Preservering van digitale AV-collecties volgens de OAIS standaard

Requirements voor een 'trusted' archief

door Annemieke de Jong, Beth Delaney en Daniel Steinmeier

Inleiding

Het Nederlandse Instituut voor Beeld en Geluid bekleedt al sinds haar oprichting in 1997 een dubbelfunctie. Allereerst fungeert het instituut als centraal bedrijfsarchief voor alle publieke omroeporganisaties (20) in Nederland. In die hoedanigheid is de organisatie verantwoordelijk voor de opslag en de beschikbaarstelling van uitgezonden radio- en televisieprogramma's. Beeld en Geluid is daarnaast het nationale AV-archief van Nederland. Dit houdt in dat de collecties tevens worden opgebouwd vanuit ander, Nederlands audiovisueel erfgoed, bijvoorbeeld wetenschappelijk AV-materiaal, documentaire film, foto's en objecten, amateurfilm en AV-collecties van bedrijven en maatschappelijke organisaties. Binnen het Nederlandse landschap van omroep en erfgoed vormt Beeld en Geluid een centraal audiovisueel knooppunt, zowel waar het gaat om digitale opslag en presentatie als om het verzamelen en verspreiden van kennis over het vakgebied. Momenteel omvatten de collecties van Beeld en Geluid meer dan 800.000 uur radio, televisie en film. Jaarlijks worden daar -vanuit de omroepproductieomgeving – meerdere duizenden uren *digital born* materiaal aan toegevoegd.

De afgelopen jaren hebben bij Beeld en Geluid in het teken gestaan van het digitaliseren van de bestaande analoge collecties. In het megaproject Beelden voor de Toekomst heeft de organisatie ongeveer de helft van zijn bestaande analoge film, video en audiocollecties omgezet in digitale files. Andere belangrijke ontwikkeling was het direct aansluiten van het archiefsysteem bij de digitale omroep-productiesystemen. Hierdoor stromen sinds 2007 alle uitgezonden RTV-programma's materiaal, met bijbehorende metadata, dagelijks in een geautomatiseerde digitale workflow het archief binnen. In de kelders van Beeld en Geluid is inmiddels voor het gedigitaliseerde materiaal en de digital borns een digitaal *respository* ingericht van ca. 6 petabyte. Iedere dag weer worden vanuit deze voorziening vele honderden gebruikers bediend met beeld- en geluidsfragmenten, zowel in de professionele omroepomgeving als daarbuiten, bij de mensen thuis, in kringen van onderwijs en bedrijfsleven en binnen het museum van Beeld en Geluid, de Media Experience.

Preservering

Nu Beeld en Geluid in staat is een groot deel van zijn collecties in digitale vorm aan te bieden en er dagelijks vele nieuwe digitale files binnenstromen, dient zich een belangrijke nieuwe uitdaging aan: de beheersing van de snel en continu groeiende digitale *storage* en de toenemende complexiteit van de processen daaromheen. Het afbreukrisico bij de bewaring van digitale materialen is groot, en het is strikt noodzakelijk om meer greep te krijgen op de levenscyclus van de digitale files. Hiertoe zullen processen, procedures en metadata bedoeld voor lange termijn preservering moeten worden geïncorporeerd in de systemen van Beeld en Geluid. Ook moeten de rollen en verantwoordelijkheden voor lange termijn bewaring van zowel de depotgevers, als het archief en de gebruikers beter en preciezer worden gedefinieerd.

Hierbij luidt de belangrijkste vraag: hoe houden we het opgeslagen, digitale materiaal blijvend toegankelijk voor onze gebruikers? De verscheidenheid aan formaten, omvang en locaties van de dagelijks aan grote groepen klanten uit te leveren files is enorm en neemt nog steeds toe.
Hoe kan het archief ervoor zorgen dat het blijvend in staat is haar materiaal aan te bieden in een format dat de *Designated Communities* van de collecties van Beeld en Geluid ook daadwerkelijk kunnen gebruiken ?

Hiermee, met lange termijn preservering dus, is (nog) niet expliciet rekening gehouden in de aansluiting bij de omroepproductiesystemen. Ook het inrichten van de infrastructuur voor de instroom (*ingest)* en het opslaan van erfgoedcollecties is - in technische zin - niet direct gebeurd vanuit het oogmerk van lange termijn preservering.

Voor zowel de rol als bedrijfsarchief van de omroepen als de positie als nationaal AV knooppunt is de garantie op betrouwbare en duurzame archivering echter strikt vereist.

