file. These properties are significant and must be tracked across actions. This does not detract from the non-conformance, however it does raise questions about the purpose of the significant properties section in PREMIS.[2]

Event Entity

PREMIS *event* information is recorded at file level in Rosetta with all semantic units except for *linkingObjectIdentifie*r, present. In addition, event outcomes are detailed separately under DNX *vs Outcome* element with sub-units such as *checkDate, agent, type, result, resultDetails, vs Evaluation* and *vs EvaluationDetails.* These record information for specific events such as validation for checksum, file format, technical metadata, virus checks and risk analysis with the clear intention of ensuring clarity in detailing these checks. There is no non-compliance for repositories to capture more detailed information for a PREMIS semantic unit

[2] We are keenly aware that there is a general agreement across DP literature with the PREMIS description. See for example [8].

than what is defined in the Data Dictionary. This type of flexibility allows for a sufficient level of consistency and encapsulates the "implementable metadata" that is the intention of PREMIS. It has enabled both libraries considerable leeway in capturing additional required information.

Agent Entity

As would be expected of an implementation system, PREMIS a*gentIdentifier* and *agentName* are implemented in Rosetta as metadata associated with individual events such as file fixity, virus check, file format, checksum and techMD outcomes. Other optional semantic components for this entity such as *agentType* and *agentNote* are not explicitly defined in Rosetta's user interface.

Rights Entity

Rights semantic units are all optional at the container level. Except for access rights, these are not explicitly defined in Rosetta's user interface although PREMIS indicated that the minimum rights information a repository should know is the rights to carry out preservation actions.

Object level data

PREMIS describes three levels of object: 'representation', 'file' and 'bitstream'. In addition to these three, NLNZ and Singapore both use the level of Intellectual Entity as the primary unit of understanding digital content. The PREMIS Editorial Committee has stated that it is looking at the level of IE for the next version. This promises to be a strong addition to the dictionary.

Examination of the bitstream level raises some important questions. This is an area that during development of the system was given a good deal of attention, but still remains a little 'fuzzy'. Which is to say, there is a large legacy of diagrams, papers and requirements that try to finesse how bitstreams should be dealt with. However, a number of factors led to this part of system being as not well-resolved as the rest of it. Where is the boundary between bitstream and file? Without this boundary it is very hard to define exactly what the required functionality is. Without an exact requirement, its importance is questioned. This in turn means that if the requirements cannot be rigorously defended, then there is no strong driver (or will) to conform. This is discussed in more detail below.

Obligation

In Rosetta, obligation (the quality of being mandatory or not) denotes that a value is required to aid in processing the object through any number of its functions. It could be argued that this is a valid interpretation of the PREMIS definition, which states "A mandatory semantic unit is something the preservation repository needs to know" [9]. Regardless of interpreting what obligation means, there are some differences. For example, it is clear to both NLNZ and Singapore that fixity is a mandatory piece of information: it is a basic unit of tracking integrity and must be included with all files. It is only optional in PREMIS. PREMIS does have the correct sentiment: "Objects that lack these features [fixity, integrity, and authenticity] are of little value to repositories that have a mission to protect evidentiary value or indeed to preserve the cultural memory" [10]; but does not make fixity a mandatory element.

This deviation does not make the implementation non-conformant, as obligation can be made more stringent without affecting conformance. But it does raise a question as to why this specific unit is optional in PREMIS.

Repeatability

In terms of repeatability, one of the more interesting differences is with format identification. PREMIS allows for repeatability of the format container. This means that multiple formats can be recorded against an object. We require however that each format coming into Rosetta is given a primary identification. This definite identification is the major driver of search, risk analysis, and preservation planning functionality. Multiple IDs with the same importance would impede this process. This is not to say that we only store one format identification. We collect and manage format identification from DROID, JHOVE, NLNZ MET, and the internal format library. But it does mean that we need to deviate from the Data Dictionary in order to be able to a) identify the definitive format identification, and b) capture the variety of format information we collect. This information is displayed in Table 1 below.

