Preservation Metadata Dictionary 2.0

Versie 0.3

Introduction

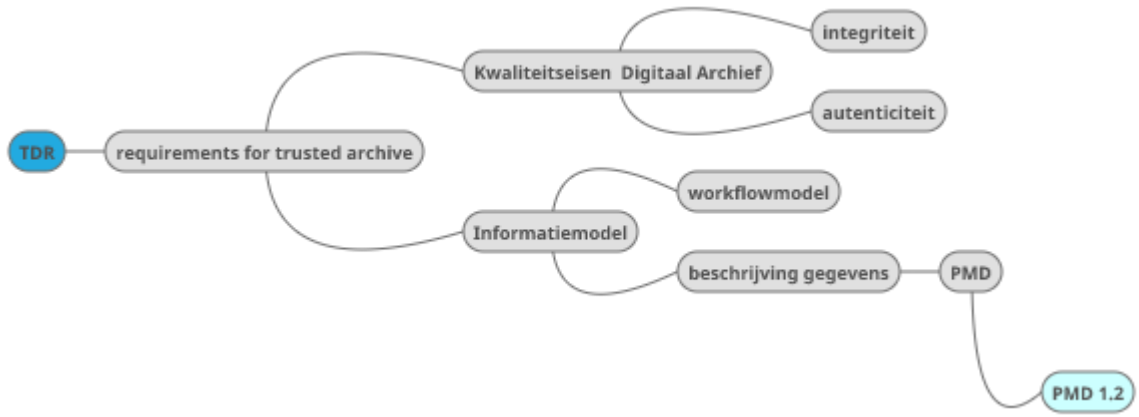
The goal of the Preservation Metadata Dictionary (PMD) is to describe the attributes of digital objects in the Digital Archive as well as how said objects are to be processed and managed[1].

This document describes version 2.0 of the PMD which replaces PMD version 1.2. A beta version of PMD version 2.0 was released at the end of 2016. Since then, it has been tested in practice and in some areas, improved.

Background

In 2016 Sound and Vision received the "Trustworthy Digital Repository" (TDR) seal. During the run-up to certification, Sound and Vision's TDR requirements[2] were defined in the set "Kwaliteitseisen Digitaal Archief" (Digital Archive Quality requirements) and in an Information Model. The information model has two components: a *workflow description* for ingest, storage and accessibility of digital files and metadata, and a *description of the data* that during the workflow is administered about the file.

Schematisch:



The two central quality requirements of a 'trusted repository': *integrity and authenticity* are illustrated in this schema. In other words: the certainty that a file is not corrupt, and the certainty that a file was not unintentionally changed and is what it proports to be. With help from the information model Sound and Vision documents how integrity and authenticity is

[1] Zie PMD V2.0 Beleidsuitgangspunten, A.de Jong, mei 2016

[2] Preservering van digitale AV-collecties volgens de OAIS standaard, Requirements voor een 'trusted' archief; Annemieke de Jong, Beth Delaney en Daniel Steinmeier, mei 2014

ensured (workflow) and how it can be demonstrated (data)[3]. With the Preservation Metadata Dictionary this last goal has been worked out in detail.

The first version of the PMD (version 1.2) was composed of three components:

- A list with attributes for each of the four PREMIS model entities (Object, Event, Agent, Rights), based on PREMIS 2.0
- Additional technical metadata originating from other standards
- A set of *events* that Sound and Vision defined based on the workflows in the information model

This version of the PMD model formed the input for the PMD GAP analysis project in 2016 and 2017. During this project, the most important file formats being ingested at the time were checked against the PMD.
The project outcome not only provided insight into gaps related to preservation metadata but also resulted in an upgrade of the PMD to version 2.0.

Design conform PREMIS

The PMD has been developed to conform to PREMIS[4]. PREMIS stands for Preservation Metadata Implementation Strategy. The PREMIS Data Dictionary for Preservation Metadata is an international metadata standard designed to support the preservation and long-term sustainability of digital objects. The dictionary contains different sorts of metadata: administrative (including rights), technical metadata, and structural metadata.

> Sound and Vision workflows are based on international standards. An important advantage of working with international standards is that they provide a reference framework for third parties. Sound and Vision can always explain how the PMD relates to a general, international standard. PREMIS is the standard upon which Sound and Vision bases its preservation metadata definitions.

