Preservation Metadata Dictionary 2.0

Versie 0.3

Inleiding

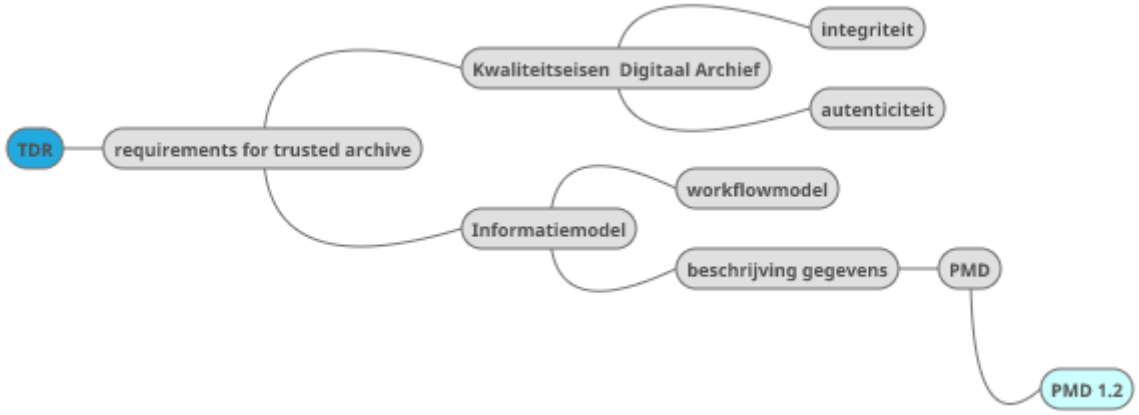
Het doel van een Preservation Metadata Dictionary (PMD) is het beschrijven van de kenmerken van de digitale objecten die in het Digitaal Archief worden beheerd en van de bewerkingen die ze ondergaan[1].

Dit document bevat een toelichting op de opzet van de *PMD, versie 2.0*. Deze versie vervangt de PMD 1.2. De upgrade PMD 2.0 is eind 2016 verschenen in een beta versie. Sindsdien is de opzet getoetst aan de praktijk en op een aantal punten verbeterd.

Achtergrond

Beeld en Geluid verkreeg in 2016 het predicaat “Trustworthy Digital Repository” (TDR). In de aanloop naar de certificering werden de requirements[2] voor TDR door Beeld en Geluid vastgelegd in de set Kwaliteitseisen Digitaal Archief en in een Informatiemodel. Dit informatiemodel kent twee componenten: een *workflowbeschrijving* voor de ingest, de opslag en de beschikbaarstelling van digitale files en metadata, en een *beschrijving van de gegevens* die tijdens de workflow en over de file worden geadministreerd.

Schematisch:



Zoals in het schema is te zien zijn de twee centrale kwaliteitseisen in het kader van ‘trusted repository’: *integriteit en authenticiteit*. Met andere woorden: de zekerheid dat een file niet corrupt is, en de zekerheid dat een file niet in de loop van de tijd onbedoeld is veranderd en daadwerkelijk is wat het voorgeeft te zijn. Met behulp van het informatiemodel wordt vastgelegd hoe door Beeld en Geluid de integriteit en authenticiteit worden behouden

[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

(workflow) en hoe dit kan worden aangetoond (gegevens)[3]. Met de Preservation Metadata Dictionary is dit laatste doel concreet uitgewerkt.

De eerste versie van de PMD (versie 1.2) werd opgebouwd uit drie componenten:

- een lijst met eigenschappen die horen bij de vier entiteiten van het PREMIS-model (Object, Event, Agent, Rights), gebaseerd op PREMIS 2.0
- een aanvulling van technische metadata afkomstig uit andere standaarden
- een uitwerking van *events* die voor Beeld en Geluid van toepassing zijn op grond van de workflowbeschrijving in het informatiemodel

Deze versie van de PMD vormde de input voor het project GAP-analyse PMD gedurende 2016 en 2017. Tijdens dit project is de reguliere instroom van de belangrijkste file-formaten van Beeld en Geluid getoetst aan de PMD.

De uitkomst van het project was niet alleen inzage in de gap met betrekking tot onze preserverings metadata, maar ook een upgrade van de PMD, naar versie 2.0.