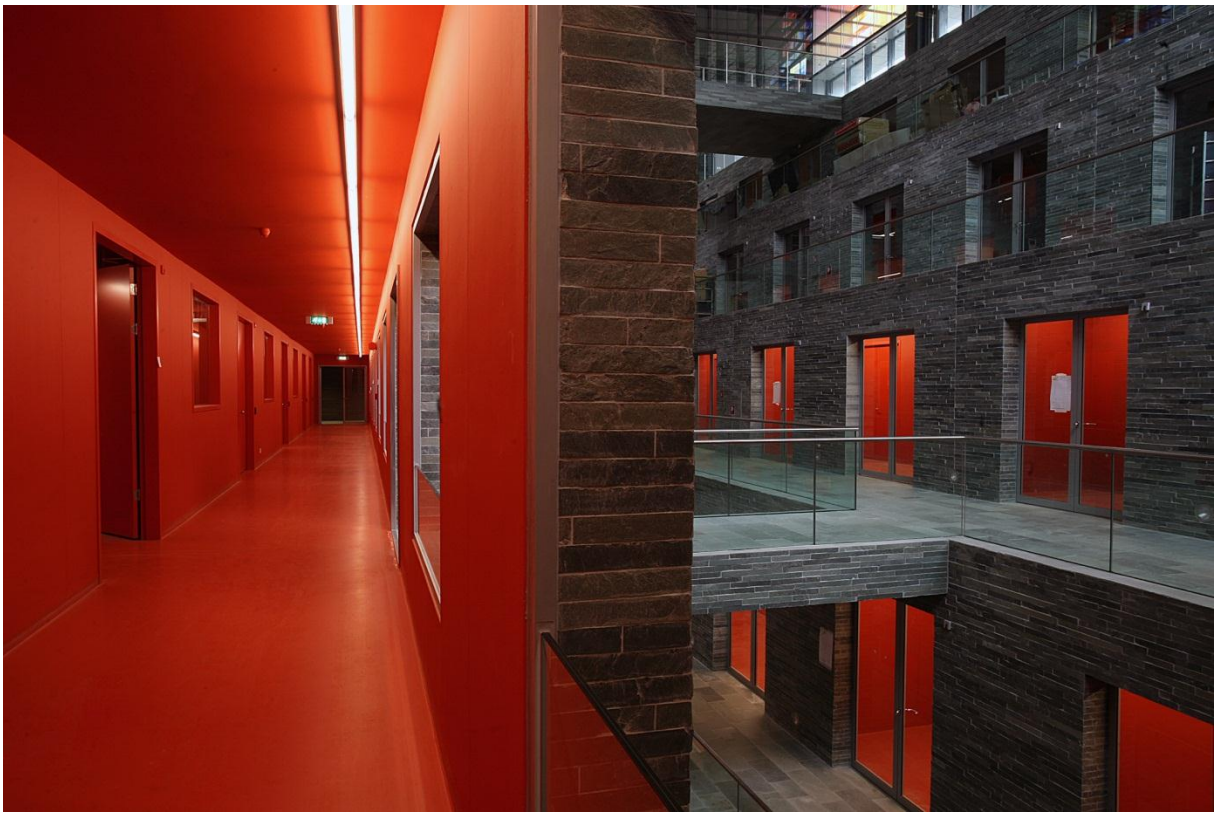

Fig.1 In de kelders van Beeld en Geluid ligt inmiddels ruim 6 petabyte aan digital born en gedigitaliseerd materiaal opgeslagen. Hoe wordt gegarandeerd dat al dit materiaal duurzaam toegankelijk blijft?

Deze nieuwe uitdaging is door Beeld en Geluid omgezet in een van haar belangrijkste strategische doelstellingen voor de komende periode. Beeld en Geluid wil "Trustworthy Digital Repository" (TDR) worden voor het Nederlandse audiovisueel erfgoed. Materiaal, of het nu gaat om publieke radio- en televisieprogramma's of om ander audiovisueel erfgoed, moet in dit archief veilig worden opgeslagen en permanent toegankelijk zijn voor wie het maar wil gebruiken. Het nieuwe strategische doel heeft in 2012 geleid tot de formulering van een project waarin de *requirements* voor een vertrouwde en betrouwbare manier van AV-archivering moesten worden bepaald.

Wat in het project opgeleverd zou moeten worden is een set normatieve beleidsdocumenten die sturing en richting kunnen geven aan de inrichting van een AV-archiveringsomgeving die in lijn is met de OAIS standaard voor digitale archieven. Een set Kwaliteitseisen Digitaal Archief Beeld en Geluid zou hierbij het centrale referentiekader moeten vormen.

Naast een aantal beleidsdocumenten voor de organisatorische en administratieve systematiek van een OAIS *compliant* archief werd het hart van het project gevormd door de requirements voor het zgn. *Digital Object Management*.

Dit onderdeel bestond uit een te ontwikkelen workflowmodel voor de ingest, de opslag en de beschikbaarstelling van digitale files en metadata.

Voor de beschrijving van de gegevens zou een *Preservation Metadata Dictionary* worden gecreëerd met een set gedefinieerde technische, administratieve en herkomst (*provenance)* gegevens. Workflow en preservation metadata samen, vormen het Informatiemodel van het digitaal archief van Beeld en Geluid. Inmiddels zijn alle projectdocumenten opgeleverd.

OAIS : Open Archival Information System

De voornaamste inspiratie bij het vormgeven van workflows en metadata in het Informatiemodel vormde het OAIS–referentiemodel. De ISO-standaard OAIS[1] biedt een basisopzet voor de inrichting van een 'trusted' digitaal archief. De standaard definieert – in een uniforme, gemeenschappelijke taal - alle processen die nodig zijn voor het duurzaam bewaren en het toegankelijkheid houden van informatie-objecten. OAIS biedt een conceptuele benadering van de inrichting van de zgn. 'business processen' van een digitaal archief d.w.z. het realiseren van duurzame preservering van de collectie. De standaard vormt aldus een essentieel hulpmiddel om de TDR status van een archief te realiseren.

Centraal in het OAIS-concept staat de garantie blijvende toegankelijkheid aan diegenen, die materialen aan het archief in beheer geven. Het model biedt richtlijnen voor het definiëren en formaliseren van processen en gegevensstructuren binnen de hele archiveringsketen: van inname, via opslag tot aan uitlevering.

Door het volgen en vastleggen van alle vooraf gedefinieerde stappen in de levenscyclus van ieder afzonderlijk ingestroomd object in zgn. 'preserveringsmetadata' kan de authenticiteit van het object zowel worden gewaarborgd als aangetoond, en wordt aldus de basisvoorwaarde voor het 'trusted' zijn vervuld. Het Archief of de repository is hierdoor in staat om te allen tijde verantwoording af te leggen aan zowel haar depotgevers als haar afnemers.