In the first version of the PMD, primarily ‘digital provenance’ (source data) and rights metadata were based on the PREMIS-model. The current version of the PMD is entirely based on PREMIS.

The PREMIS dictionary is built around four entities: They are:

- **Objects** the subject of digital preservation. It focuses on four types of digital content units
  - A bitstream
  - A file
  - A combination of files, needed to playback a program
  - An abstract description of a program (the photos, a fragment, etc)
- **Rights** describe the *acts* that Sound and Vision are allowed to carry out upon the objects. PREMIS focuses on acts from an archive management viewpoint.

[3] Informatiemodel Digitaal Archief Beeld en Geluid. Annemieke de Jong, Beth Delaney, Daniël Steinmeier, Yvette Hollander, Pol Hoffman. Netherlands Institute for Sound and Vision, 2013
[4] PREMIS Data Dictionary for Preservation Metadata, version 3.0, June 2015

- The **Events** are *activities* that have been carried out on a digital object while under archive management.
- An **Agent** is an external help entity. For example, for Rights, it would be the party with which the archive makes rights agreements, for Events it might be the software program that carries out a particular activity on the object.

The figure below illustrates how the entities relate to each other.

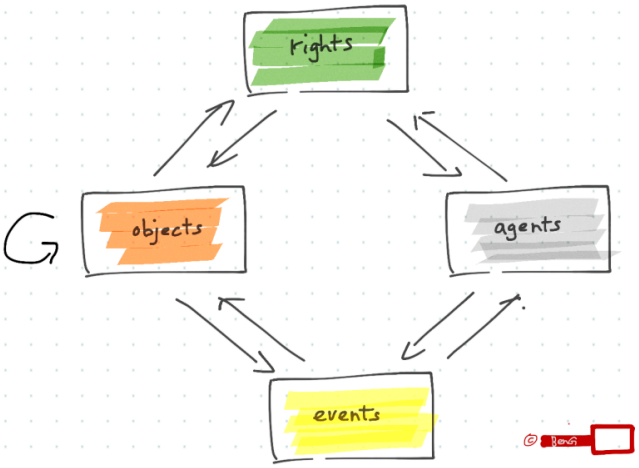


Figure: entity scheme PREMIS 3.0

In the figure it is clear, among other things:

- that an Object can have a relation with Rights and/or Events
- that Object entities can point to one another
- that Rights as well as Events can both have Agents

Objects

Objects are the subject of digital preservation. Digital content is analyzed in four ways. This adds an added dimension to the model. Each category has its own set of attributes and relationships. This means that the figure above can appear four different ways, each time demonstrating different relationships between entities[5].

Sound and Vision has made a few design decisions in its application and has not implemented all the relationships. These decisions are inherent in the way the categories have been interpreted and applied in practice. This interpretation is described below.

Bitstream

- The Bitstream is the data in the body of the file, whether the file content is streaming or still. The bitstream and the file are inextricably linked.

File

- File refers to all digital content; containers of digital data, primarily recognized by their extension, with their own specific technical metadata attributes. These files are managed

[5] See also the Conceptual view between object categories, PREMIS 3.0, blz 9, figure 2, for all possible relationships between the *categories* among and within themselves

in the storage management system (DIVA among others). *Almost all* preservation events are carried out on this level.
- files do not refer to each other, although there may be a relationship indirectly. When speaking of a derived file (a copy, a transcoded version, etc) there is an indirect relationship via an event. The event involves an input file (mother) and an output file (daughter). There can also be an indirect relationship with a Representation, namely, between the main file and one or more sub-files (for example, the STL that is part of an MXF). Main files and subfiles come together in packages. During import the entire package is analyzed; this is an event carried out at the Representation level.

Representation

- Often multiple files are needed in order to playback meaningful archive material. Together these files form a Representation. For example, a DPX in combination with its related WAV. Or 1 program made out of two successive WAV files.

Intellectual Entity

- The descriptive metadata concerning the content of the material forms the Intellectual Entity. Think of radio- or TV programs, a film, a photo, an audio recording. The Intellectual Entity can consist of one or more Representations. For example, a high resolution copy (archive copy) or a low resolution version (access copy).