Opzet conform PREMIS

De PMD is opgezet conform PREMIS[4]. PREMIS staat voor Preservation Metadata Implementation Strategy. The PREMIS Data Dictionary for Preservation Metadata is een internationale standaard voor metadata die de preservering van digitale objecten ondersteunt en duurzaam gebruik borgt. De dictionary bevat verschillende soorten metadata: administratief (waaronder rechten), technische metadata, en metadata die structuren vastlegt.

> Beeld en Geluid werkt op basis van internationale standaarden. Een belangrijk voordeel van het gebruik van een internationale standaard is dat het een referentiekader biedt voor derden; Beeld en Geluid kan altijd uitleggen hoe zijn PMD zich tot de algemene, internationale standaard verhoudt. PREMIS is voor Beeld en Geluid de standaard die wordt gehanteerd bij het definiëren van de preservation metadata.

In de eerste versie van de PMD waren vooral de ‘digital provenance’ (herkomst gegevens) en rechtenmetadata geënt op het PREMIS-model. De huidige versie van de PMD is integraal gebaseerd op PREMIS.

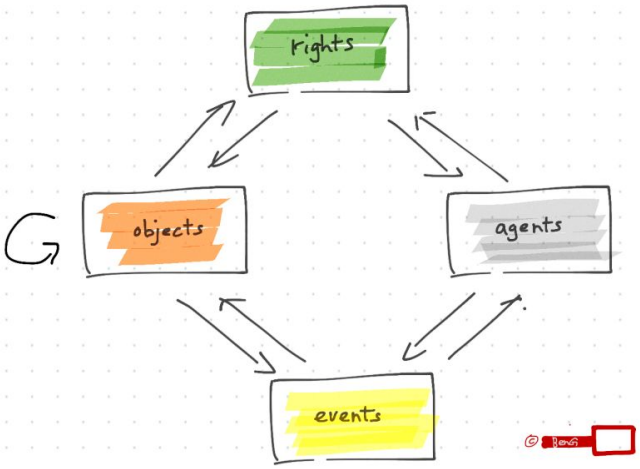
De PREMIS dictionary is opgebouwd uit een aantal entiteiten. De betekenis van deze entiteiten is als volgt:

- **Objecten** vormen het onderwerp van de digitale preservering. Het gaat om informatie-eenheden over de digitale content, op vier manieren:
    - een reeks bytes
    - een file

[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

    - een combinatie van files, nodig om een programma af te spelen
    - een abstracte beschrijving van het programma (de foto, het fragment enz)
- **Rights** beschrijven de *handelingen (acts)* die Beeld en Geluid met het materiaal mag verrichten. PREMIS focust daarbij op handelingen uit oogpunt van archiefbeheer.
- De **Events** zijn *gebeurtenissen* die in de loop van de tijd met een object hebben plaatsgevonden. Het gaat om gebeurtenissen uit oogpunt van archiefbeheer.
- Een **Agent** is een externe/hulp-entiteit. Voor Rights is dit bijvoorbeeld de partij met wie de afspraken over rechten zijn gemaakt, voor Events kan dit bijvoorbeeld het softwareprogramma zijn dat de gebeurtenis uitvoert.

De onderstaande figuur bevat het schema hoe de entiteiten zich tot elkaar kunnen verhouden.



Figuur: schema van entiteiten PREMIS 3.0

Uit het schema blijkt ondermeer:

- dat een Object een relatie kan hebben met Rights en/of Events
- dat de entiteit Object onderling naar elkaar kan verwijzen
- dat zowel Rights als Events elk Agents kunnen hebben

Hierna wordt de samenhang van de entiteiten nader toegelicht.

Objecten

Objecten vormen het onderwerp van de digitale preservering. De digitale content wordt op vier manieren ontleed. Hierdoor krijgt het model een extra dimensie. Elke manier (*category*) heeft een eigen verzameling eigenschappen en relaties. Dat betekent dat de bovenstaande figuur zich vier keer herhaalt, waarbij de relaties naar de andere entiteiten zich telkens opnieuw kunnen voordoen[5].

In de uitwerking van Beeld en Geluid komen niet alle relaties overal voor. Beeld en Geluid heeft hierin enkele ontwerpkeuzes gemaakt. De ontwerpkeuzes zijn inherent aan de interpretatie van de categorieën, wanneer deze worden toegepast in de praktijk van Beeld en Geluid. Deze interpretatie wordt hieronder toegelicht.