OAIS is een wijdverbreid referentiemodel dat tot nu toe m.n. opgeld doet in de wereld van de *digital libraries* en de traditionele 'papieren' archieven. Dit geldt ook voor de belangrijkste standaard voor preserveringsmetadata: PREMIS (Preservation Metadata Implementation Standard)[2] . Toepassing en implementatie van deze standaarden in het archivale mediadomein – waar de nadruk van oudsher ligt op toegang en hergebruik - zijn nog schaars.

Hoe snel groeiende digitale volumes op een rationele en verantwoorde manier te blijven managen en preserveren, is uiteraard ook een vraag die zich in toenemende mate opdringt aan omroeparchieven en andere grote audiovisuele collectiehouders.

Het Informatiemodel Digitaal Archief Beeld en Geluid 1.0 formuleert voor Beeld en Geluid een eerste, normatief antwoord op deze vraag. De in OAIS en PREMIS beschreven processen en gegevens zijn in dit model waar noodzakelijk aangepast aan de eigen aard van AV-files en naar de specifieke workflows en behandeling van AV-bestanden in een dynamische media-productieomgeving. In het bijzonder is gekeken naar toepassing binnen het domein waarin Beeld en Geluid fungeert als bedrijfsarchief van de Nederlandse Publieke Omroep (NPO), als nationale AV-erfgoedbeheerder en als online-aanbieder van AV-content aan een variëteit van gebruikersgroepen.

[1] ISO 14721:2003. The Open Archival Information System Reference Model
[2] http://www.loc.gov/standards/premis/

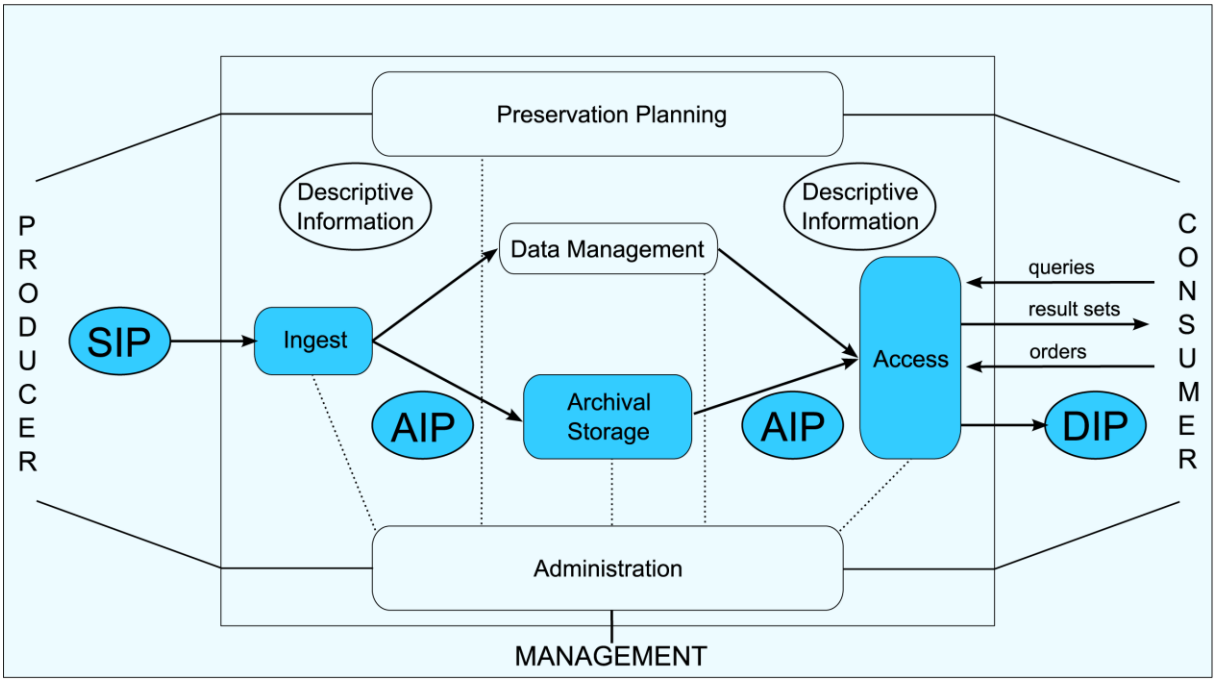

Fig. 2 OAIS Schema met gemarkeerd de workflow van de informatiepakketten SIP, AIP en DIP met content en metadata : van instroom (ingest) , via opslag (storage) tot het aanbieden aan gebruikers (access).

Authenticiteit en integriteit

Beeld en Geluid is tot nu toe nog niet aantoonbaar uitgerust om depotgevers en gebruikers van de collectie de belangrijke garanties te bieden die onlosmakelijk horen bij een status als Trusted Digital Repository : de integriteit en de authenticiteit van de digitale objecten. Wat is er voor nodig om deze garanties te kunnen gaan verlenen? In het algemeen zal een repository die is ingericht volgens OAIS de integriteit en authenticiteit moeten kunnen aantonen door middel van een vastgelegd 'spoor': een zgn. *audit trail*. Zo'n spoor is het samengesteld resultaat van alle zgn. *preservation events* (i.c. de bewerkingen van/aan digitale objecten, uitgevoerd om ze duurzaam te kunnen bewaren) zoals die plaatsvinden als onderdeel van een gedefinieerde preserveringsworkflow.