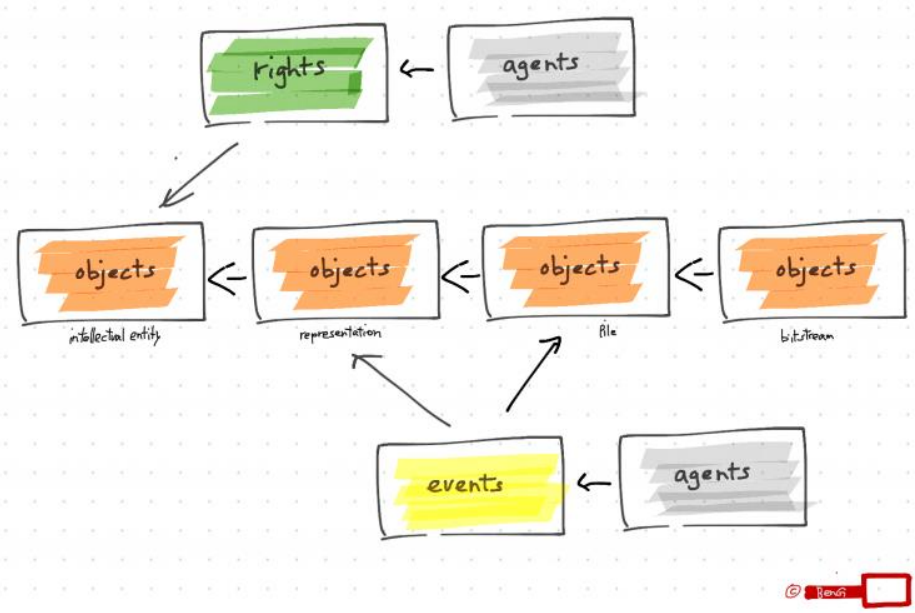
Connection with the MAM metadata model.
Sound and Vision acquired a new MAM system in 2018. The MAM system recognizes different entities that do not always correspond to the categories in PREMIS.

- Program: description of the content; comparable with Intellectual Entity
- Item: technical metadata of the master file
- File: attributes based on header information found in each file, relates to the master-file
- Package: a combination of items when multiple items are required in order to properly playback one program
- Logtrack item: information about a segment of a master-file, based on position information

Multiple files can be 'hidden' behind each *item*. For example, different resolutions, multiple backups etc. The MAM divides files into: Highres, Proxy, Auxiliary, Other (keyframes, indexes). In this case Highres and Proxy files are independent representations of the same program.

According to Sound and Vision's interpretation, relationships between events and rights do not always occur at every level. Even relationships between Objects on the same level do not always occur. Thus, at Sound in Vision, the events (preservation acts), in practice take place mostly at the file level. While at the same time, rights agreements (rights management) made at the intellectual level are applicable to all underlying files.

This leads, for now, to the following *logical data model* for the PMD:



Translation of the Data Dictionary in Sound and Vision practice

PREMIS-compliancy

Implementation according to level 1A

The PREMIS standard can be implemented in different ways. For example, by the way in which archive systems are based on PREMIS categories. Sound and Vision has chosen to comply with *Conformance level 1*, which means that the archive ensures that the PMD fields can be mapped to the preservation metadata as documented in its systems.

This Level 1 can be split into:

Level 1A: Object entity mapping
Level 1B: Object, Event and Agent entity mapping

The PMD 2.0 meets Level1A requirements (Mapping at the Object Entity Level). In the future, reaching level 1B is possible, after the MAM system is implemented and the recording of events is crystalized.

Further details Objects

PREMIS is generically designed in order to ensure broad applicability. This paragraph explains how Sound and Vision has developed its PMD within a variety of PREMIS model[6] generic starting points.