[5] Zie ook de Conceptual view between object categories, PREMIS 3.0, blz 9, figure 2, voor alle mogelijke relaties tussen de *categories* onderling en met zichzelf

Bitstream

- de Bitstream is de data in de body van de file, ongeacht of dit een streaming of still is. De bitstream en de file zijn onlosmakelijk met elkaar verbonden.

File

- met File worden alle digitale bestanden bedoeld; containers van digitale data, in beginsel herkenbaar aan hun extensie, met hun specifieke technische eigenschappen. Deze files worden beheerd in het storagemanagementsysteem (onder andere DIVA). *Bijna alle* events die nodig zijn in het kader van preservation, richten zich op het niveau van de file.
- files verwijzen onderling niet naar elkaar, hoewel er indirect wel sprake kan zijn van een relatie. Wanneer sprake is van een afgeleide file (een kopie, een transcodering etc) ligt deze indirecte relatie via een event. Vanuit de event is sprake van een inputfile (moeder) en een outputfile (dochter). Ook kan een indirecte relatie worden vastgelegd via de Representatie. Namelijk tussen een hoofd-file en een of meer sub-files (bijvoorbeeld de STL bij een MXF). Hoofd- en subfiles komen als pakketje binnen. Tijdens de import zal de volledigheid van het pakketje worden gecontroleerd; dit is een event op het niveau van de Representatie.

Representation

- om archiefmateriaal betekenisvol af te spelen (of te tonen) zijn vaak meerdere files nodig. Deze files vormen samen een Representation. Bijvoorbeeld DPX in combinatie met de bijbehorende WAV. Of twee opvolgende WAV-files die samen 1 programma bevatten.

Intellectual Entity

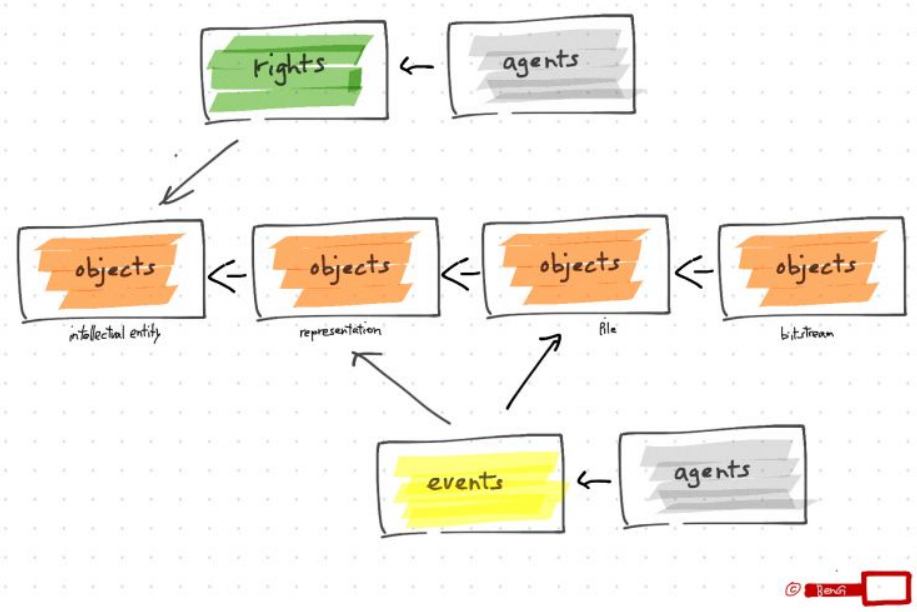
- de inhoudelijke metadata over de content van het materiaal vormt de Intellectual Entity. Denk aan radio- of TVprogramma's, een film, een foto, een geluidsopname. De Intellectual Entity kan één of meer Representations hebben. Bijvoorbeeld een op hoge resolutie (archiefkopie) en een op lage resolutie (raadpleeg-kopie).