Zo vereist het kunnen aantonen van de integriteit van een file bijvoorbeeld dat zeker is gesteld dat *datastream* niet corrupt is geworden gedurende een transport of tijdens de opslag. Mechanismen voor validatie en foutcontrole *(error checking*) stellen de repository in staat om steeds weer na te gaan of er inderdaad geen corruptie is opgetreden. Dit gebeurt vooral op basis van zgn. *fixity* informatie - bijvoorbeeld in de vorm van een *checksum* of controlecijfer - die een file op bitniveau karakteriseert. Eenmaal gegenereerd (bij voorkeur vóór instroom in de repository) wordt deze checksum, na elke verplaatsing of bewerking van de file vergeleken met het oorspronkelijke cijfer. Op basis hiervan kan de integriteit van het object door de tijd heen worden aangetoond.

De authenticiteit van een file wordt bepaald op grond van bewijs. De centrale vragen die moeten kunnen worden beantwoord zijn : is een object werkelijk wat het voorgeeft te zijn? Is het niet onbedoeld veranderd? En als het dan gewijzigd is, is dat wel gedocumenteerd?
Het aantonen van authenticiteit door de tijd heen vereist dat de herkomst (*provenance)* van een object door het volledige bewerkingsproces *(chain of custody*) wordt vastgelegd : vanaf creatie en instroom via de opslag tot en met uitlevering.

Voor betrouwbare, duurzame bewaring van digitale files zijn dus twee essentiële zaken vereist. Allereerst is dat een set gedefinieerde business processen gericht op preservering, die zeker stellen dat aan preservering gerelateerde events ook daadwerkelijk plaatsvinden. Vervolgens gaat het om het hebben van een mechanisme met behulp waarvan een audit trail kan worden gegenereerd en in stand gehouden, zodat het archief in staat is de uitkomst van de preserverings events daadwerkelijk aan te tonen.

Het TDR project van Beeld en Geluid werd opgezet om deze processen te definiëren alsmede om een robuuste set technische en preservation metadata te ontwikkelen, waarmee de belangrijkste *life cycle management* informatie van ingestroomde digitale files kan worden gedocumenteerd en gemanaged.

Informatiemodel

In het Informatiemodel wordt beschreven welke workflows onderscheiden worden binnen de processen rondom de functies instroom (ingest), opslag (storage) en het verlenen van toegang (access). Het model legt alle acties of *events* vast die plaats kunnen vinden op een object en beschrijft aldus de totale levenscyclus van dat object. Het vastleggen van deze levenscyclus gebeurt door middel van (zoveel mogelijk automatisch gegenereerde) metadata die de uitkomst is van bewerkingen aan het object. Deze informatie wordt vervolgens aan het object toegevoegd. Ze documenteert aldus de bewerkingsgeschiedenis van het object en vormt daarmee het herkomstdeel (provenance) van de preserveringsmetadata.

Waar in de workflow de acties of events plaatsvinden en in welke provenance metadata dit resulteert, is vastgelegd in het informatiemodel. Het vooraf definiëren is noodzakelijk om een referentiekader te hebben waarmee geverifieerd kan worden dat alle events in de levenscyclus van een object, ook voldoen aan het preserveringsbeleid van het archief. Door provenance-metadata en de events in het informatiemodel met elkaar te vergelijken kan dus vastgesteld worden dat er - als de workflow correct is verlopen - geen onverwachte acties zijn uitgevoerd op het object. Dit moet uiteindelijk voor de eindgebruiker bijdragen aan een gevoel van authenticiteit van het object.

Bij het definiëren van de processen en workflows voor duurzame bewaring hoort ook de creatie van hanteerbare digitale objecten, noodzakelijk om de files en bijbehorende metadata binnen de processen te kunnen identificeren en managen. Pakketten of “packages” van metadata en digitale content worden gevormd tijdens de diverse workflowprocessen. In het algemeen hangt de typering van deze packages in een op OAIS gebaseerde informatie-omgeving af van de plaats in het proces waar deze objecten voorkomen. Het digitaal archief wordt opgedeeld in verschillende functionele processen. Er zijn processen rondom de instroom (ingest), de opslag (storage) en het verlenen van toegang (access).

De digitale objecten met de bijbehorende metadata vormen samen een Information Package (IP). De IP die wordt aangeleverd door een producer of depotgever bij de instroom heet de *Submission Information Package* (SIP). De versie van de IP die in de repository wordt opgeslagen en gepreserveerd is de *Archival Information Package* (AIP). De versie die uiteindelijk wordt uitgeleverd aan de gebruikers is de *Dissemination Information Package* (DIP). De inhoud (content en metadata) van deze verschillende typen packages kunnen verschillen : wat instroomt wordt in verrijkte vorm (bijvoorbeeld voorzien van bepaalde toegevoegde metadata) opgeslagen. En wat aan gebruikers wordt uitgeleverd is vaak maar een deel van wat wordt opgeslagen (bijvoorbeeld alleen een viewingkopie van de content, zonder alle opgeslagen metadata).

