Preserveringsmetadata bij Beeld en Geluid

De Preserverings Metadata Dictionary als praktisch instrument

De implementatie van een standaard voor preserveringsmetadata draagt bij aan de duurzaamheid van het archief. In het algemeen is een standaard generiek opgezet teneinde een zo groot mogelijk toepassingsbereik te bedienen. Om de standaard te implementeren is een vertaalslag nodig naar de eigen praktijk. Dit artikel vertelt hoe deze vertaalslag door Beeld en Geluid is gemaakt. Het resultaat is een praktisch instrument, de Preserverings Metadata Dictionary, dat ons dagelijks helpt bij het inrichten en managen van archiveringsprocessen.

Tekst **Marjolein Steeman**

Het archief

Het Nederlands Instituut voor Beeld en Geluid in Hilversum is een van de grootste archieven ter wereld en bewaart diverse soorten media, zoals radio- en televisieprogramma's, video(games), geschreven pers, politieke prenten, gifjes, websites en objecten. Vanuit de wettelijke opdracht tot het verzamelen, bewaren en ontsluiten van audiovisueel erfgoed van nationaal belang volgt de verplichting om het materiaal dat aan de organisatie wordt toevertrouwd, duurzaam in stand te houden, zodat alle gebruikersgroepen er blijvend toegang toe kunnen hebben.

Het instituut is in 1996 opgericht, als het nationale audiovisuele archief van Nederland. In Beeld en Geluid zijn de collecties van de Nederlandse publieke omroepen, de Stichting Film en Wetenschap, de filmcollectie van de RVD en het Omroepmuseum samengebracht. De belangrijkste depotgever van Beeld en Geluid, de Nederlandse publieke omroep, startte in 2006 met het digitaal produceren van televisie- en radioprogramma's. Daarnaast digitaliseerde Beeld en Geluid in zeven jaar tijd meer dan 100.000 uur aan film- en videomateriaal. De omvang van de unieke audiovisuele content in het archief bedraagt intussen 14,7 Petabyte.

Digitale preservering

Al dit materiaal wordt duurzaam gepreserveerd. Dat wil zeggen dat Beeld en Geluid zorgt voor het behoud van de integriteit en authenticiteit van de digitale objecten. *Integriteit* houdt in dat het object aantoonbaar ongewijzigd is. Het is een begrip dat gaat over de files en de samenstelling van de file op *bitlevel*. Het regelmatig controleren van de file met behulp van een daarvoor specifiek ontworpen algoritme (checksum) is hiervoor de belangrijkste waarborg. *Authenticiteit* betekent dat het object is wat het voorgeeft te zijn; het is aantoonbaar ongewijzigd sinds de aanlevering of er kan worden gedemonstreerd dat na transformatie alle kenmerkende eigenschappen zijn behouden.

Authenticiteit grijpt aan bij de wijze waarop de audiovisuele content zich voordoet op moment van afspelen of tonen. Het gaat hier om het in stand houden van de authentieke beleving. Er zijn drie voorwaarden waaraan moet worden voldaan:[1]

1. Het object is wat het voorgeeft te zijn. Hiertoe wordt een kwaliteitsanalyse op het object uitgevoerd waarbij technische eigenschappen worden gecontroleerd, en waarbij tevens kan worden vastgesteld of de specificaties van een bestand voldoen aan de standaard van het archief.
2. Het object is bruikbaar, afspeelbaar en heeft betekenis voor de gebruiker. Hiertoe worden de essentiële kenmerken van het object behouden. Dit kunnen technische, esthetische of intellectuele eigenschappen zijn, die – door de tijd en door alle technologische veranderingen heen – bewaard moeten blijven.
3. Het object is niet ongeautoriseerd of onbedoeld veranderd. Hiertoe wordt de levenscyclus van het object vastgelegd met behulp van procesmetadata (*provenance* of *audit trail*).

   Deze toelichting maakt duidelijk dat een van de belangrijkste sleutels tot een duurzaam digitaal archief ligt

1 A. de Jong, 'Digitale Preservering Beeld en Geluid: Beleid, Standaarden en Procedures' (2016) via http://publications.beeldengeluid.nl/pub/387, geraadpleegd op 27 mei 2019.

in de controle over de technische metadata en procesmetadata. Samen vormen deze het fundament van digitale duurzaamheid.

Praktijk

In de dagelijkse praktijk van Beeld en Geluid worden in een continue stroom materialen, afkomstig van de omroep en uit andere bronnen dan de omroep, *digital-born* en gedigitaliseerd, ingenomen, beheerd en gepresenteerd vanuit het Digitaal Archief. Bij het innemen komt metadata mee met de file zelf of uit de bronsystemen van de depotgevers. Tijdens de inname en het beheer genereren de systemen automatisch procesmetadata en technische gegevens. *File-headers* worden uitgelezen en in metadatavelden geregistreerd. Controles van bestandstype en van metadata zijn een vast onderdeel van de workflows. Ook tijdens filmscanning- en encoderingsacties wordt de nodige technische en procesmetadata gecreëerd, zoals analysegegevens en de opmerkingen van medewerkers over het bronmateriaal.

