Preservation Health Check: Monitoring Threats to Digital Repository Content

#phcpilot

Wouter Kool
Metadata Specialist

Titia van der Werf
Senior Program Officer

Brian Lavoie
Research Scientist

April 2014

OCLC Research
Dublin, Ohio 43017 USA
www.oclc.org

ISBN: 1-55653-472-8 (978-1-55653-472-0)
OCLC Control Number: 876081156

Please direct correspondence to:
Titia van der Werf
Senior Program Officer
Titia.vanderwerf@oclc.org

Suggested citation:
Kool, Wouter, Titia van der Werf and Brian Lavoie. 2014. *Preservation Health Check: Monitoring Threats to Digital Repository Content*. Dublin, Ohio: OCLC Research.
http://www.oclc.org/content/dam/research/publications/library/2014/oclcresearch-preservation-health-check-2014-a4.pdf.

#phcpilot

Contents

Acknowledgments 4
Introduction 5
Preservation Metadata: Costs and Benefits 5
Preservation Metadata for Threat Assessment 6
Methodology 7
Some Notes about the Mapping Exercise 10
Results of the Mapping Exercise 11
Next Steps 18
Notes 19
References 20

Acknowledgments

We'd like to thank our partners, the Open Planets Foundation and the Bibliothèque Nationale de France for their participation in and support of this work. We would especially like to thank Ed Fay (Open Planets Foundation) and Sébastien Peyrard (Bibliothèque Nationale de France) for reviewing an earlier draft of this paper.

Introduction

The Open Planets Foundation (OPF) has suggested the need for digital preservation repositories to perform periodic "health checks" as a routine part of their preservation activities. In the same way that doctors monitor basic health properties of their patients to spot indications of infirmity, repositories should monitor a set of properties associated with "preservation health" to provide an early warning of potential threats to the ongoing security of the archived digital objects in their care. The Preservation Health Check (PHC) project, undertaken as a joint effort by OPF and OCLC Research, aims to evaluate the usefulness of the preservation metadata created and maintained by operational repositories for assessing basic preservation properties.

The PHC project seeks to develop an implementable logic to support preservation health checks of this kind, and to test this logic against the store of preservation metadata maintained by an operational preservation repository. The Bibliothèque Nationale de France has agreed to share their preservation metadata in support of this project. Our goal is to advance the use of preservation metadata as an evidence base for conducting preservation health checks according to a standardized, widely-applicable protocol. Doing so opens up possibilities for internal or third-party threat assessment services that can be used for internal repository planning and auditing/certification. From a broader perspective, we hope that the PHC project will also provide evidence of the benefits well-maintained preservation metadata can confer on the day-to-day planning and operations of digital repositories.

The remainder of this paper provides background on the problem addressed by the PHC project, our approach for operationalizing the concept of a preservation health check, some preliminary findings, and next steps.

Preservation Metadata: Costs and Benefits

The PREMIS Data Dictionary is the de facto international standard for preservation metadata. The Data Dictionary (v2.2) is 272 pages long (including supplementary materials), and defines 195 semantic units (containers and components) that constitute the information most preservation repositories need to know in some form to support the long-term preservation of digital materials (PREMIS 2012). Of course, not all semantic units will be necessary or relevant to all repository contexts, nor does PREMIS conformance require full implementation of the Dictionary. Nevertheless, the size and scope of the PREMIS Data Dictionary suggests that a reasonably comprehensive implementation requires significant effort.

The serious resource commitment needed to create and maintain preservation metadata begs the question: what are the benefits received by repositories in return for this investment? The PREMIS Data Dictionary defines preservation metadata as "*the information a repository uses to support the digital preservation process*. Specifically, . . . the functions of maintaining viability, renderability, understandability, authenticity, and identity in a preservation context" (PREMIS 2012, 3). This suggests that the benefits from preservation metadata are, broadly speaking, associated with their ability to *support the maintenance of a set of preservation properties*. In more concrete terms, preservation metadata accomplishes this by documenting the key features of an archived digital object, as well as its provenance, the process by which it is preserved, and other information relevant to its long-term stewardship. In this way, preservation metadata helps a digital object become "self-documenting"— a feature that is especially important if the digital object becomes separated from its original curatorial context—e.g., transferred into the custody of a new steward in a new repository setting, or in the extreme, passing out of archival retention entirely.

