Toronto, Ontario

2 October 2012

Preservation metadata as an evidence base for risk assessment

Brian Lavoie

Research Scientist
OCLC

Roadmap

Preservation metadata as an evidence base for risk assessment

PREMIS & SPOT

Mapping & examples

Preservation metadata

PREMIS Data Dictionary 2.2 (p.1):

The Data Dictionary defines preservation metadata that:

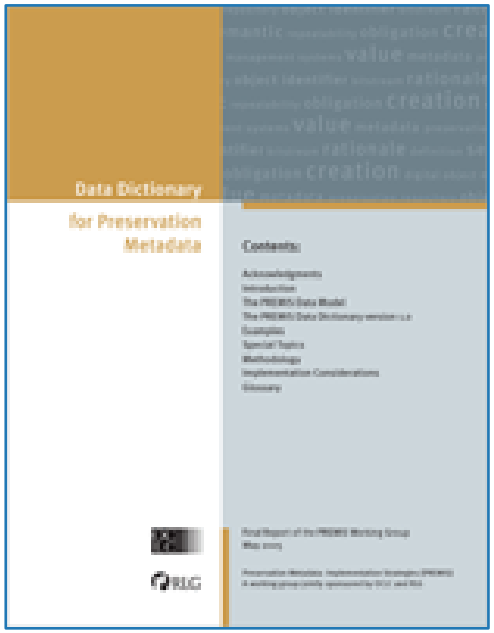
Supports the viability, renderability, understandability, authenticity, and identity of digital objects in a preservation context

Represents the information most preservation repositories need to know to preserve digital materials over the long term

Threat models

"Digital preservation strategies must address the threats relevant to the specific repository context in which they are expected to operate; this in turn requires an understanding of the full range of potential threats so repository staff can evaluate the likelihood and impact of each in the context of local circumstances, and take appropriate steps to address those threats representing significant risk."

Vermaaten, Lavoie, Caplan (2012) "Identifying Threats to Successful Digital Preservation: The SPOT Model for Risk Assessment" *D-Lib Magazine*



Data about viability, renderability, understandability, authenticity, identity of archived digital objects:

In the form of semantic units defined as properties of Objects, Events, Rights, Agents

SPOT

DRAMBORA

TRAC

…

Risks to achieving viability, renderability, understandability, authenticity, identity:

In the form of enumerations of threats potentially impacting digital preservation process

Preservation Metadata

Evidence Base

INFORMS

Risk Assessment (within threat model context)