Table 2 shows how this information could be presented in PREMIS. Across the two tables, the example is of a TIFF file being identified. Until currently, DROID has identified TIFF files with multiple identifiers.[3] So in the PREMIS example, all the identifiers could be put into PREMIS. But it allows us no concept of primacy, and also does not give us other details that Table 1 does. For example, in Rosetta we also capture format information from the MD extraction process (in this case, JHOVE suggests that the file is TIFF version 5). Crucially though, the issue that in Table 1, the ID value that is used by the system is the *formatLibraryID*. In table 2, there is no clear field that would be used. The purpose of tables 1 and 2 is not to describe in detail the flows that lead to a given result in each unit (for Rosetta, these flows are complex and require more space than available to describe), but rather to show that while PREMIS allows repeatability of the format container, it does not allow us to specify which container is to be used as the definitive format identification.

[3] DROID has recently added a new classifier for TIFFs that is a generic container for these multiple hits.

Table 1: Format Identification Information in DNX

NLNZ/NLB Unit	Value	Description
generalFileCharacteristics.fileMIMEType	audio/tiff	From Format Library
generalFileCharacteristics.formatLibraryID	Fmt/7	Definitive Format identification
generalFileCharacteristics.fileExtension	tif	From file name
fileFormat.agent	DROID	The tool used to identify the file
fileFormat.formatRegistry	PRONOM	The registry from which the registry ID comes from
fileFormat.formatRegistryID	Fmt/7	The registry ID
fileFormat.formatRegistryRole	-	The purpose of the registry
fileFormat.formatName	Fmt/7	The format ID as negotiated between the agent and the Format Library
formatVersion	-	The format version as negotiated between the agent and the Format Library
fileFormat.formatDescription	Tagged Image Format	The textual name of the format
fileFormat.formatNote	-	Notes as assigned during manual identification
fileFormat.exactFormatIdentification	FALSE	Notes whether identification was by tool or not.
fileValidation.Format	TIFF	Format identification from the MD extractor
fileValidation.Version	5	Format version from the MD extractor
fileFormat.mimeType	audio/tiff	Mime-type from the Format Library
fileFormat.agentVersion	5	Version of the identification tool used.
fileFormat.agentSignatureVersion	50	Version of the signature used during identification

Table 2: Format Identification in PREMIS

PREMIS Unit	Rosetta Equivalent	PREMIS Value
formatRegistryName	formatRegistry	PRONOM
formatRegistryKey	formatRegistryID	fmt/7
formatRegistryRole	formatRegistryRole	specification
formatName	formatName	audio/tiff
formatVersion	formatVersion	3
formatNote	formatNote	-
formatRegistryName	formatRegistry	PRONOM
formatRegistryKey	formatRegistryID	fmt/8
formatRegistryRole	formatRegistryRole	specification
formatName	formatName	audio/tiff
formatVersion	formatVersion	4
formatNote	formatNote	
formatRegistryName	formatRegistry	PRONOM
formatRegistryKey	formatRegistryID	fmt/9
formatRegistryRole	formatRegistryRole	specification
formatName	formatName	audio/tiff
formatVersion	formatVersion	5
formatNote	formatNote	-
formatRegistryName	formatRegistry	PRONOM
formatRegistryKey	formatRegistryID	fmt/10
formatRegistryRole	formatRegistryRole	specification
formatName	formatName	audio/tiff
formatVersion	formatVersion	6
formatNote	formatNote	

Case study of unclear conformance

As noted above, the PREMIS model has three levels of objects. These are 'representation', 'file' and 'bitstream'. At these levels, different types of semantic units are retained, and it is this information that is used to manage the objects in the preservation repository. The conformance statement is clear when it states that if a repository chooses to store information about the file level (for example), that all the mandatory semantic units for that level must be recorded, but that mandatory information for the other levels are not required to be recorded. [11]

There is arguably though, a gap in the conformance documentation around the 'correctness' of the level of information. The data dictionary is very specific about what constitutes a file and what constitutes a bitstream. However, there is no statement of conformance expressing that implementers must follow the correct classification of these object types.

For example, PREMIS states that audio data within a WAVE file is bitstream level. [12] NLNZ currently chooses to write this information at the file level. Correspondence on the PREMIS listserv has indicated that other implementers write the audio information also to the file level. Should this be classed as non-conformance?

Further examples have brought to light other areas of practice that deviate from the PREMIS examples. The data dictionary uses two more examples to highlight the difference between file and bitstream. A single image TIFF file should have all information about the image written to the file level, but in a multi-image TIFF, all the image information should be written to discrete bitstream levels (one for each image within the file) [13]. Again, discussion on the listserv showed that practice deviated from the prescribed classification of the object sub-types. Some institutions wrote all image information, irrespective of whether or not it came from a multi-page TIFF, to the bitstream level. NLNZ and NLB currently write it all to the file level.