Preservation metadata is often viewed as "contingency information" that is stored and maintained in the event of future need. Such information certainly imparts an important benefit to preservation repositories, since long-term digital stewardship is a process fraught with uncertainty, and preservation metadata helps guard against some of the potential consequences of that uncertainty. However, the benefits from implementing preservation metadata also extend to the support of near-term, "day-to-day" operations of the repository, as the use case addressed by the PHC project suggests.

Preservation Metadata for Threat Assessment

Our goal in formulating the PHC project was to explore a use case where preservation metadata could be used to support a particular aspect of digital preservation repository operations. It is important to emphasize that the goal was not to define new forms of preservation metadata that could be used for this purpose; instead, it was to find new uses for information already defined in the existing standard for preservation metadata (the PREMIS Data Dictionary)[1]. We recognize, and our preliminary findings confirm this, that a threat assessment would likely involve information that extends beyond what is currently defined in the PREMIS Data Dictionary, or even beyond what falls within the scope of preservation metadata. Issues also arise as to whether the information needed for assessment is under the direct control of the repository itself, or whether it is created and maintained by parties external to the repository. However, threat assessment is the responsibility of a repository, as is obtaining sufficient control of the information required to perform such assessments. Our starting point is that the PHC should be able to obtain the necessary information from the repository and that a core part of this information consists of preservation metadata.

We chose *preservation monitoring based on threat assessment* as our use case. By preservation monitoring, we mean identifying and tracking changes impacting a set of key properties of digital preservation. A useful framework for understanding the type of preservation monitoring envisioned by the PHC project is a process involving sensors, thresholds, and triggers. Metadata records the results of the latest measured values (sensors); policies provide guidance of critical thresholds (e.g., frequency of media refreshment; frequency of checking bit integrity, etc.); events are preservation actions triggered by matching sensor information and thresholds. Events can lead to recording new metadata values and resetting thresholds.

Threat assessment involves the identification of certain contingencies that can potentially interfere with a repository's ability to achieve its goals, and evaluating the likelihood or imminence of these threats in regard to the repository's current operations. Threat assessments should be undertaken periodically as a regular feature of repository management practice. In this sense, they are part of the operational workflow of the repository. Our goal was to determine if the preservation metadata recorded and maintained by the repository could serve as a useful evidence base to support this workflow.

Methodology

Our goal is to explore the opportunities for using preservation metadata to support threat assessment exercises. We are not proposing a complete implementation suitable for immediate application, but rather, testing a hypothesis that preservation metadata can indeed be used to support the day-to-day operational needs of digital repositories. We expect that the analysis will reveal a variety of gaps in current preservation metadata coverage, which might be filled by other metadata schema used in conjunction with PREMIS, or with new semantic units defined within the PREMIS Data Dictionary itself. Given that our analysis is exploratory, we are confining its scope to PREMIS alone, which is, in a sense, the least common denominator of many preservation metadata implementations. We do not extend our analysis to include format-specific technical metadata schema (such as MIX), or packaging schema (such as METS), since use of these schema will depend on specific implementation decisions, and will vary from repository to repository. The scope of our analysis is confined to establishing a general approach for using preservation metadata to support threat assessment, and to evaluating the utility of PREMIS preservation metadata as an evidence base for such assessments.

We chose the following methodology to address this question. In the first phase of the project, a mapping was constructed between the information defined in the extant preservation metadata standard (i.e., the PREMIS Data Dictionary) and an enumeration of significant

threats that could potentially manifest in the day-to-day operations of a "typical" digital preservation repository. Given this mapping, the second phase of the project constructs examples of how the PREMIS semantic units, mapped to a particular threat, could be organized into an automated assessment workflow that yielded an actionable result to repository managers, in the form of a binary indicator that signaled whether the threat was imminent or not.

In order to carry out this plan, a threat model for digital preservation was needed. We chose the Simple Property-Oriented Threat (SPOT) Model for this purpose (Vermaaten, Lavoie and Caplan 2012). The SPOT Model enumerates sets of threats associated with six properties of successful digital preservation (availability, identity, persistence, renderability, understandability, and authenticity). In addition to providing an organized context for the threat aspect of the preservation-metadata-to-threats mapping, the SPOT Model has a number of other features that recommended it for our purposes: the properties of successful preservation used to organize the SPOT Model are based on a wide-community consensus; the model shares with PREMIS (and preservation metadata generally) a focus on the preservation aspect of digital lifecycle management, and in particular, the technical aspects relating to the ingest, storage, maintenance, and dissemination of archived content. The model is relatively light-weight and designed for ease of practical use. It explicitly does not address threats to the custodial organization itself, such as the economic and legal aspects of a repository. The scope of the SPOT Model corresponds to the scope of our preservation monitoring use case.