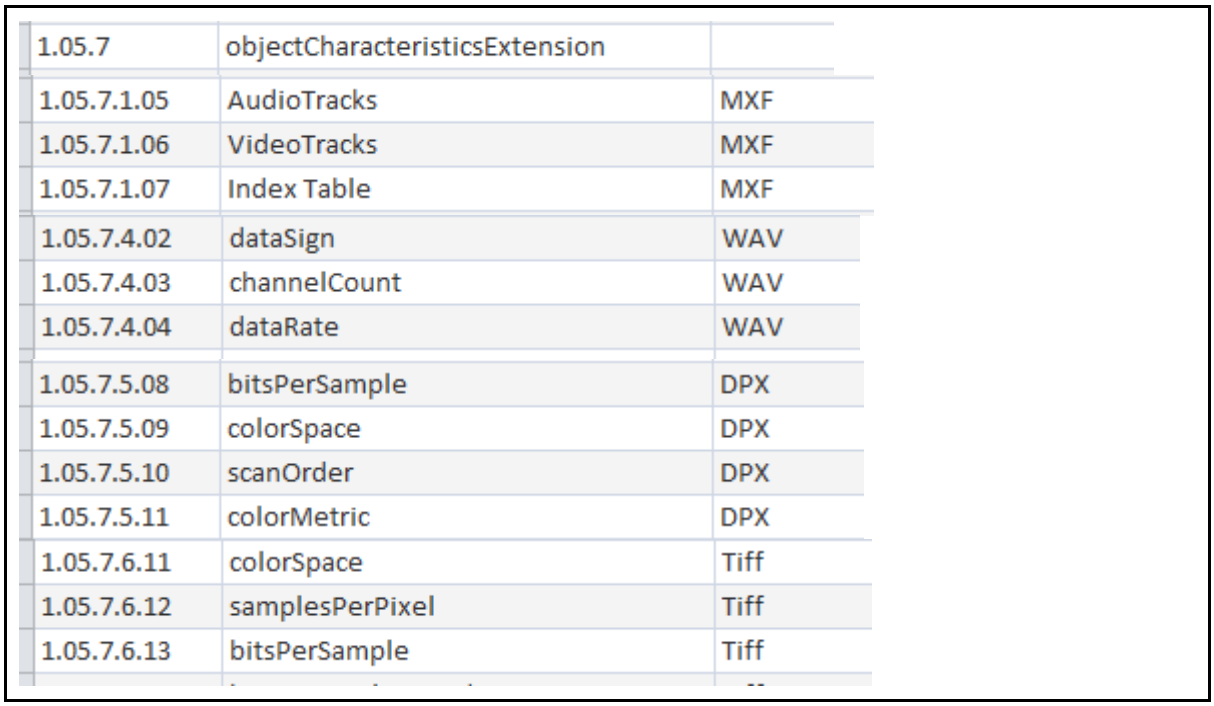
[6] Digital Preservation Metadata for Practitioners, Angela Dappert, Rebecca Squire Guenther, Sébastien Pyrard, Editors, chapter 3.6 PREMIS Goals and Principles, blz 27-30

1 extensibility

Technical attributes are outside the scope of PREMIS. However, in order to demonstrate the integrity and authenticity of an object, technical attributes are of great importance. Especially when there is talk of migration to another format. In such a case, the original checksum is no longer usable. With help from a few technical specifications it must be possible to document that a format migration was successful. Also, in the case when no checksum is delivered with the original object, technical attributes must be checked.

PMD 1.2 includes a list of technical attributes based on a variety of specific standards such as PBCore, EBUcore, AES, LC VideoMd, AudioMD and NARA's reVTMD. In the PMD 2.0 these attributes have been incorporated by using the *Extensibility* of the model.

The objectCharacteristicsExtension contains technical attributes. Sound and Vision has specified which attribute applies to which type of file.



1.05.7	objectCharacteristicsExtension	
1.05.7.1.05	AudioTracks	MXF
1.05.7.1.06	VideoTracks	MXF
1.05.7.1.07	Index Table	MXF
1.05.7.4.02	dataSign	WAV
1.05.7.4.03	channelCount	WAV
1.05.7.4.04	dataRate	WAV
1.05.7.5.08	bitsPerSample	DPX
1.05.7.5.09	colorSpace	DPX
1.05.7.5.10	scanOrder	DPX
1.05.7.5.11	colorMetric	DPX
1.05.7.6.11	colorSpace	Tiff
1.05.7.6.12	samplesPerPixel	Tiff
1.05.7.6.13	bitsPerSample	Tiff

A separate extension is available for attributes that are especially meaningful for audiovisual file performance. A limited number of attributes have been accommodated in this extension in the PMD.

Here again, some of the attributes are format specific, or are generally valid for more formats.