Deze gegevens landen in de diverse systemen en systeemmodules. Soms worden eigenschappen op meerdere plekken en momenten geregistreerd. Naast het feit dat de gegevens op verschillende plekken in het systeemlandschap landen, gebeurt dit niet bij alle instroom workflows op dezelfde wijze. Er is dus sprake van een grote variëteit.

De noodzaak van orde in deze *preservation* metadata voor digitale duurzaamheid was in een vroeg stadium van het digitale archief duidelijk. Het selectief formeel aanwijzen van gegevens als *preservation* metadata werd gezien als een onmisbare eerste stap op weg naar een duurzaam digitaal archief. Een groot voordeel daarbij was dat Beeld en Geluid slechts een beperkt aantal bestandstypes (file formaten) in het archief opnam. Deze formaten werden in samenspraak met de omroepen vastgesteld en golden de facto als standaard voor het archief.

Deze eerste stap resulteerde in een groslijst van technische en *provenance* metadata die mogelijk een rol kunnen spelen bij preservering. Voor de *provenance* metadata baseerde Beeld en Geluid zich op de internationale standaard voor preserveringsmetadata, PREMIS 2.0. Voor de technische metadata werd een keus gemaakt uit attributen uit onder andere de EBU P-meta, PBCore, the Library of Congress VideoMD and AudioMD, NARA reVTMD en de ANSI/NISO Z39.87 Data Dictionary Technical Metadata for Digital Still Images. De groslijst kreeg de naam Preserverings Metadata Dictionary (PMD). De daarop volgende versie van de PMD werd volledig gebaseerd op PREMIS 3.0.[2]

PREMIS

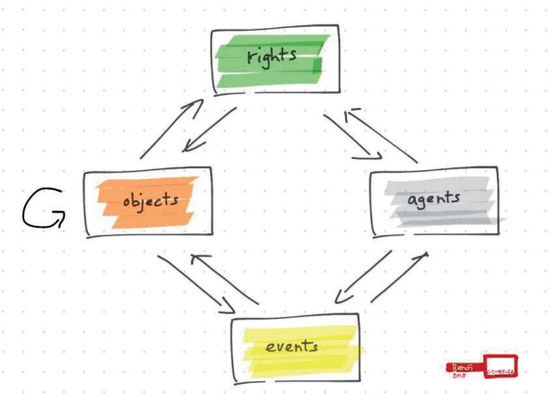
PREMIS staat voor *Preservation Metadata: Implementation Strategies*. De PREMIS Working Group publiceerde in juni 2005 zijn *Data Dictionary for Preservation metadata*. Deze basislijst van metadata-gegevens werd aanbevolen voor alle archieven, ongeacht het type materiaal of gekozen werkwijze van preserveren. De lijst was geheel systeemonafhankelijk en streefde naar een verbetering van uitwisseling tussen systemen met behulp van gestandaardiseerde exports. Een belangrijk uitgangspunt bij het opstellen van de lijst was alle gegevens op te nemen waarvan het *waarschijnlijk* was dat een digitaal archief die moet willen weten om duurzaam te kunnen preserveren.

De implementatie van PREMIS biedt veel ruimte voor maatwerk. Deels in het toevoegen van formaatspecifieke kenmerken, deels in de reikwijdte waarmee de Dictionary wordt gebruikt.

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 niveaus:
  - de bitstream: een reeks bytes in een file (videostream, audiostream);
  - een file;
  - representatie: een combinatie van files, nodig om één programma af te spelen;
  - *intellectual* item: een beschrijving van het programma (de foto, het fragment, enzovoort).
- **Rights** vormen de *beperkingen en/of toestemming* om bepaalde preserveringshandelingen (acts) met het materiaal te verrichten.
- De **Events** zijn *gebeurtenissen* die in de loop van de tijd met een object hebben plaatsgevonden.
- 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.



Schema van entiteiten PREMIS 3.0

Uit het schema blijkt onder meer dat een 'object' een relatie kan hebben met 'rights' en/of 'events', dat de entiteiten 'object' onderling naar elkaar kunnen verwijzen en dat zowel 'rights' als 'events' elk 'agents' kunnen hebben.

De implementatie van PREMIS vindt plaats in een aantal stappen, waarin het model wordt vertaald naar een logisch

2 PREMIS, Data Dictionary for Preservation Metadata, version 3.0, june 2015 via http://www.loc.gov/standards/premis/v3/, geraadpleegd op 27 mei 2019.

datamodel[3] dat geschikt is voor de concrete organisatie waarin het wordt toegepast.

Stap 1: de objecten die onderwerp zijn van archivering worden in lijn gebracht met de vier objectniveaus van PREMIS. PREMIS kent vier niveaus, variërend van de bitstream in de file tot het intellectuele abstracte object waarvan de beleving bewaard moet worden. De gelaagdheid van het model maakt het mogelijk om de 'events' en de 'rights' op verschillende manieren aan objecten te koppelen. Tijdens de eerste stap wordt bepaald op welke manier de objecten van het archief tot uitdrukking worden gebracht op deze vier verschillende niveaus.