In the first phase of the project, the semantic units defined in the PREMIS Data Dictionary were mapped to the high-level threats defined in the SPOT Model, as illustrated in figure 1:

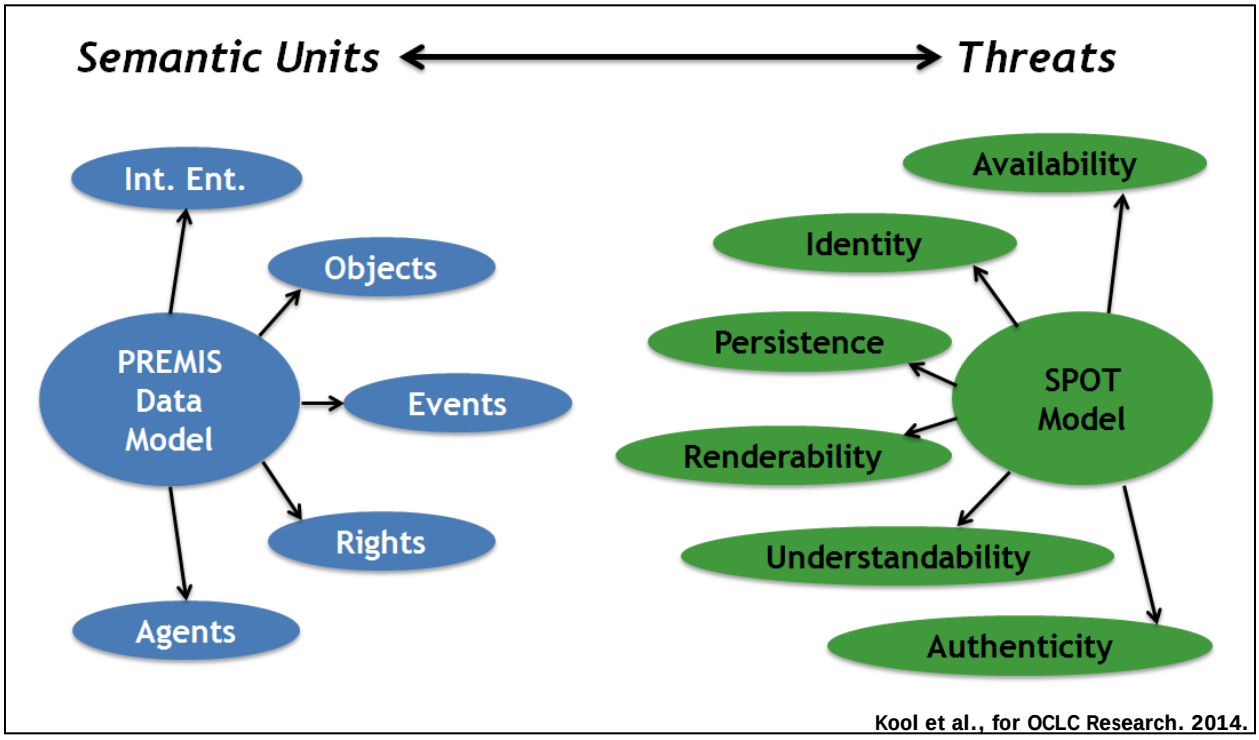

Figure 1: Mapping PREMIS semantic units to SPOT properties

In the PREMIS Data Dictionary, semantic units are organized into a five-part data model, including Intellectual Entities, Objects, Events, Rights, and Agents. Each PREMIS semantic unit may be understood as a property of one of these data model entities—for example, semantic units associated with the Object entity are properties of archived digital objects. Similarly, the SPOT Model is divided into six properties of successful digital preservation: availability, identity, persistence, renderability, understandability, and authenticity. Each property is associated with a set of high-level threats which negatively impact the ability to achieve the property. The mapping between PREMIS and SPOT, therefore, is a mapping between the properties of entities relevant to the digital preservation process, and threats to the properties of successful digital preservation. Depending on the nature of the SPOT property in question, the mapping will potentially include semantic units related to some or all of the entities defined in the PREMIS data model: Objects, Events, Rights, and Agents.[2]

The PREMIS Data Dictionary was designed to provide a set of core preservation metadata—that is, preservation metadata that would likely be applicable to most repository contexts. As such, it does not directly encompass information like format-specific technical metadata, but instead is extensible to include the use of complementary metadata schema such as MIX or textMD whose usage will likely vary from repository to repository. Moreover, information

defined in PREMIS may be duplicated in other standards used in conjunction with PREMIS, such as METS; repositories may choose to implement this information outside of PREMIS. In cases where information relevant for threat assessment falls outside the scope of PREMIS-based metadata, it will need to be gathered from other sources such as complementary metadata or packaging schema, or registries.

