Lisbon, September 5th 2013

IPRES 2013

Preservation Health Check: work in progress

Workshop
PREMIS implementation fair 2013

Titia van der Werf

Senior Program Officer
OCLC

biblioteca de praia

Horário:

De segunda a sexta
Das 10 às 13 e das 15 às 18h

Contactos:

Biblioteca Municipal de Sesimbra
Av. da Liberdade, n.º 46 | 2970-635 Sesimbra
Tel.: 21 228 85 88 | Fax: 21 228 87 01
E-mail: biblioteca@cm-sesimbra.pt

Pólo de Leitura da Quinta do Conde
Av. Cova dos Vidros - Junta de Freguesia
2975-333 Quinta do Conde
Tel.: 21 210 10 22 | Fax: 21 210 24 42
E-mail: plqc@cm-sesimbra.pt

Sesimbra um mar de emoções... todo o ano.

Sesimbra
câmara municipal
www.cm-sesimbra.pt

What is the Preservation Health Check Pilot?

Joint initiative:

- Open Planets Foundation (OPF)

  A community hub for digital preservation whose main goal is to jointly manage and improve tools and research outcomes for practical use.

- OCLC Research

  A community resource for shared R&D that addresses challenges facing libraries and archives in a rapidly changing information technology environment.

The proposition

As part of their preservation management task, repository managers need to be able to monitor the preservation status of the content of their repository.

We are looking at regular “routine check-ups” that can support this monitoring task.

- Monitoring should be made easy (automatically generated reports or dashboard)
- Monitoring should be based on objective data, generated by the repository (e.g. preservation metadata)

The analogy

The research question

If a Preservation Health Check is a monitoring activity to be performed on a repository with digital content

1. What are empirical indicators (i.e. measures) for PHCs?
2. Are preservation metadata recorded by repositories useful as health indicators for PHCs?

Monitoring is about tracking change ... intentional and unintentional change.

The analogy

Analogy with a car dashboard, involving sensors, thresholds, and triggers.

The research methodology

The design of the Preservation Health Check is based on both top-down & bottom-up approaches:

Top-down:

work with existing models (PREMIS and SPOT) that define properties of successful preservation and indicators (threats) of what theoretical could go wrong;

Bottom-up:

work with real metadata and assess their applicability for sensing what needs attention and for triggering preventive actions.

The research methodology

Goal:

To develop an implementable logic (or protocol) to support PHCs, and to test this logic against the store of preservation metadata maintained by an operational preservation repository.

The pilot site

The BnF runs a fully operational trusted digital repository (SPAR). They volunteered to become a PHC-pilot site.

The empirical data consists of:

1. A sample (200 GB) of the PREMIS data (AIP-METS files), covering the following collections:
   - Gallica = digitised periodicals, monographs, still images and manuscripts (TIFF + OCR-files)
   - Legal deposit Web harvests (warc files)
   - 3rd party collection (Centre Pompidou)

The pilot site

The empirical data consists of (continued):

2. All the Reference Information packages in SPAR that contain reference information/code/specifications of (external) tools used during INGEST (ex. JHOVE) and of formats ingested;
3. Per collection: SLAs defining policy agreements with SIP suppliers concerning the preservation regime to be applied at the INGEST and ARCHIVAL STORAGE stages.

PHC-pilot stage 1: Top-down approach

Mapping PREMIS semantic units to SPOT properties: which semantic units address each of the 6 basic properties of successful preservation defined in SPOT?