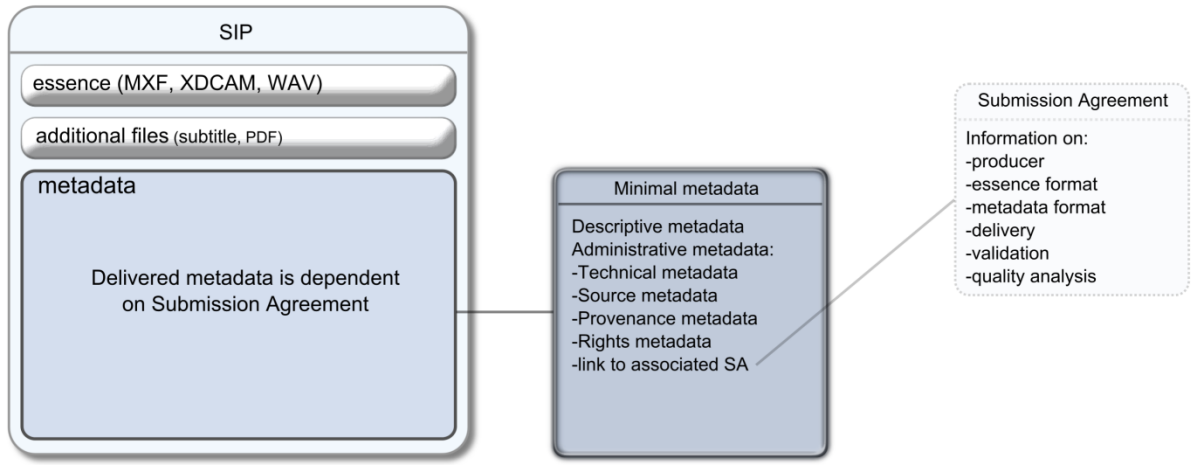

Fig. 3 De inhoud van het Submission Information Package (SIP) dat door de depotgever volgens afspraak wordt aangeboden aan het archief.

In het model worden zowel de acties die een rol spelen in de levenscyclus van het object beschreven als de eigenschappen van het object zelf. Dit heeft geresulteerd in twee verschillende soorten schema's. Het workflowschema beschrijft de levenscyclus van een object vanuit een chronologisch/lineair perspectief terwijl het objectenschema weergeeft welke metadata in welk stadium van het object aanwezig moeten zijn. Inhoudelijk zijn de verschillende objecttypes gedefinieerd als sets van bronbestanden, eventuele ondersteunende files en metadata.

In de SIP-fase is de metadata nog beperkt tot hetgeen aangeleverd is door de depotgever. Dit is in ieder geval een set van minimale beschrijvende metadata, rechtendata en eventueel aanvullende gegevens zoals informatie over de brondrager of andere technische metadata. De AIP is qua metadata het meest uitgebreid. Deze bevat zowel beschrijvende metadata die is toegevoegd door de repository zelf, als de volledige set aan preserveringsmetadata. Deze preserveringsmetadata valt uiteen in technische informatie, provenance-informatie, broninformatie en rechten. De provenance-informatie bevat alle events die een rol hebben gespeeld gedurende de levenscyclus van het object.

De DIP tenslotte, het pakket dat wordt uitgeleverd aan de gebruikers, bevat alle velden die ook in de AIP zitten. Doorgaans bestaat de DIP uit een stukje van de AIP. Immers, gebruikers zullen doorgaans geen preserveringsmetadata mee uitgeleverd willen krijgen met de file, maar vooral gericht zijn op het metadatadeel dat voor toegang en/of hergebruik belangrijk is.

Hoe de DIP er voor de verschillende gebruikersgroepen precies uitziet is afhankelijk van wat er met hen is afgesproken in zgn. *Order Agreements,* een contract tussen archief en gebruikers over de wijze van uitlevering van de files en de metadata.

Een aantal acties in het Informatiemodel zijn vanuit de Kwaliteitseisen en vanuit *best practices* gedefinieerd als een soort basis aan voorzorgsmaatregelen om een verantwoorde verwerking van een digitaal object te garanderen. De eerste versie van het Informatiemodel van Beeld en Geluid is vooral gericht op audio-en videomateriaal zoals dat aan het archief wordt aangeboden door omroepen en culturele instellingen.

Het workflowmodel werd echter gebaseerd op overkoepelende principes die ook kunnen worden toegepast op andere soorten workflows vanuit andere typen depotgevers.

Workflow

De eerste stap in de workflow vindt feitelijk plaats vóór de instroom of ingest en wordt gevormd door een onderhandelingsfase met de depotgevers van het materiaal. Deze stap resulteert in het opstellen van zgn. *Submission Agreements*, contractuele documenten waarin afspraken zijn vastgelegd over zaken als formaten, rechten, validering, metadata en foutmeldingen.

Tijdens het eigenlijke ingestproces vindt allereerst een *viruscheck* plaats, om eventuele schade aan de omgeving van het digitaal archief te voorkomen. Hierna word een *fixitycheck* gedaan, opdat gecontroleerd kan worden of het bestand goed is ontvangen. Deze check moet garanderen dat het bestand volledig en correct is afgeleverd door de depotgever. De fixitycheck kan niet plaats vinden wanneer een depotgever niet in staat is een checksum of een ander controlecijfer mee te leveren. In dat geval zal het archief - om de integriteit van de file bij navolgende acties te kunnen blijven controleren- zelf een checksum genereren, nadat het materiaal is ingestroomd. Bij het ontbreken van een meegeleverde checksum wordt dan vooraf, in de Submission Agreement, aan de depotgever gemeld dat het archief geen aansprakelijkheid aanvaardt indien materiaal tijdens de transfer bij het ingestproces corrupt is geraakt.