It is useful to place this approach in a larger context. The PREMIS Data Dictionary essentially represents an evidence base of information that potentially is of use in making assessments of the immanency of threats to the archived content in a digital repository. The SPOT Model serves as a framework for organizing and assessing the evidence supplied by PREMIS-based preservation metadata. Finally, a risk model is a means of interpreting and, ultimately acting on, the intelligence provided by the PREMIS/SPOT assessment. The risk model should incorporate the repository's policies, practices, and tolerance relating to risk, and help determine whether action is required. The analysis in this paper focuses on the first two elements of this three-part process: the evidence base (PREMIS), and the organizing threat framework (SPOT). We feel that these two elements are, by and large, generalizable over many repository contexts. The third element (risk modeling) will tend to be repository-specific, and is beyond the scope of this study.

Some Notes about the Mapping Exercise

The mapping was done at the highest level for the PREMIS semantic units—in other words, at the container level. Individual semantic components were not evaluated at this stage of the project; the idea was that if the information scoped within the container is relevant to a particular threat, the semantic components, by inheritance, will generally also be relevant. The exception to this was the objectCharacteristics semantic unit. Given the variety of properties that are gathered under this semantic unit, as well as the different levels of granularity at which they apply (i.e., file-level, bit- level, etc.), it would not be useful to map it as a single entity to the SPOT threats.

The PREMIS data model entities Events, Agents, and Rights, and their respective semantic units, were evaluated monolithically, rather than splitting them up into their component semantic units. By and large, if any of these were relevant to a particular threat, all of their semantic units would potentially be relevant in describing the Event, Agent or Right in question.

From the perspective of the SPOT Model, the mappings were done at the level of the six properties of successful preservation, each of which subsumes a list of specific threats. In other words, a PREMIS semantic unit was mapped to a SPOT property if it was deemed relevant to one or more of the high-level threats associated with that property.

A fairly relaxed standard was used for determining a semantic unit's relevance to a SPOT property. Even if the association seemed very indirect, it was included, with the idea that it can always be dropped later. It is worth emphasizing that the mapping is not necessarily one-to-one in either direction. A property of a PREMIS entity—i.e., a semantic unit—can potentially map to multiple threats across different properties of successful preservation. Conversely, a SPOT property can map to multiple PREMIS semantic units distributed across multiple PREMIS entities.

Results of the Mapping Exercise

Several interesting findings emerged from the PREMIS-to-SPOT mapping exercise.

Coverage

A key issue is how much evidence PREMIS provides relevant to threats associated with the SPOT preservation properties. The mapping exercise confirmed that some portion of the PREMIS Data Dictionary is relevant to each of the six SPOT properties (table 1). In other words, an evidence base consisting of PREMIS metadata would appear to offer useful evaluative information for threats associated with each of the six preservation properties addressed in the SPOT Model.

Table 1. Number of relevant PREMIS semantic units per SPOT property

SPOT property	Number 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*

A large number of relevant semantic units does not necessarily mean that a SPOT property is "well-covered" within the PREMIS Data Dictionary; the more important criterion is that the semantic units directly address the threats associated with the property. For example, while the SPOT property of persistence has the fewest number of semantic units mapped to it (10), it still appears to be "well-covered" *vis-à-vis* other SPOT properties, because the information recorded in these semantic units speaks directly to the threats associated with the

persistence property. In contrast, the SPOT property of understandability has more semantic units mapped to it then the persistence property, but it is nevertheless probably less "well-covered": the semantic units mapped to this property, while relevant, tend to fall around the edges of the threats associated with understandability. For example, PREMIS does not explicitly define semantic units that provide information that aids in the understanding or interpretation of the content of the archived digital object (although the repository can establish a link to such information if it is recorded in another digital object).

Documenting policies

The mapping exercise indicated that one of the most significant gaps in the PREMIS "evidence base" in terms of its relevance to supporting threat assessment is its lack of coverage of policy-related issues, which will likely be essential in conducting preservation monitoring based on threat assessment. 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.