Aansluiting bij het metadatamodel van het MAM.
In 2018 neemt Beeld en Geluid een nieuw MAM-systeem in gebruik. Het MAM-systeem kent verschillende entiteiten die niet altijd overeenkomen met de *categories* van PREMIS.
. Program: beschrijving van de content; vergelijkbaar met de intellectual entity
. Item: technische metadata van de master-file
. File: eigenschappen gebaseerd op headerinformatie van elke file, gerelateerd aan de master-file
. Package: combinatie van Items wanneer deze opeenvolgend moeten worden afgespeeld bij één programma
. Logtrackitem: informatie over een stukje van de master-file, op basis van positie-informatie
Achter elk *item* kunnen dus meerdere Files 'schuil gaan'. Bijvoorbeeld verschillende resoluties, meerdere backups enzovoort. Het MAM deelt de files in in: Highres, Proxy, Auxiliary, Other (keyframes, indexen). Hiervoor geldt dat de Highres en de Proxy elk afzonderlijke representations van hetzelfde programma zijn.

Uit de interpretatie van de categorieën bij Beeld en Geluid blijkt dat relaties met events en rights niet op ieder niveau voorkomen. Ook de onderlinge verwijzing van Objecten binnen

één niveau komt niet overal voor. Zo zullen de events (gebeurtenissen voor preservation) in de praktijk van Beeld en Geluid meestal plaatsvinden per file. Terwijl afspraken over rights (rechten archiefbeheer) in één keer per intellectual entity worden gemaakt om vervolgens voor de onderliggende files van toepassing te zijn.

Dit leidt voorlopig tot het volgende *logische data model* voor de PMD:



Vertaling van de Data Dictionary naar de praktijk bij Beeld en Geluid

PREMIS-compliancy

Implementatie volgens level 1A

Implementatie van de PREMIS standaard kan op meerdere manieren vorm krijgen. Bijvoorbeeld in de mate waarin de systemen van het archief zijn geënt op de PREMIS-indeling. Beeld en Geluid kiest hier voor *Conformancy level 1,* dat inhoudt dat het archief ervoor zorgt dat de velden van de PMD stuk voor stuk gemapt kunnen worden naar preservation metadata zoals die in de systemen is vastgelegd.

Dit Level 1 kan nog worden uitgesplitst in:

- Level 1A: Mapping op Object entiteit
- Level 1B: Mapping op Object, Event en Agent entiteiten

De PMD 2.0 voldoet aan level 1A (Mapping op Object entiteit). In de toekomst is uitbreiding naar level 1B mogelijk, wanneer na de implementatie van het MAM-systeem de vastlegging van events is uitgekristalliseerd.

Nadere uitwerking Objecten

De opzet van PREMIS is heel generiek zodat het breed toepasbaar is. In deze paragraaf wordt uitgelegd hoe Beeld en Geluid de PMD heeft uitgewerkt binnen enkele generieke uitgangspunten van het PREMIS-model[6].

1 uitbreidbaarheid

Technische eigenschappen vallen buiten de scope van PREMIS[7]. Om de integriteit en authenticiteit van een object te kunnen aantonen zijn technische eigenschappen echter van groot belang. In het bijzonder wanneer sprake is van migratie naar een ander formaat. Op dat moment is de checksum niet meer bruikbaar en zal aan de hand van een aantal technische specificaties moeten kunnen worden vastgesteld of de formaatmigratie succesvol is geweest. Ook wanneer een checksum niet wordt meegeleverd is het controleren van technische eigenschappen noodzakelijk.

De PMD 1.2 bevat een lijst met technische eigenschappen afkomstig van diverse specifieke standaarden zoals PBCore, EBUcore, AES, LC VideoMd, AudioMD and NARA's reVTMD. In de PMD 2.0 zijn deze eigenschappen ondergebracht in PREMIS door gebruik te maken van de *Extensibility* van het model.

De objectCharacteristicsExtension omvat technische eigenschappen. Per eigenschap heeft Beeld en Geluid gespecificeerd voor welk type file de eigenschap geldt.

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

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

[7] "Detailed, format-specific technical metadata is clearly necessary for implementing most preservation strategies, but the group had neither the time nor the expertise to tackle format-specific technical metadata for various types of digital files." Data Dictionary for Preservation Metadata: PREMIS version 3.0, blz 32

1.05.7.6.11	colorSpace	Tiff
1.05.7.6.12	samplesPerPixel	Tiff
1.05.7.6.13	bitsPerSample	Tiff

Voor eigenschappen die grote betekenis hebben voor de audio visuele prestatie van de file is een afzonderlijke extensie beschikbaar. In de PMD is een beperkt aantal eigenschappen in deze extensie ondergebracht.
Ook hier zijn sommige velden formaatspecifiek, of geldig in het algemeen (vb use) of voor meerdere formaten.