Na deze serie stappen wordt het formaat van het object bepaald en wordt technische metadata geëxtraheerd. Dit betreft materiaal-eigenschappen van de AV-file zoals bijvoorbeeld *aspect ratio, color space*, gebruikte *codecs* en *bitrate*. Technische metadata-extractie is met name noodzakelijk om te allen tijde overzicht te kunnen houden over de verschillende technische formaten die in het archief worden opgeslagen. Op deze wijze kunnen toekomstige risico's verbonden aan bepaalde fileformaten (bijvoorbeeld het niet meer afspeelbaar zijn), tijdig wordt gedetecteerd en kan er rekening mee worden gehouden bij de planning van migratieacties naar nieuwe bestandsformaten. Het extraheren van technische metadata is ook van belang om te kunnen verifiëren of de formaten voldoen aan de afspraken zoals gemaakt in de Submission Agreements met de depotgevers.

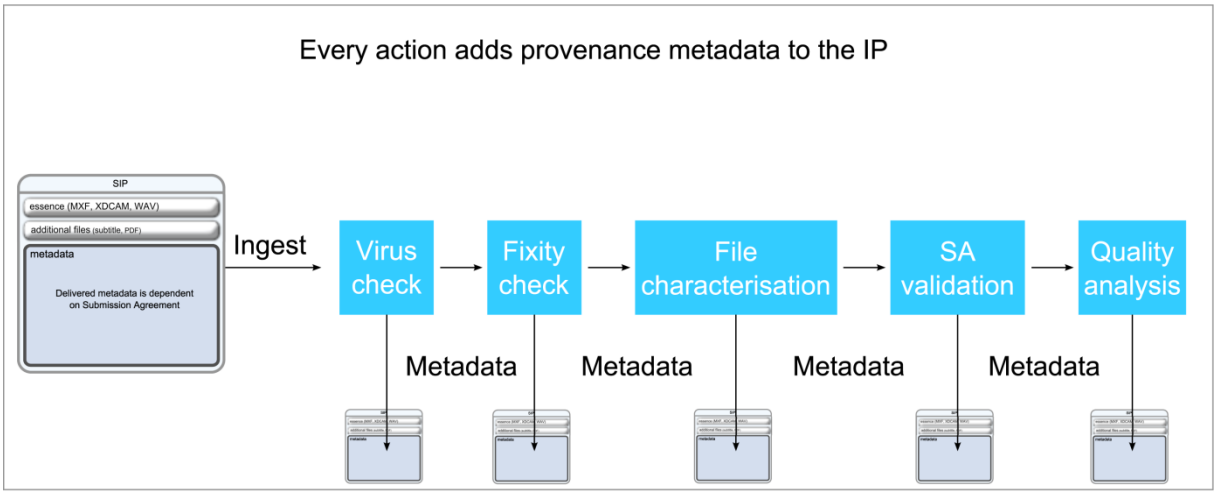

Fig. 4 De INGEST workflow van de Submission Information Package (SIP)en de diverse acties die in deze fase op het pakket worden uitgevoerd.

Optioneel onderdeel van de workflow is vervolgens de handmatige of geautomatiseerde Quality Assurance waarbij de files worden gecontroleerd op kwaliteit.

Alle ingestroomde pakketten van content en metadata krijgen tenslotte vervolgens een *persistent identifier,* een uniek label dat en permanente verwijzing vormt naar een digitaal object, onafhankelijk van zijn bewaarlocatie. Met deze stappen eindigt de ingestfase en zijn SIPs gereed voor definitieve opslag.

Samen met additionele files (zoals bijvoorbeeld ondertitels en contextinformatie over de inhoud van een bepaalde file) plus de geëxtraheerde technische metadata en de eventueel manueel toegevoegde inhoudelijke metadata wordt de SIP omgevormd tot een AIP. Dit pakket stroomt dan het opslagdomein in van de repository: de Archival Storage. Het aantal workflowstappen voor opslag is in het Informatiemodel van Beeld en Geluid beperkt. De AIP krijgt alleen nog een definitieve storagelocatie toegewezen, die wordt opgeslagen in de metadata