Autonomy of the repository

Another potential difficulty with using PREMIS as an evidence base for threat assessment is that 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". There are, however, good reasons for a repository to "outsource" parts of the digital preservation process to third parties. Situations where the digital preservation process is distributed over multiple organizations will likely introduce different perspectives on the threats associated with the SPOT properties of successful digital preservation, and place different demands on a supporting evidence base of preservation metadata.

An example of this issue arises in connection with threats to the SPOT property of identity. A repository is not necessarily one system, but instead can be an integration of several sub-systems (e.g., database, file-system, web server, etc.). An identifier generated automatically by a sub-system (e.g., the database) is local and usable in that sub-system and is important for system-dependent actions (retrieval, copy, delete, etc.) but not for digital preservation processes that span more than one sub-system, nor for third parties providing digital preservation services to the repository. Moreover, the identifier automatically generated by a repository sub-system is only as persistent as the sub-system itself. When it becomes a legacy system and digital assets are migrated to another sub-system, the identifier is no longer usable by the new sub-system—which will generate its own local system-specific identifiers.

So persistency of the identifier should not be an attribute of the local identifiers used by repository sub-systems, but should be an attribute of the system-independent objectIdentifier used through the repository as a whole. At the repository level it is therefore important to assign system-independent identifiers to objects.

Another example concerns the SPOT property of persistence. Bit preservation is mainly a storage function, which can be—and often is—out-sourced to an internal IT department or external party. In this sense, some preservation-metadata-like information relevant to the persistence property may not be collected and maintained directly by the digital repository service itself, but instead may be found in the various audit trails of IT processes that are themselves external to the repository service's processes (see discussion of threat assessment logic for persistence, beginning on page 14).

On the other side of the mapping, it is not clear that the SPOT Model adequately addresses threats associated with coordinating digital preservation processes over multiple sources. It is possible that future versions of both PREMIS and SPOT will need to re-consider this issue.

Explicit encoding

Yet another practical challenge to constructing generalized threat assessment logic based on PREMIS preservation metadata is that PREMIS does not require explicit encoding of metadata; nor does PREMIS require that there be a one-to-one mapping between semantic units and metadata elements. This makes it difficult to construct logic that generalizes over many repository contexts, and suggests that significant local adaptation of "high-level logic" will have to be done to implement threat assessment workflows. It also impedes provision of a threat assessment service 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. One approach for mitigating this problem would be the requirement that PREMIS implementations be "externally conformant", which facilitates the exchange of PREMIS-based metadata between repositories or other parties. A detailed discussion of this issue is beyond the scope of this paper, but interested readers are referred to the PREMIS conformance statement for more information. (PREMIS 2010, 4-5)

Level of threat assessment

In considering a threat assessment exercise conducted through the expediency of a model like SPOT, one question which arises is the level at which the assessment is to be performed. Will it be done at the level of individual digital objects? Collections of objects? Repository-wide? The answer will likely depend on both the particular context of the repository involved in the assessment, as well as the nature of the specific threat (and preservation property) being

evaluated. For example, threats to authenticity would likely be evaluated at the level of individual objects; threats to renderability might be evaluated at the collection level, where collections are defined as classes of objects sharing the same hardware/software environments; threats to identity might in some circumstances be a repository-wide issue, if for example identifiers are created and maintained through a single internal repository system, or distributed through a shared external registry.

Because of the multiplicity of levels at which preservation monitoring and threat assessment can take place, it is important to clearly indicate in the supporting logic not only which PREMIS semantic units map to a particular SPOT-defined threat, but also at what level (or levels)—object, collection, repository—the assessment should be conducted. The SPOT Model does not explicitly specify this "granularity of analysis" for the properties and threats it covers.

Logic for threat assessment: persistence

To illustrate the type of threat assessment logic we envision for a preservation health check, we present an example based on the SPOT property of persistence. As mentioned above, this property is relatively "well-covered" by the PREMIS Data Dictionary, and therefore serves as a useful exemplar of the use of preservation metadata as a means of supporting automated threat assessment.