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

Tenslotte is gebruik gemaakt van een derde extension, namelijk voor de creatingApplication. Hierin liggen kenmerken vast die te maken hebben met de creatie van de digitale file: met welk apparaat is de file gemaakt en eventueel: wat waren belangrijke eigenschappen van de analoge drager. Het spreekt voor zich dat ook hier sprake is van formaatspecifieke kenmerken.

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 maatwerk

Bij de mapping is gebleken dat niet alle velden kunnen worden gemapt naar een eigen systeemveld. Toch is dan niet altijd sprake van een tekortkoming[8]. Sommige eigenschappen volgen impliciet uit de PMD zelf of uit algemene beleidsdocumenten. Er zijn ook verplichte velden die niet in de metadata, maar wel in bijvoorbeeld de header beschikbaar zijn.

In dit verband wordt ook gesproken van de *Technical Neutrality* van PREMIS. De velden bepalen slechts welke informatie nodig is ten behoeve van preservering, geheel los van gebruik van systemen en metadata-records. De wijze waarop door het archief in de informatie voorziet kan afwijken, mits de preserveringsdoelen voldoende worden gediend.

In de PMD heeft Beeld en Geluid een aanzet gedaan om onderscheid te maken naar de vorm waarin de metadata beschikbaar mag/moet zijn. In de PMD is dit uitgewerkt in de varianten implicit, metadata en header. Voor elke eigenschap is dit per *category* gespecificeerd. Metadata is vereist wanneer een eigenschap voor een groep files moet kunnen worden opgevraagd. Voor de individuele file volstaat de header. Het vastleggen van gegevens in textfiles (vb batonrapport) is in dit verband gelijkgesteld aan headerinformatie.

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 consistente implementatie

PREMIS geeft voor elk data-element aan bij welke *category* (Bitstream, File, Representation, Intellectual Entity) deze voor kan komen (*applicability*). De schrijvers van PREMIS benadrukken dat het model de metadata beschrijft die archieven in het algemeen *waarschijnlijk* willen weten in het kader van digitale preservering[9]. Dit brengen zij tot uitdrukking in het feit dat niet alle data-elementen verplicht zijn: het model onderscheidt *mandatory* en *optional* elementen.

De PMD bevat voor elke *category* die is geïmplementeerd (bitstream en file), alle mandatory velden die daarbij applicable zijn. Optional elementen zijn afwisselend geïmplementeerd op nul, één of beide *categories*. Eenmaal geïmplementeerd maken alle onderliggende

[8] "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

[9] 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

verplichte velden deel uit van de PMD. Hiermee voldoet de PMD aan de *Degrees of freedom* zoals die zijn gedefinieerd in het *statement of conformance*.

Het onderstaande voorbeeld toont de *Obligation* en *Object Category* (applicable) die door PREMIS is voorgeschreven. Op grond daarvan is CompositionLevel zowel voor File als Bitstream geïmplementeerd. Fixity is een optional veld en alleen voor File geïmplementeerd. Twee subvelden van fixity zijn vervolgens mandatory voor de 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

Bruikbaarheid PMD 2.0

De PMD legt daarmee de preservings metadata vast waarmee Beeld en Geluid de duurzame toegang tot de digitale objecten borgt. De borging vindt op verschillende wijzen en niveaus plaats. De preserverings metadata wordt gebruikt om:

a. digitale objecten te groeperen ten behoeve van:
    i. specifieke preserveringsplannen en acties
    ii. data management; data kwaliteit, efficient and consistent beheer
    iii. collectie vormen, toepassen retentiebeleid and toegangsbeleid
b. beheer van de levenscyclus van digitale objecten
c. controleren migratie van assets naar nieuwe formaten