Premis_v3	Name	Attribute of
1.04	SignificantProperties	
1.04.3	significantPropertiesExtension	
1.04.3.02	use	
1.04.3.04	Frame Position	DPX
1.04.3.05	sequence length (frames)	DPX
1.04.3.12	Sound	MXF; WAV
1.04.3.06	duration	MXF; WAV
1.04.3.07	pixels	MXF; DPX; Tiff

Finally, a third extension is used for creatingApplication. Here lie the attributes that have to do with the creation of a digital file: with which machine has the file been created and eventually: what were the important attributes of the analogue carrier. It goes without saying that here too there are format-specific characteristics.

Premis_v3	Name	Attribute of
1.05.5	creatingApplication	
1.05.5.1	creatingApplicationName	
1.05.5.2	creatingApplicationVersion	
1.05.5.3	dateCreatedByApplication	
1.05.5.4	creatingApplicationExtension	
1.05.5.4.3.2	sourceAspectRatio	DPX;MXF
1.05.5.4.3.3	sourceColor	Tiff;DPX;MXF
1.05.5.4.3.4	sourceDuration	WAV;MXF;DPX
1.05.5.4.3.5	sourceFramerate	WAV;MXF;PDP
1.05.5.4.4	digitizationRemarks	

2 customization

During the mapping it became clear that not all PMD fields could be mapped to a system field. Nevertheless this is not considered a shortcoming[7]. Some attributes implicitly follow either the PMD or general policy documents. There are also required fields that are not available in system metadata but are, for example, available in the header.

Here one can also speak of the *Technical Neutrality* of PREMIS. The fields determine only what information is needed to preserve the file, completely independent of system used and metadata-records. The way in which the archive provides information may differ, provided the preservation goals are adequately served.

[7] "A mandatory semantic unit is something that the preservation repository needs to know, independent of how or whether the repository records it. The repository might not explicitly record a value for the semantic unit if it is known by some other means (e.g., by the repository's business rules)." Data Dictionary for Preservation Metadata: PREMIS version 3.0, blz 31

In the PMD, Sound and Vision has made an initial attempt to distinguish the form in which the metadata may/must be available. In the PMD this has been elaborated in the choices *implicit, metadata* and *header*. For each property, this is specified per category. Metadata is required when a property for a group of files must be retrievable. The header suffices for the individual file. The recording of data in text files (e.g. baton report) is equated with header information in this respect.

Premis_v3	Name	Fimpl
1.05.2	fixity	
1.05.2.1	messageDigestAlgorithm	implicit
1.05.2.2	messageDigest	header
1.05.2.3	messageDigestOriginator	header
1.05.3	size	metadata

3 consistent implementation

PREMIS states which data-element is applicable in each *category* (Bitstream, File, Representation, Intellectual Entity). PREMIS authors emphasize that the model describes the general metadata archives will *probably* want to know for digital preservation[8]. This is expressed by the fact that not all data elements are mandatory: the model differentiates between *mandatory* and *optional* elements.

For each *category* (bitstream and file) that has been implemented, the PMD contains all applicable mandatory fields. Optional elements are alternately implemented at zero, one or both categories. Once implemented, all underlying mandatory fields are part of the PMD. This means that the PMD meets the *Degrees of freedom* as defined in the *statement of conformance.*

The example below shows the *Obligation* and *Object Category* (applicable) recommended by PREMIS. Based on this, CompositionLevel as well as File and Bitstream have been implemented. Fixity is an optional field and has only been implemented for File. Two Fixity subfields are mandatory for File.

Premis_v3	Name	Obligation	Object Category	Implemented
1.05	objectCharacteristics	M	FB	FB
1.05.1	CompositionLevel	M	FB	FB
1.05.2	fixity	O	FB	F
1.05.2.1	messageDigestAlgorithm	M	FB	F
1.05.2.2	messageDigest	M	FB	F
1.05.2.3	messageDigestOriginator	O	FB	F

[8] PREMIS uses this practical definition: *things that most working preservation repositories are likely to need to know in order to support digital preservation."* Data Dictionary for Preservation Metadata: PREMIS version 3.0, blz 3