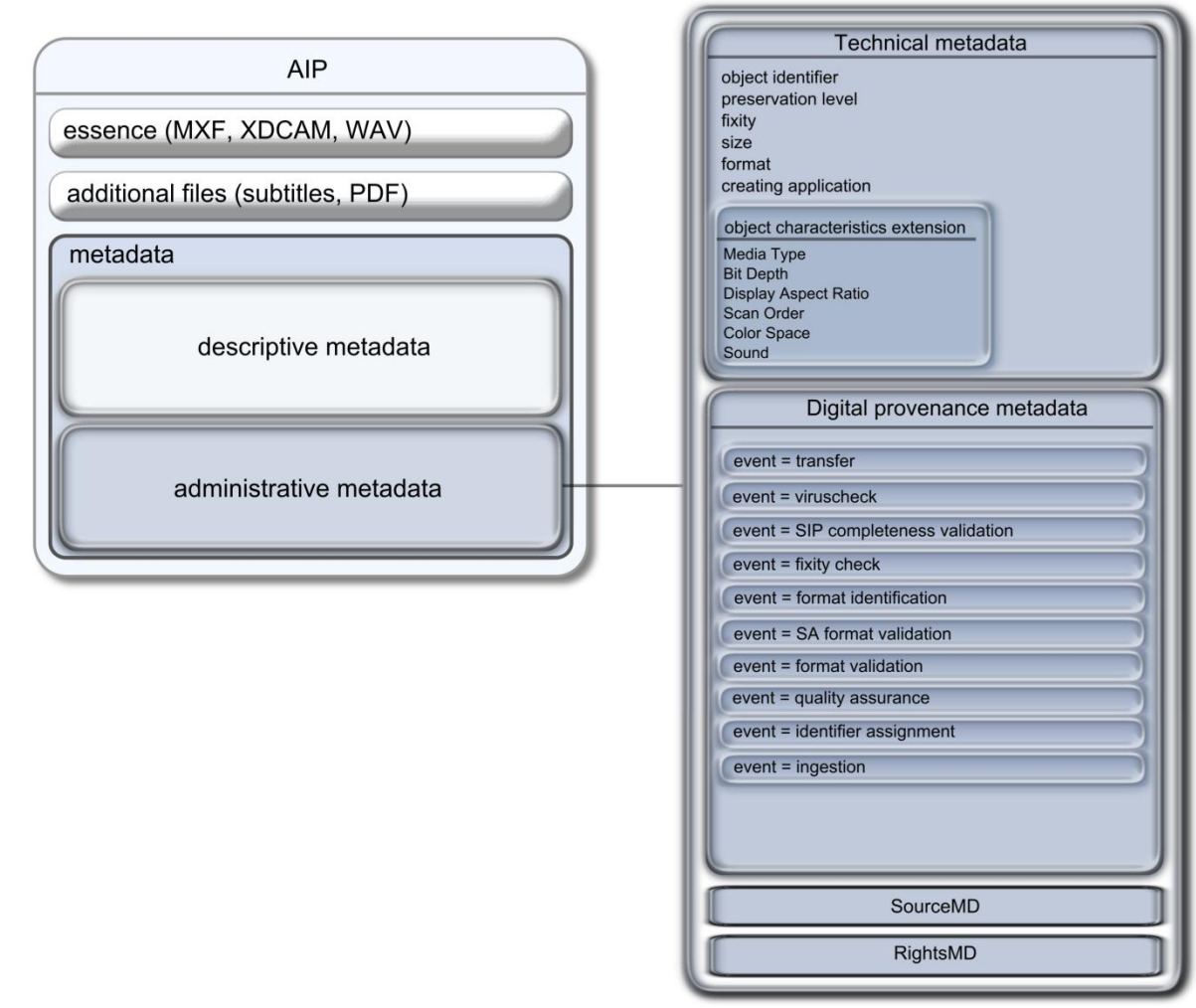

Fig 5 De inhoud van een Archival Information Package (AIP), klaar voor opslag in de repository.

Het laatste onderdeel van de workflow beschrijft de acties rond de Dissemination Information Package, de DIP. De DIP-workflow begint standaard met authenticatie van de gebruiker, waarbij wordt vastgesteld wat de gebruiker die zich op een bepaald moment aanmeldt, wel en niet mag. Hierna volgt een verzoek om een bepaald type materiaal bestemd voor een bepaald soort gebruik. Wanneer het systeem het verzoek toestaat, volgt uitlevering van de DIP. In het geval dat de *volledige* AIP wordt opgevraagd volgt een fixitycheck om te kunnen garanderen dat het materiaal goed is afgeleverd.

Wordt een *transcode* (ander formaat van file dan het opgeslagen formaat) of een *partial retrieve* (deel van een AIP, i.c. fragment van de AV-file en/of de bijbehorende metadata) aangevraagd dan wordt de checksum van deze nieuwe versie door het systeem berekend en meegeleverd met het bestand. In het Informatiemodel is de workflow generiek opgesteld zodat 'een verzoek' om uitlevering zowel om alleen 'zoeken op metadata' kan gaan als om het bestellen van een specifiek digitaal AV-file, zoals bijvoorbeeld een MXF-formaat.

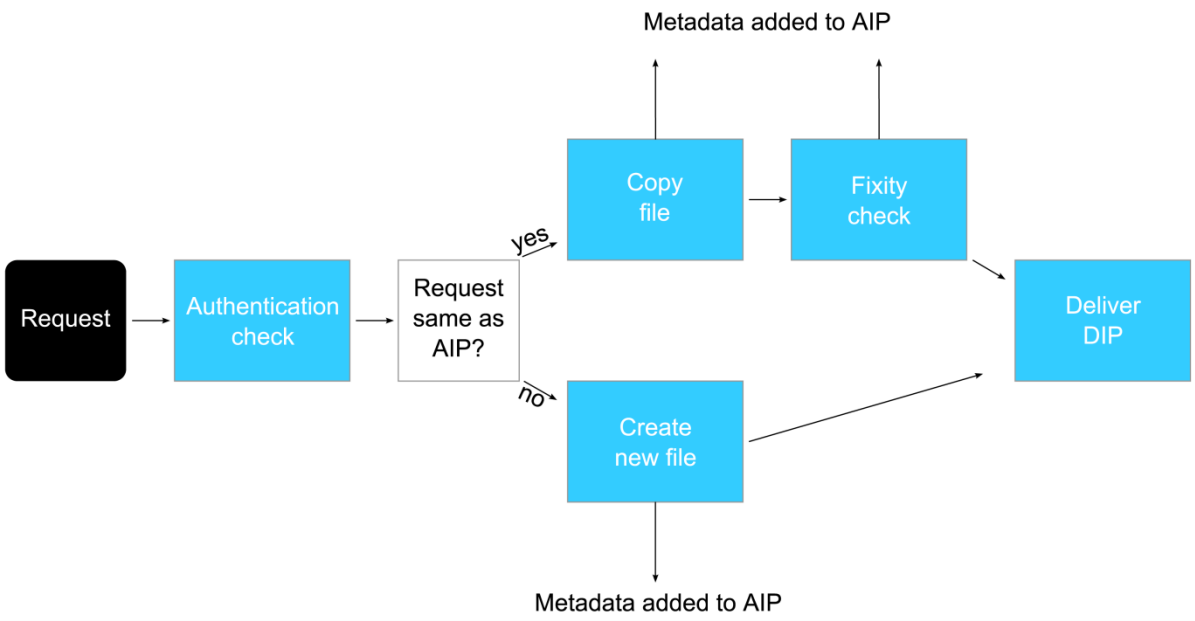

Fig 6 De ACCESS workflow van de Dissemination Information Package (DIP) nadat een gebruiker een verzoek om uitlevering heeft gevraagd.