According to the SPOT Model, the property of persistence is defined as follows:

> "Persistence is the property that the bit sequences comprising a digital object continue to exist in a usable/processable state, and are retrievable/processable from the medium on which they are stored. All digital objects exist as a series of bits stored on some form of physical medium, such as magnetic tapes, optical discs such as CDs and DVDs, or hard drives on servers or personal computers. In order for the digital object to remain useful over time, it is essential that the bit sequences are not corrupted in any way, and that they can be read in their entirety from the physical media on which they are stored. Persistence is achieved when these two conditions are met." (Vermaaten, Lavoie and Caplan. 2012)

Figure 2 presents a schematic representation of the threat assessment logic for persistence. It is primarily based on the evidence provided by the PREMIS preservation metadata recorded in a repository. It is designed as a high-level flow diagram, identifying the PREMIS semantic units at their highest level of relevance only, not going into the details of the underlying sub-units and corresponding flow steps. It consists of a generic decision-flow, applicable to any object in the repository, and it does not imply any specific implementation for executing the threat assessment.

The automated assessment flow yields an actionable result to repository managers, in the form of indicators (-1; 0; 1) that inform about the likelihood of a threat associated with that property. We distinguish between 3 indicators: (-1) signals that relevant metadata is missing; (0) signals one or more threats have been detected; (1) signals that no threat has been detected. If no indicator is specified, the flow stops without yielding an actionable result—for example when the Object under scrutiny is neither a File nor a Bitstream.

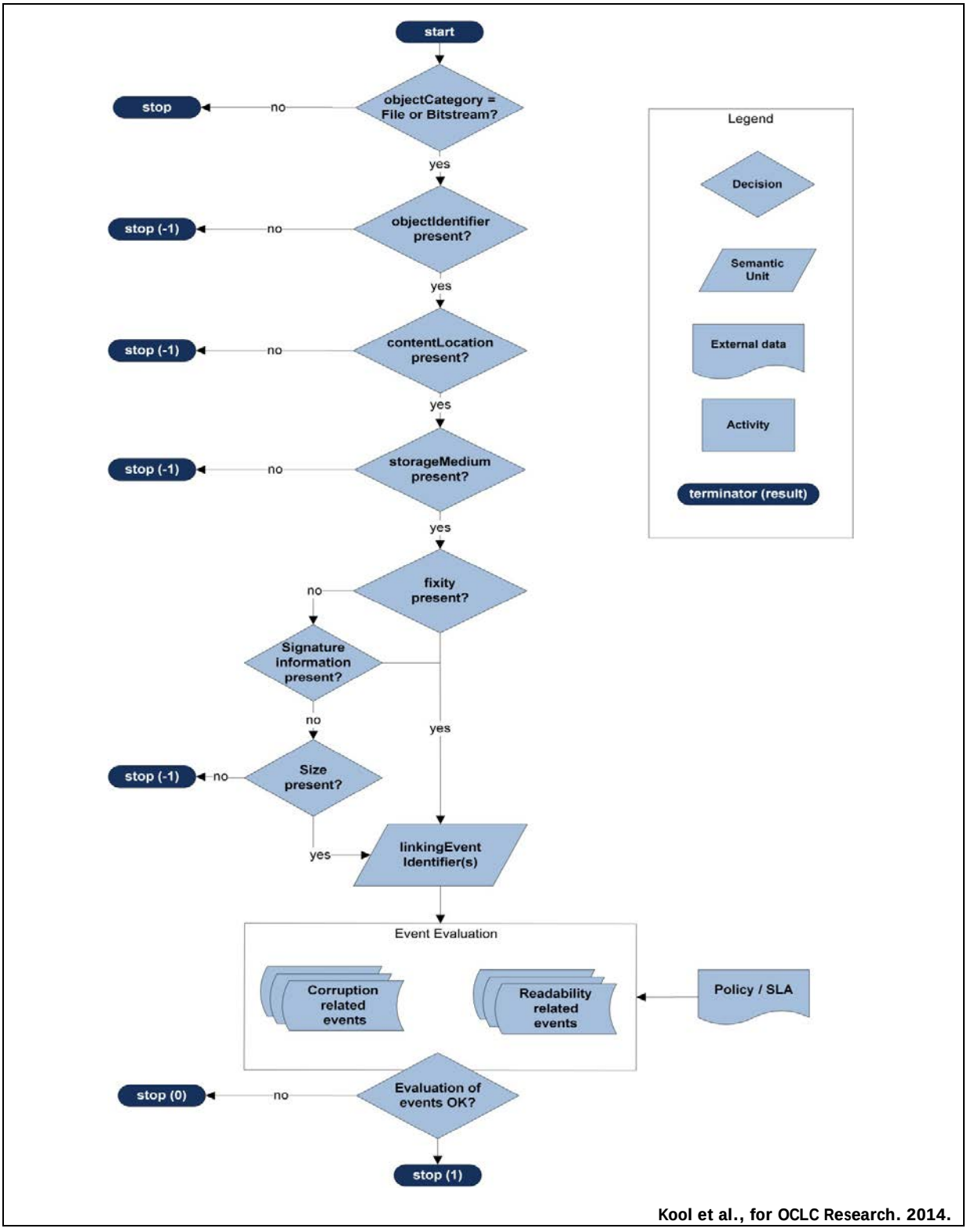

Figure 2. Preservation Health Check diagram for Persistence

Logic for threat assessment: persistence (continued)