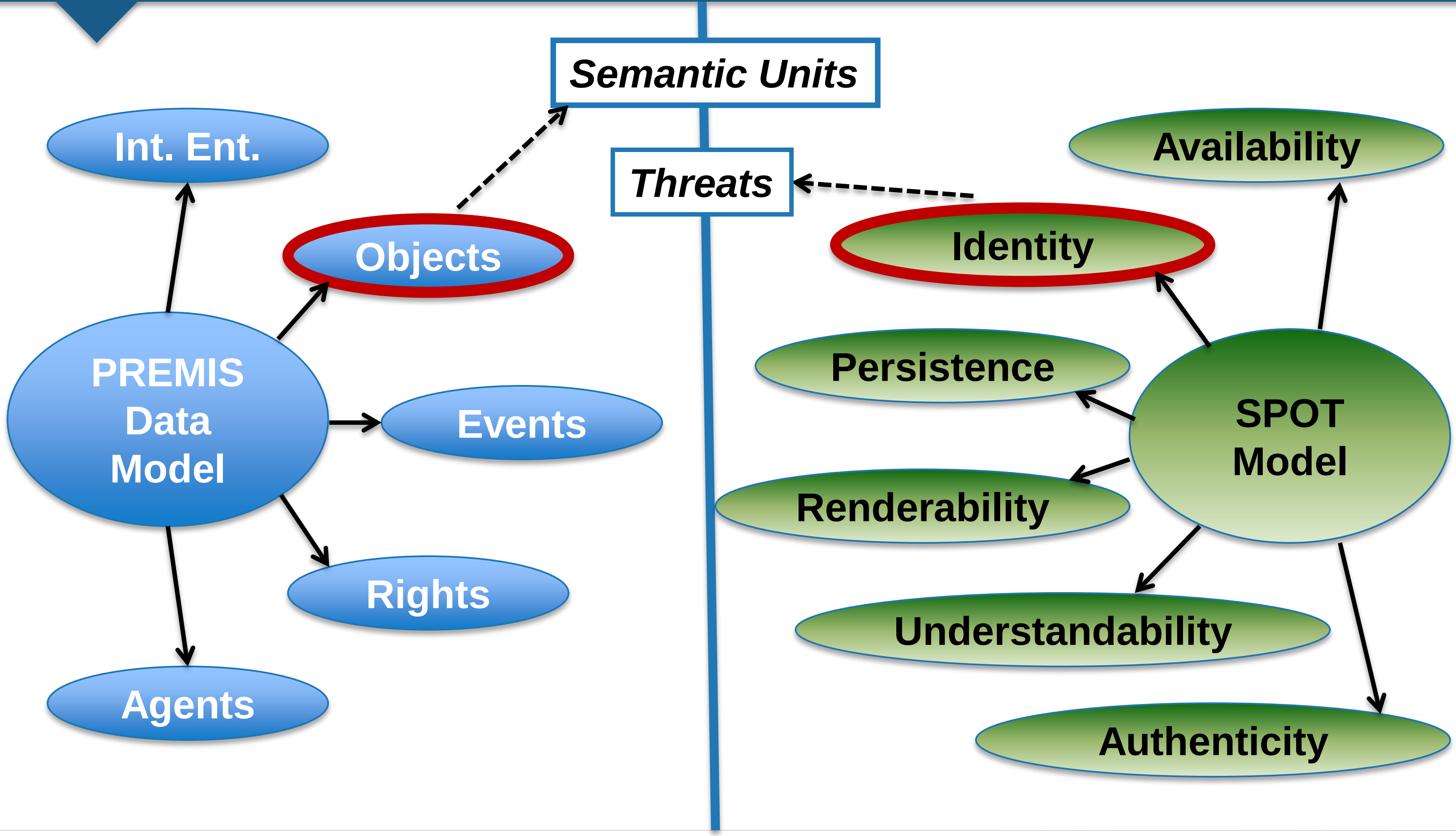
Example:

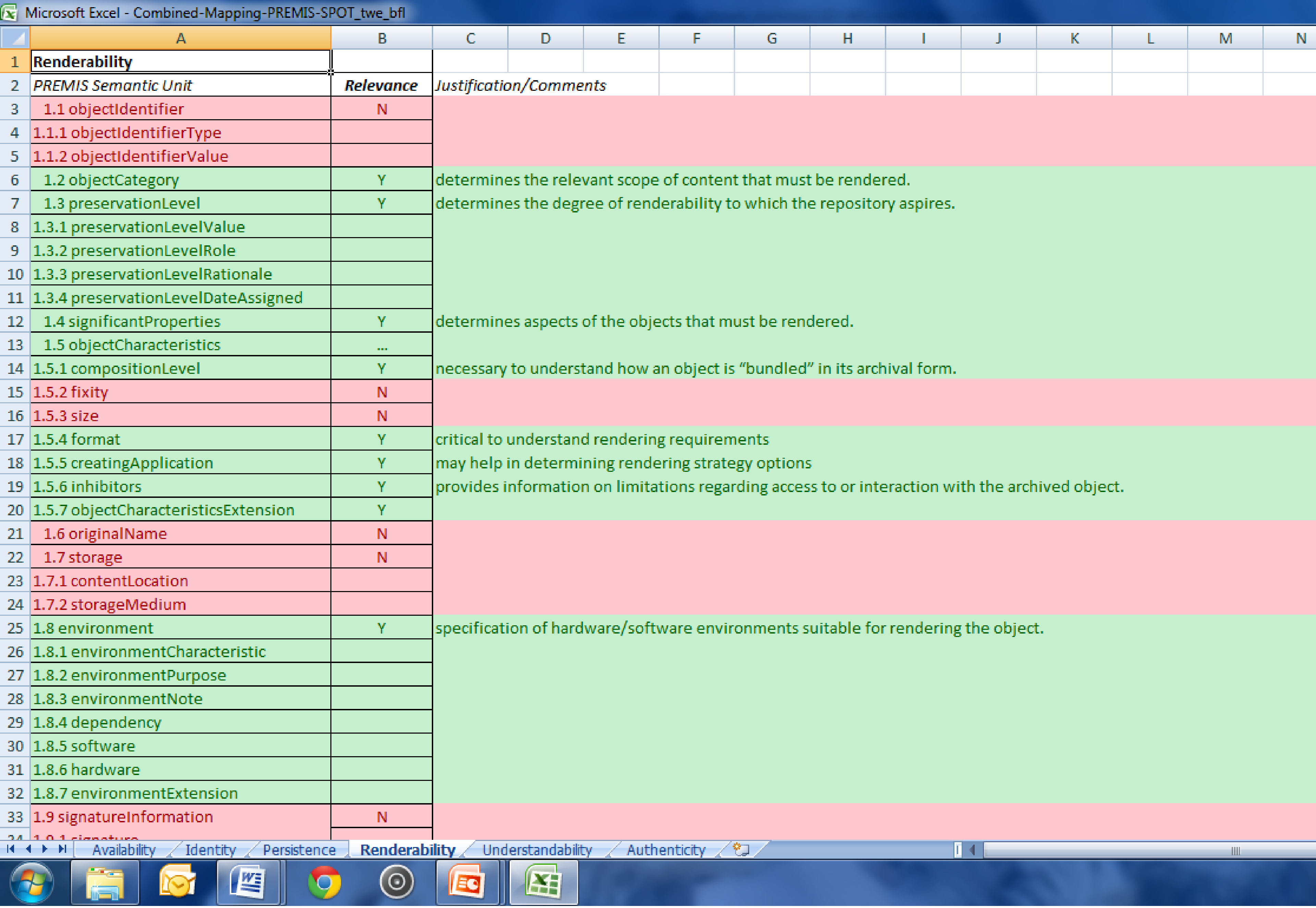
SPOT property: **Persistence →** `associated threats (e.g. storage medium deterioration)`

PREMIS semantic units

storageMedium = `magneticTape`
eventType = `mediaRefreshment`
eventDateTime = `1998-07-31`