Stap 2: per niveau wordt bepaald welke metadata nodig zijn om de objecten en eventueel gekoppelde entiteiten te beschrijven. De PREMIS veldlijsten zijn daarbij een vertrekpunt. Niet alle velden zijn verplicht; het archief bepaalt zelf welke attributen van toepassing zijn. Daarbij kan ook gebruik worden gemaakt van extensies die PREMIS biedt: uitbreidbare delen van de lijst, bijvoorbeeld om formaatspecifieke attributen toe te voegen.

Het resultaat is een normatieve lijst met attributen per entiteit (bijvoorbeeld object of event), op elk niveau dat eerder in stap 1 is gedefinieerd, voor elk specifiek formaat dat voor het archief van toepassing is. De Preserverings Metadata Dictionary 2.0 is zo'n lijst.[4]

De volgende stap in de implementatie van PREMIS is de confrontatie met de werkelijke registratie in de systemen. Zijn voor elk veld daadwerkelijk waarden beschikbaar, voor elk denkbare instroom van archiefmateriaal? Beeld en Geluid onderzocht op deze wijze alle attributen voor de gangbare fileformaten (MXF, WAV, TIFF en DPX). De attributen werden gemapt met de velden die daadwerkelijk in de bestaande systemen aanwezig zijn.

Dankzij de mapping met gedefinieerde en geformaliseerde *preservation metadata* in de PMD kunnen gegevens worden geïdentificeerd die nodig zijn uit oogpunt van duurzaam archief. Op elk moment, voor elke subset van assets. Door deze wijze van implementatie (via mapping naar systeemvelden) voldoet de PREMIS-implementatie van Beeld en Geluid aan de basiseisen van implementatie.[5]

Een volgend *conformancy level* zou zijn wanneer de gegevens geautomatiseerd uit de systemen kunnen worden gegenereerd. Beeld en Geluid werkt momenteel aan een *business intelligence tool* waarmee de verschillende systemen worden bevraagd en de gegevens daaruit in onderlinge samenhang kunnen worden gerapporteerd. Daarmee komt het volgende *conformancy level* in zicht.

PMD ten dienste van duurzaamheid

Wat begon als een hulpmiddel om de preserveringsmetadata te ordenen heeft zich ontwikkeld tot een instrument dat drie belangrijke functies vervult in het kader van preservering:

1. Controlemechanisme na transformatie-acties
   We moeten achteraf kunnen vaststellen of bepaalde eigenschappen van een digitaal object na een transformatie (bijvoorbeeld een migratie naar een nieuw formaat) behouden zijn gebleven dan wel zijn gewijzigd. Hiervoor is het nodig vorm en inhoud van deze eigenschappen vooraf te identificeren, standaardiseren en definiëren.

2. Selecties van een 'groep' objecten op grond van bepaalde kenmerken
   Er moeten (ten behoeve van overzichten, rapportages en planningen voor bijvoorbeeld preserveringsactieplannen, ICT-beheer, collectie, retentie- en accessbeleid) willekeurige groeperingen van objecten gemaakt kunnen worden op basis van (technische) overeenkomsten tussen die objecten (zoals bestandstype, opslaglocatie, instroomproces). Hiertoe is het noodzakelijk deze eigenschappen te identificeren en ze gestandaardiseerd vast te leggen.

3. Overzichten van de levenscyclus van een object of groep objecten
   Gebeurtenissen (binnen fasen) in de levenscyclus van een object of een groep objecten (transfer, *ingest*, opslag, preservering, *access*) moeten kunnen worden weergegeven als *audittrail*. Hiervoor is het nodig om alle relevante gebeurtenissen vooraf te definiëren als preserveringsevent. Met een *audittrail* kan worden aangetoond dat (a) de acties die zijn uitgevoerd overeenkomen met het *ingest*-, storage-, backup-, preserverings- en *access*beleid en (b) dat het object/de groep niet ongeautoriseerd gewijzigd is.

Conclusie

De Preservation Metadata Dictionary vormt voor het bewust managen en beheren van de digitale collectie een onmisbaar centraal referentiekader. De dictionary biedt een formeel overzicht van de technische en procesmetadata en legt precies vast hoe deze moeten kunnen worden uitgedrukt en wat ze betekenen. Hiermee draagt de dictionary bij aan de praktische, systematische beheersing van alle preserveringsprocessen. ●

Marjolein Steeman

Senior Mediamanager bij Beeld en Geluid

3 Angela Dappert, Rebecca Squire Guenther, Sébastien Peyrard (eds.), Digital Preservation Metadata for Practitioners, 2016.

4 M. Steeman (ed.) Preservation Metadata Dictionary 2.0, Netherlands Institute for Sound and Vision, 2018, via http://publications.beeldengeluid.nl/pub/615, geraadpleegd op 27 mei 2019.

5 PREMIS Conformance, dd. 20 november 2017, http://www.loc.gov/standards/premis/premis-conformance-20150429.pdf, geraadpleegd op 27 mei 2019.