The Object under scrutiny in this specific diagram should be a File or a Bitstream. Other object categories (Representations) or higher-level entities (e.g. Intellectual entity) are not analyzed for Persistence, because they represent mental/logical constructs, not real/physical instances. We argue that risks to the Persistence of digital objects occur at the physical (bit) level, not at the logical level. We recognize the fact that if a part of the logical whole is corrupt, the whole is corrupted as well. More specifically, because Representations are made up of files, if a file runs a particular Persistence risk, then the whole Representation runs that same risk. The relationships between the parts and the whole can be inferred by following the PREMIS structural relationships. This exercise however, is not considered core to the risk analysis flow and is therefore not included in the diagram.

The PREMIS semantic units that we have mapped as "relevant" for the SPOT property of Persistence are included in the decision flow: objectCategory, objectIdentifier, contentLocation, storageMedium, fixity/size/digitalSignature, linkingEventIdentifiers. These units are checked. This could include a variety of checks (is the unit present in the metadata? is there a corresponding value? is the value correctly encoded?). The level of detail of these checks will depend on the particular context and specific practices of the repository. A PHC could only check the presence of the semantic units in the metadata files and carry on without any further inspection, assuming the units contain valid values—or a PHC could go as far as validating the values, going into the nitty-gritty details of validation, error handling, etc. More importantly however, we note that the semantic units identified in the flow-diagram are a baseline for the event evaluation further on in the assessment flow.

There are two categories of semantic units that are core to Persistence: those relating to storage and those relating to objectCharacteristics:

- Storage: According to the PREMIS usage notes, the semantic unit storage is mandatory, but both of its subunits are optional. For the PHC it is arguably not necessary to have access to the storage content location, because the check is not carried out on the actual objects (files or bitstreams). However, digital storage media vary in how long they last and require active management, including regular migration of the content from older storage media to newer ones to avoid data loss. Therefore, knowing the specific storage medium is necessary to be able to make threat assessments in relation to Persistence. In addition, for generating automatic threat assessments, the real values or references that lead to the real values are necessary as well. If this information is not available in PREMIS metadata, the PHC will need to take other information sources into account—such as audit reports generated by storage management systems.[3]
- objectCharacteristics: The semantic unit objectCharacteristics consists of a very diverse set of sub-units. For Persistence, only fixity is relevant. However, we considered that if fixity was absent, we should allow for alternative ways to detect corruption at the bit-level. We therefore included the semantic unit digitalSignature, which is a good alternative to fixity, and the sub-unit "size", which is far less precise.

In principle, threat assessments cannot be made on the basis of the values of the above semantic units per se, but the information will be of use in the next step of the logic, which is to evaluate the metadata about the events that the repository has carried out to prevent or detect threats to Persistence.

The SPOT definition of persistence highlights two important conditions for achieving this property:

> "…it is essential that the bit sequences **are not corrupted in any way**, and that they can be **read in their entirety** from the physical media on which they are stored." [emphasis added] (Vermaaten, Lavoie and Caplan 2012)

We therefore define two categories of events relevant to assessing persistence: 1) Corruption-related events and 2) Readability-related events. These are broad classes of events. The first category corresponds to events that detect or prevent/mitigate corruption at the file or bit-level (involving Fixity Checks, Virus Checks, Backup & Restore procedures, etc.). The second category corresponds to events that detect risks or prevent risks at the storage medium level (involving medium refreshment and other processes such as replication, etc.). These two classes of events probably have blurred delimitations and might in some cases overlap (e.g. Backup event). The differences between detection events and mitigation events might even be sufficiently distinctive to warrant differentiation.

We note in this context that there are no pre-defined events in PREMIS[4], which means that the repositories need to define their own events. For PHC purposes we can expect a vast array of different yet similar events (e.g. backup and replication). In addition, the same type of event can be implemented in different ways. PREMIS is very open and flexible when it comes to defining preservation events,[5] but for preservation threat assessment applications like the PHC, which rely on automated data evaluation, a certain level of predictability and harmonization is necessary. The practical way forward would be to establish some form of interoperability with IT storage practices and routines for "data integrity and availability".