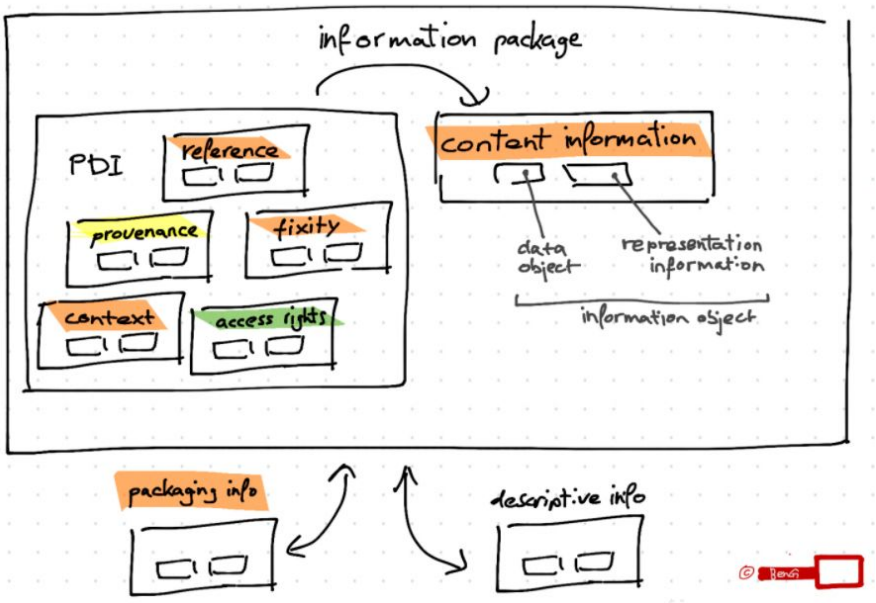
Met behulp van de PMD wordt vastgelegd welke informatie minimaal beschikbaar moet zijn om deze taken uit te kunnen voeren. De PMD legt daarmee een basis voor de beschrijving van Information Packages conform OAIS.

Archival Information Package (AIP)

Door per veld, waar nodig per filetype/instroom, vast te leggen waar deze waarden in de systemen van Beeld en Geluid worden vastgelegd, wordt de Archival Information Package van OAIS beschreven.

De onderstaande figuur toont de opzet van een Information Package volgens OAIS. Voor de *Archival* Information Package geldt dat alle onderdelen van de Preservation Description Information (PDI) verplicht zijn. Met de huidige PMD zijn alle oranje gearceerde delen gedefinieerd. Elke mapping legt de AIP vast voor die onderdelen, wat betreft die formaten of specifieke instroom.

In een volgende versie zullen *Events* (provenance) en *Rights* (access rights) kunnen worden toegevoegd. Met de Object categories *Representation* en *Intellectual Entity* wordt dan ook de *Packaging info* compleet.



Samenstelling van een Information Package volgens OAIS

Submission Information Package (SIP)

PREMIS geeft suggesties hoe de waarden voor een eigenschap kunnen worden verkregen en/of ge-update (creation/maintenance notes). Daarmee geeft PREMIS een aanwijzing of de waarden in de SIP moeten worden aangeleverd.

In de PMD is dit per eigenschap vertaald naar een aanduiding of de SIP deze informatie op één of meer *categories* afzonderlijk moet aanleveren.

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 dit voorbeeld is de checksum en size als onderdeel van de SIP genoteerd. Dat betekent dat bij een ingest met de file ook een checksum wordt verwacht en dat de size apart moet worden aangeleverd.
Het algorithme dat daarbij is gebruikt (bijvoorbeeld MD5) kan als onderdeel van de Submission Agreement worden aangeleverd, en geldt dan voor de volledige instroom.
Zowel checksum als size kunnen door Beeld en Geluid na/tijdens ingest worden afgeleid van de file. De velden in de SIP dienen dus als controlemechanisme.

Het voorbeeld laat zien dat een nadere uitwerking nodig is met betrekking tot de vraag of de waarde mag blijken uit de agreement, moet worden meegeleverd in de metadata, of mag worden ingebed in de header van de file. Bij de nadere uitwerking van de events zal ook blijken welke velden onderwerp zijn van toetsing (bijvoorbeeld met behulp van een baton-profiel).

De PMD heeft dus ook een belangrijke rol in het vastleggen van minimumeisen voor nieuwe ingest, tenminste, wat betreft de preservation metadata[10]. Dit vraagt nog wel om verdere uitwerking in een volgende versie.

[10] Daarnaast kunnen eisen worden gesteld aan de descriptive metadata.