The question here is the importance of such differences across institutions. What does it mean if the same information is being written to different levels by different institutions? This question is of particular merit when juxtaposed with the benefit of sharing. The authors are of the view that the difference should not exist Every institution should be writing the same information to the same level. It goes against the purpose of uniformity; of trying to adhere to community guidance and practice, and it most certainly constrains sharing.

There should be clear guidance on classifying files and bitstream. Accompanying this guidance, there should be a statement of conformance. Such a statement would determine that conformance requires implementers to be correctly classifying files and bitstreams. It is clear that this is no small task, not least because it requires a strong and clear description of what exactly the differences between file and bitstream are. Guidance would need a multitude of examples to clarify fully. A simple example helps display this point. The formatting of this paper follows a very well-defined template. But even this relatively simple task of guiding how to format a paper for Archiving 2012 has issues. There are some methods of writing that the authors use which are not covered by the template (let's say, footnotes). We have then had to make a decision on what to do, informed by our experience so far, but without formal guidance. If we move this into the realm of files and bitstreams, where the problem is far more complex, what chance is there then of being able to generate full guidance on what is a file and what is a bitstream? How many examples are needed? Should there be minimal examples, allowing institutions to make decisions by inference? Or, should there be full and incontrovertible rules?

A critical role in this differentiation could be taken by tools such as DROID, JHOVE, and NLNZ Metadata Extractor. These are the tools that generate a lot of the technical information that is found within the files and bitstreams. If these tools generated their outputs in terms of PREMIS mappings, then implementers would have no issues in determining where information should be written. We (the authors) have a route to make this happen for the NLNZ extractor, but it is difficult to devote development time if there is no clear benefit in doing so.

Discussion on benefits

Neither NLNZ nor NLB are conformant with PREMIS, but implement a version of PREMIS through their use of Rosetta.

The issue is whether there are degrees of non-compliance. What does it mean that objectIdentifier is an optional unit for both Libraries, but mandatory in PREMIS? How important is it that the significant properties section is used to store technical characteristics that should be maintained in the objectCharacteristicsExtension unit? While details are good to have in an implementation, the level of details provided by repositories can vary. Unless there is a minimum level of conformity and subjectivity is kept to a minimum, compliance statements will still result in data inconsistencies.

Linking identifiers and most PREMIS extension units are not explicitly indicated in Rosetta and consequently unused by both libraries. Linking identifiers are important for associating entities where relationships are expressed as linking information in the form of identifiers. Extension units are perhaps not as critical but are still important in affording both libraries a level of freedom in adding required descriptors from other schemas. For such information, effort will need to be put in by the libraries to express existing relationships built into the system and incorporate the extensibility required.

As noted above, the key benefits of standards and compliance with them are consistency, consensus, sharing, and history. This final section explores these benefits in terms of the conformance exercise undertaken by the National Library Board, Singapore and National Library of New Zealand.

Consistency

This benefit looks at consistency both within single repositories and across repositories. The former allows for better management and processes, the latter supporting sharing and the development of tools that can take advantage of the consistencies. Despite any deviations from PREMIS, internal consistency is achieved in both organisations. There is a sound data model with documentation covering usage. Many of the processes of creating and updating metadata are covered by the system, and the robust testing by the vendor and institutions ensures that data is written to the correct places, with the correct values, in the correct form.

This benefit begins to lessen when discussed in terms of consistency across institutions, particularly institutions that are not using the same preservation system. The key issues here are the misuse of the significant properties unit, and the lack of clarity on files and bitstream boundaries.

Consensus

Deviation from PREMIS means deviation from the expert group that guided the creation of the Dictionary. It is clear that in some cases, deviation has been well-reasoned internally during development (for example, fixity being mandatory). The PREMIS committee is preparing version 3 of the Data Dictionary, so the full picture shape of preservation metadata has not yet been fully drawn. However, we do believe that the picture is very close to completion, with only some highlights and detail required to be added. Some of the deviation by NLNZ and NLB is driven by system need (the difference in obligation for example).