Having said that, the repository is responsible for the bit-level preservation, and the role of the IT service provider—in-house or outsourced—should be one of implementation. In other words, the bit-level preservation or Persistence policies need to be defined by the repository and the Events need to be informed by the Policies. The repository should therefore have policies in place that prescribe frequencies of fixity checks, of medium refreshment, backup policy, etc. The PREMIS semantic unit preservationLevel does not address such policies. The PHC flow thus needs to get the policy information from other sources.

Finally, there is another category of events relevant to Persistence: security measures. Data integrity, confidentiality and availability can be compromised by computer crime and malpractice, such as failing to keep a password secure, hacking, not following procedures, etc. We consider security as out-of-scope for Persistence, because these measures and policies affect the repository as a whole and do not speak to specific PREMIS units.

Next Steps

Our preliminary findings suggest that there is indeed opportunity to use PREMIS preservation metadata as an evidence base to support a threat assessment exercise based on the SPOT Model. In Phase 2 of the project, we will explore this issue further by constructing additional generalized logic sequences/diagrams that demonstrate how PREMIS metadata can be used to assess the threats defined in the SPOT Model. Once these examples are constructed, we will test their efficacy on a data set of "real-world" preservation metadata. We would like to acknowledge and thank our partner in this effort, the Bibliothèque Nationale de France, which has generously provided us with a sample of preservation metadata from their SPAR digital repository system. We will use this metadata to validate the practical utility of our evidence-based threat assessment logic in light of the actual preservation metadata a particular digital repository records and maintains. As part of this exercise, we will continue to evaluate the extent and scope of the gaps we discover between the preservation metadata needed for threat assessment, and the information defined in the current version[6] of the PREMIS Data Dictionary.

Notes

1. We do plan to make recommendations regarding perceived gaps in the PREMIS Data Dictionary where appropriate.
2. Note that Intellectual Entities, while included in the PREMIS data model for completeness, do not actually have semantic units (properties) assigned to them in the PREMIS Data Dictionary. See discussion of descriptive metadata in PREMIS Data Dictionary 2.2 (2012, 23-24), for more information on this point.
3. If such is the case, we note that the PREMIS assumption "that the repository ultimately is in control" of the preservation status of its digital objects is not quite the reality, if this control is based on PREMIS metadata only.
   > "In some cases this can be masked from direct repository management by storage management systems but the underlying assumption is that the repository ultimately is in control and needs to manage for technological obsolescence.
   > In some cases the value may not be the specific medium, but the system that knows the medium, e.g., Tivoli Storage Manager (TSM).
   > Knowing the storage medium is an internal requirement in order to trigger preservation actions. However, since this is not information that is used for exchange purposes, it is optional." (PREMIS 2012; usage notes p. 79)
4. PREMIS does provide a list of recommended event labels for the semantic unit eventType (PREMIS 2012; usage notes p. 134-135), but it is just a "suggested starter list."
5. The PREMIS Data Dictionary 2.2 (2012), states for example: "Whether or not a preservation repository records an Event depends upon the importance of the event. Actions that modify objects should always be recorded. Other actions such as copying an object for backup purposes may be recorded in system logs or an audit trail but not necessarily in an Event entity." (p. 130) and further on p. 135: "In general, the level of specificity in recording the type of event . . . is implementation specific and will depend upon how reporting and processing is done."
6. The changes for version 3.0 have recently been published on the PREMIS website. We have based our work on v2.2. In a next stage, we will look into the changes for v3.0 to assess in how far they affect our preliminary conclusions.

References

PREMIS (Editorial Committee). 2010. *Conformant Implementation of the PREMIS Data Dictionary.* October. Washington DC: Library of Congress. http://www.loc.gov/standards/premis/premis-conformance-oct2010.pdf.

PREMIS (Editorial Committee). 2012. *Data Dictionary for Preservation Metadata: PREMIS version 2.2.* Washington DC: Library of Congress. http://www.loc.gov/standards/premis/v2/premis-2-2.pdf.

Vermaaten, Sally, Brian Lavoie and Priscilla Caplan. 2012. "Identifying Threats to Successful Digital Preservation: the SPOT Model for Risk Assessment" *D-Lib Magazine* 18 (9/10). http://www.dlib.org/dlib/september12/vermaaten/09vermaaten.html.