Mapping PREMIS on to SPOT





Renderability		
PREMIS Semantic Unit	***Relevance***	*Justification/Comments*
1.1 objectIdentifier	N	
1.1.1 objectIdentifierType		
1.1.2 objectIdentifierValue		
1.2 objectCategory	Y	determines the relevant scope of content that must be rendered.
1.3 preservationLevel	Y	determines the degree of renderability to which the repository aspires.
1.3.1 preservationLevelValue		
1.3.2 preservationLevelRole		
1.3.3 preservationLevelRationale		
1.3.4 preservationLevelDateAssigned		
1.4 significantProperties	Y	determines aspects of the objects that must be rendered.
1.5 objectCharacteristics	...	
1.5.1 compositionLevel	Y	necessary to understand how an object is “bundled” in its archival form.
1.5.2 fixity	N	
1.5.3 size	N	
1.5.4 format	Y	critical to understand rendering requirements
1.5.5 creatingApplication	Y	may help in determining rendering strategy options
1.5.6 inhibitors	Y	provides information on limitations regarding access to or interaction with the archived object.
1.5.7 objectCharacteristicsExtension	Y	
1.6 originalName	N	
1.7 storage	N	
1.7.1 contentLocation		
1.7.2 storageMedium		
1.8 environment	Y	specification of hardware/software environments suitable for rendering the object.
1.8.1 environmentCharacteristic		
1.8.2 environmentPurpose		
1.8.3 environmentNote		
1.8.4 dependency		
1.8.5 software		
1.8.6 hardware		
1.8.7 environmentExtension		
1.9 signatureInformation	N	
1.9.1 signature		

Findings: coverage

SPOT property	# of PREMIS semantic units*
• Availability	16
• Identity	19
• Persistence	10
• Renderability	15
• Understandability	14
• Authenticity	16

**Container level only; Agents, Events, Rights considered one semantic unit*

Findings: coverage

- What does the question “*Is this SPOT property well-covered by PREMIS*” mean?
- More meaningful: Do the PREMIS semantic units address the threats associated with a SPOT property?

Example of a gap between SPOT and PREMIS:
SPOT property: **Understandability**
We found no PREMIS semantic units that provide information that aids in the understanding or interpretation of the *content* of the archived digital object.

Findings: coverage of policies

A repository usually implements a large number of explicit and implicit policy decisions; however, PREMIS currently makes few provisions for recording these in preservation metadata (the semantic unit *preservationLevel* being a notable exception).

The issue is exacerbated if there are numerous policies applied at the collection level, rather than repository wide.

Findings: autonomy of the repository

The PREMIS Data Dictionary seems to be designed around an implicit assumption that the repository is a self-contained system, and that all digital preservation processes are controlled "in-house".

Example:

SPOT property: **Identity**

"Recommended practice is for repositories **to use identifiers automatically created by the repository as the primary identifier in order to ensure that identifiers are unique and usable by the repository**. *Externally assigned identifiers can be used as secondary identifiers in order to link an object to information held outside the repository." [PREMIS DD 2.2 p.28]*

Findings: explicit encoding

PREMIS conformance does not require explicit encoding of metadata if the information applies to all objects in the repository.

This impedes the provision of automated PHC services (by a third-party provider) because efficient provision of this service would likely require the information in semantic units to be explicitly recorded, and implemented in a standard way.

Findings: assessment level

What is the entity we are monitoring during a PHC?

Digital object? Collection? Repository?

We observe that the threat assessment level depends on the nature of the specific threat.

Examples: Identity => repository-wide;

Renderability => collection of objects sharing same HW/SW environment

The SPOT model does not explicitly specify this “granularity of analysis” for the properties and threats it covers.

Conclusion

Despite some gaps in both models used, there is indeed opportunity to use PREMIS preservation metadata as an evidence base to support a threat assessment exercise based on the SPOT model.

We will continue this work as a basis for the PHC design.

Next steps

- Choose a SPOT property that is well addressed by PREMIS (persistence?)
- Develop a generalized logic that makes threat assessment statements

Example of Logic :

Compute elapsed time between last media refreshment event and current date.

- If (MTtoF – elapsed time) > Critical period, return Green
- If (MTtoF – elapsed time) < Critical period, return Red

Next steps

- Test the “implementability” of this logic on a set of “real-world” preservation metadata
- Construct a decision-tree-based PHC-dashboard

The analogy

Wikipedia entry for PHC

Primary health care, often abbreviated as "PHC", has been defined as "essential health care based on practical, scientifically sound and socially acceptable methods and technology, made universally accessible to individuals and families in the community. It is through their full participation and at a cost that the community and the country can afford to maintain at every stage of their development in the spirit of self-reliance and self-determination“

World Health Organization. Declaration of Alma-Ata. Adopted at the International Conference on Primary Health Care, Alma-Ata, USSR, 6–12 September 1978.

Titia van der Werf
titia.vanderwerf@oclc.org

Q&A