Sharing

By not conforming exactly to PREMIS, some avenues of sharing are, perhaps not lost, but made more difficult. What is important here is to understand the contexts in which sharing is envisaged to take place. Sharing from NLNZ is seen only in extreme circumstance where the National Library is not able to function any more. Where possible, preservation of the content would hopefully then be taken on by another institution. It is well-known, and tragically expressed very recently, that New Zealand sits on an area of great seismic activity. However, it is still considered highly improbable that the National Library will ever be in a position that it would not, nor could not, care for content in its collections. This benefit is therefore a little less immediate than for organisations that have as part of their routines regular sharing.

History

This benefit is entirely based on good documentation. Good documentation that is reliable and complete allows future users to have a clear picture on the shape of the information they are looking at.

It is clear that the NLNZ and NLB implementations are not entirely conformant with PREMIS. However, they are most of the way there. In addition there is clear documentation detailing which parts of PREMIS are used in line with PREMIS, where deviations exist, why those deviations exist, and how to correctly understand the data written into such fields.

Conclusions

This paper is a necessarily brief foray into the data models as used by National Library Board, Singapore and the National Library of New Zealand and how they relate to the PREMIS Data Dictionary. We have highlighted some of the areas that we thought are most interesting and raised questions that we believe the community could help answer. Implementation does not take place in a vacuum. In the case of NLNZ and NLB it took place during a period of highly pressurised development. This development was done when the PREMIS Data Dictionary was a complete and accepted community touchstone. It also took place in parallel with the development of the Dictionary (NLNZ began working on preservation metadata around 2001, and the PREMIS Data Dictionary came out of work begun by OCLC and RLG in 2001). It has been suggested that the PREMIS implementation process "involves adaptation, and this adaptation involves many steps that are deeply iterative in nature" [14]. This has certainly been the case in our respective organisations, and we are still undertaking this process. Preservation flows have now been business-as-usual in the Libraries for the last three years, and we are still learning about the information we need and how that information should be written. Constant evolution is the stage we are at, and the stage we will remain at. The variety of data we collect, the changing requirements for process driven from the preservation data, and outputs from community research all mean that we cannot remain static in terms of the preservation metadata we all use.

References

[1] PREMIS Editorial Committee, Conformant Implementation of the PREMIS Data Dictionary, (October 2010), pg. 6.

[2] PREMIS Editorial Committee, PREMIS Data Dictionary for Preservation Metadata, version 2.1, (January 2011), pg. 1.

[3] PREMIS Editorial Committee, Conformant Implementation of the PREMIS Data Dictionary, (October 2010), pg. 1.

[4] National Library of New Zealand (Te Puna Mātauranga o Aotearoa) Act 2003, http://legislation.govt.nz/act/public/2003/0019/latest/DLM191962.html.

[5] National Library Board Act, 1996. See http://statutes.agc.gov.sg.

[6] CDNLAO Newsletter, No. 56, July 2006 Special topic: legal deposit system, http://www.ndl.go.jp/en/cdnlao/newsletter/056/564.html.

[7] PREMIS Editorial Committee, PREMIS Data Dictionary for Preservation Metadata, version 2.1, (January 2011), pg. 39.

[8] Andrew Wilson, Significant Properties Report, http://www.significantproperties.org.uk/wp22_significant_properties.pdf.

[9] PREMIS Editorial Committee, PREMIS Data Dictionary for Preservation Metadata, version 2.1, pg. 23.

[10] *Ibid*, pg.209.

[11] PREMIS Editorial Committee, Conformant Implementation of the PREMIS Data Dictionary, (October 2010), pg. 3.

[12] PREMIS Editorial Committee, PREMIS Data Dictionary for Preservation Metadata, version 2.1, pg. 7.

[13] *Ibid.,* pg. 7.

[14] Devan Ray Donaldson and Paul Conway, "Implementing PREMIS: a case study of the Florida Digital Archive," Library Hi Tech, vol 28. No. 2, pg. 282. (2010)

Author Biographies

Haliza Jailani is Asst Director (Metadata Services) at the National Library Board of Singapore. She was responsible for the implementation of metadata for digital objects and had served as Project Manager and implementer of the Digital Preservation System for NLB. She received her MSc from the University of Central England.

Peter McKinney received his MA and MPhil from the University of Glasgow. He has worked for HATII at the University of Glasgow, and taken part in the ERPANET and ESPIDA projects. He drove vans before moving to New Zealand. Currently, he is the Digital Preservation Policy Analyst at the National Library of New Zealand.