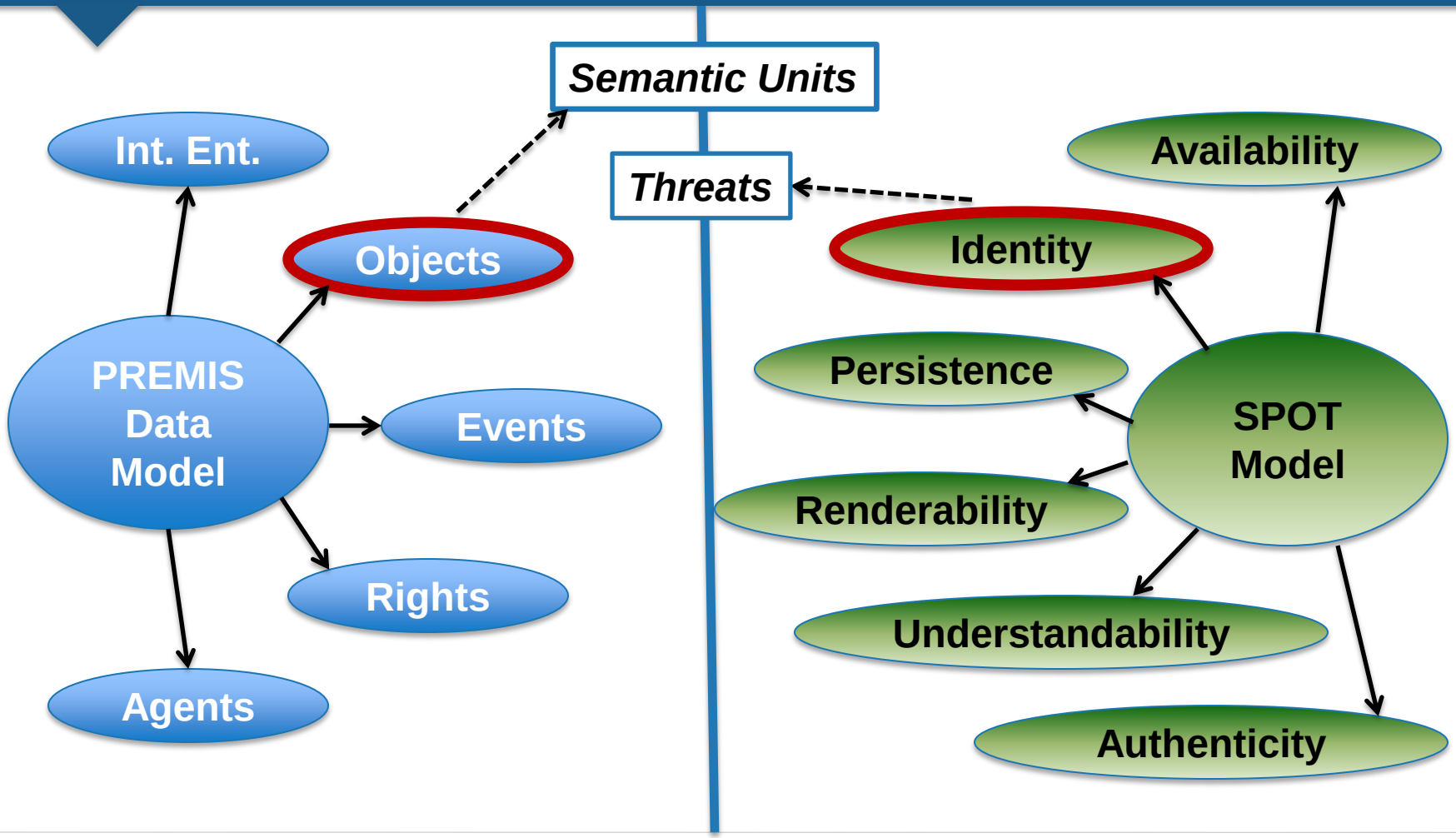
PHC: Goals

- Standardize use of preservation metadata as a basis for conducting risk assessment exercises
  - Map established preservation metadata standard (PREMIS) to a threat model (SPOT)
  - “Protocol” for metadata-based risk assessment
- Focus on actionable intelligence & automated analysis
- Analyze gap between metadata recorded “in practice” and metadata needed “in theory”
- Establish use case for well-maintained preservation metadata

PREMIS & SPOT

- **PREMIS:** international, *de facto* standard
  - Widely-used
  - Implementation-neutral; applicable in many contexts
  - Comprehensive
  - Focused on practical application
- **SPOT:** property-oriented threat model
  - Properties of successful preservation defined in SPOT reflect community consensus (e.g., Waters & Garrett, OAIS, PREMIS)
  - Conceptual clarity, appropriate detail and consistent granularity, comprehensiveness, simplicity
  - Focused on digital preservation (technical)
  - Focused on practical application

Finding the right level of abstraction



Mapping

- Map *properties of entities relevant to digital preservation process* (PREMIS) to *threats to properties of successful digital preservation* (SPOT)
- Mapping not necessarily one-to-one in either direction:
  - A property of an Entity (semantic unit) can map to multiple threats across different properties of successful preservation
  - A threat to a property of successful preservation can map to multiple semantic units across multiple Entities
- “Bottom-up” vs. “top-down” approach:
  - B-U: given a set of metadata, what threats can we assess?
  - T-D: given a set of threats, what metadata is needed to assess them?
  - Both approaches useful

Example: Media degradation

SPOT:

Persistence → Useful life of storage medium is exceeded (mean time to failure approaching zero)

PREMIS:

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

(i.e., last media refreshment occurred roughly 14 years ago)

Suppose average shelf-life of tape is 15 years:

- Automated analysis of metadata can flag this threat for attention

Example: Significant properties

METS package (MS Word doc & metadata) arrives for ingest via inter-repository transfer

- Standard procedure: normalize Word to plain ascii text

SPOT:

Renderability → Object characteristics important to stakeholders are incorrectly identified and therefore not preserved

PREMIS:

significantProperties

significantPropertiesType = `behavior`

significantPropertiesValue = `hyperlinksTraversable`

Example: Inhibitors

A representation (2 files & metadata) submitted for ingest

PREMIS:

objectIdentifierValue: `...\...\file1`

inhibitors

inhibitorType = `PGP`

inhibitorTarget = `Content`

SPOT:

Availability → Only part of the digital object is available for preservation; the rest has deteriorated, was not selected, or is otherwise unavailable for preservation.

Missing metadata? inhibitorKey = [PGP private key for decryption]

Comments on examples

- Gap analysis: not necessarily the case that repository actually records necessary metadata
  - Identify “core” preservation metadata needed to support risk assessment …
  - … which can be compared against metadata actually collected
  - Similar in concept to PREMIS itself
- Repository policies/mission essential for establishing context for risk assessment (e.g., significantProperties)
- Two categories of “threat detection”:
  - Threats that have already manifested themselves: e.g., check-sums
  - Threats that potentially can manifest: e.g., media refreshment

Concluding thoughts

- Metadata analysis for risk assessment already done at many repositories
  - So our concept is not new …
  - … but we are taking it one step further to see whether such analysis can be **standardized** into a widely-applicable protocol

- Benefits of project can flow in two directions:
  - Bottom-up approach (consider potential threats in light of available metadata)
    - Highlight gaps in threat model
  - Top-down approach (consider metadata in light of enumerated threats)
    - Highlight gaps in Data Dictionary