Usability PMD 2.0

The PMD thus documents the preservation metadata with which Sound and Vision guarantees sustainable access to digital objects. Assurances are found in different ways and on different levels. The preservation metadata is used to:

a. Group digital objects in order to enable:
    i. Specific preservation planning and actions
    ii. Data management; data quality, efficient and consistent management
    iii. Collection forming, application of retention policy and access policy
b. Manage the lifecycle of digital objects
c. Control the migration of assets to new formats

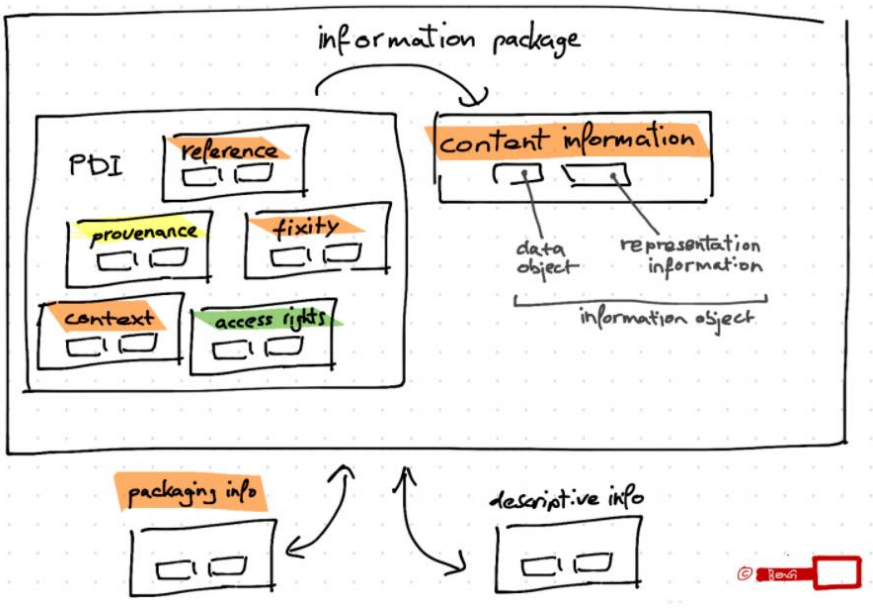
The PMD helps document which information must be minimally available in order to carry out these tasks. The PMD thus documents the basis description for OAIS conformant Information Packages

Archival Information Package (AIP)

The OAIS Archival Information Package is described by documenting in which system at Sound and Vision the values per field, and where needed, per filetype/ingest are found.

The figure below shows the structure of an Information Package according to OAIS. All subparts of the Preservation Description Information (PDI) are required in an *Archival* Information Package. Within the current PMD, the orange colored units are defined. Each mapping documents the AIP subunits by format or specific ingest.

In the following version *Events* (provenance) and *Rights* (access rights) can be added. The *Packaging info* will then also be complete for the Object categories *Representation* and *Intellectual Entity.*



Composition of an Information Package according to OAIS

Submission Information Package (SIP)

PREMIS gives suggestions on how attribute values can be acquired or updated (creation/maintenance notes). With this, PREMIS gives an indication of what values should be delivered in a SIP.

In the PMD, this is translated per attribute into an indication of whether the SIP must provide this information separately for one or more categories.

Premis_v3	Name	SIP
1.05.2	fixity	
1.05.2.1	messageDigestAlgorithm	F
1.05.2.2	messageDigest	F
1.05.2.3	messageDigestOriginator	
1.05.3	size	F

In this example, it is noted that the checksum and size forms a part of the SIP. That means that when ingesting a file, a checksum is expected and the size should be delivered separately.
The algorithm that is used (for example, MD5) can also be delivered if specified in the Submission Agreement, and that can be applicable for the complete ingest. Both checksum and size can be derived from the file by Sound and Vision after / during ingest. The fields in the SIP thus serve as a control mechanism.

The example demonstrates that further development is needed concerning the question of whether the value can be determined from the agreement or must be delivered in the metadata, or may be embedded in the file header. Future work on events will also determine which fields will be filled as a result of testing (for example, with help from a baton-profile).

Thus, the PMD plays an important role in documenting the minimum requirements for new ingest processes, at least concerning preservation metadata[9]. This will require further elaboration in the following version.

[9] In addition requirements may be developed for descriptive metadata.