Het ontwikkelen van een Preservation Metadata Dictionary

In de verschillende applicaties die deel uitmaken van de huidige IT architectuur van Beeld en Geluid zijn uiteraard bepaalde technische metadata aanwezig. Deze metadata worden echter niet bewust gemanaged en kunnen dus niet worden gezien als vooraf gedefinieerde, structurele component van een preserveringsysteem. Een van de opdrachten binnen het TDR project bestond uit het ontwikkelen van een Preservation Metadata Dictionary waarin zowel essentiële technische karakteristieken van AV-files zouden worden benoemd als de zgn. preserveringsmetadata , die zich vooral richten op de herkomst *(provenance*) van de files. Dit betreft die metadata die het creatieproces van de materialen vastleggen en de wijzigingsgeschiedenis ervan, samen de *chain of custody* genoemd. Rechtenmetadata , voor zover deze betrekking hebben op het recht van de respository om het materiaal te preserveren, vormen ook onderdeel van een Preservation Metadata Dictionary.

De Beeld en Geluid Preservation Metadata Dictionary bestaat uit technische metadata voor audio, video en foto's. PREMIS werd gekozen als standaard voor het vastleggen van provenance metadata en de metadata die gaat over het recht om te preserveren. De Dictionary bevat ook de definitie van de complete set events die uitgevoerd worden als onderdeel van de workflow die werd uitgewerkt in het Informatiemodel.

Voor het vaststellen van de technische metadata werden diverse AV- specifieke metadataschema's bestudeerd zoals PBCore, EBUcore, AES, LC VideoMd, AudioMD and NARA's reVTMD. Ook metadataschema's ontwikkeld voor digitale repositories binnen academische omgevingen (o.a. Rutgers University en de Universiteit van Texas) werden in de research betrokken. Dit gebeurde vooral vanwege de goed gedefinieerde woordenlijsten hier, gebaseerd op bestaande standaarden. Ook vormen deze schema's goede voorbeelden van daadwerkelijk geïmplementeerde metadatasystemen in instituten die AV-collecties beschouwen al een essentieel onderdeel van hun digitale preserveringsverantwoordelijkheid.

De grootste uitdaging bestond uit het mappen van de diverse geraadpleegde metadataschema's. Deze schema's verschilden namelijk niet alleen in de gekozen namen en definities van attributen, er waren ook afwijkingen waar het ging om het niveau van toepassing. Waar het ene schema bijvoorbeeld 'fileniveau' aangaf voor een attribuut , nam het ander voor hetzelfde attribuut het niveau van de streams.

Deze zaken maakten vergelijking en synchronisatie lastig. Daarnaast liepen de meningen over het belang van sommige metadatavelden uiteen: attributen die in sommige schema's als essentieel werden aangemerkt, kregen in andere slechts het predicaat 'nice to have'. Dit alles maakte het proces van het selecteren van de juiste attributen voor een Beeld en Geluid Dictionary nog complexer. Uiteindelijk is er een balans gevonden tussen het opnemen van ieder attribuut dat in de toekomst mogelijk belangrijk kan worden en de serie attributen waarvan we in ieder geval op dit moment het belang inzien. De Dictionary werd vervolgens zo ingericht dat nieuwe attributen in toekomstige versies eenvoudig kunnen worden ingevoegd.

Attribute of	Name	Definition	Value type Text/Numeric/Date/Binary	Obligation	Repeatable	Data Constraint	Values allowed, or link to CV-list.
Moving Image Audio	duration	The elapsed time of the entire item or track in playback	Text	M	NR	Structured form.	
Moving Image Audio	dataRate	Also known as bit rate; the rate at which data is presented within the codec. Data rate of the compressed data over time expressed in bytes per second.	Numeric	M	NR	None	
Moving Image Audio	dataRateMode	Indicates whether the stream data has been processed to achieve a fixed (constant) or variable bit rate.	Binary	M	NR	CV	Allowed values (LC): Fixed; Variable.
Moving Image	timecodeInitialValue	Starting value for timecode.	Text	M	NR	Structured form.	
Moving Image	timecodeRecordMethod	Method for recording timecode on the video source item					See also: http://rucore.libraries.rutgers.edu/open/projects/openmic/index.php?sec=guides&sub=metadata&pg=t_time-code
Moving Image	timecodeRecordType	Type of timecode recorded on video source item, e.g., SMPTE dropframe, SMPTE nondropframe, etc..	Text	M	NR	CV	longitudinal (LTC); vertical interval (VITC); Other

Fig.7 Uittreksel uit de Beeld en Geluid Preservation Metadata Dictionary V1.0 : voorbeelden van technische metadata van AV files.

De huidige Beeld en Geluid Preservation Metadata Dictionary bevat nog geen uitgebreide metadata voor het vastleggen van processen voor het opnieuw formatteren van materiaal , d.w.z. voor files die zijn gecreëerd als onderdeel van een digitaliseringsproject.

PREMIS biedt weliswaar ruimte aan metadata m.b.t. een zgn. 'creating application' en voor bepaalde andere informatie die met het ontstaan van een file samenhangt, deze standaard omvat niet het complete scala aan *reformatting* metadata van bijvoorbeeld de NARA's reVTMD of de AES 57-2011 schema's (N.B. om in deze lacune te voorzien heeft NARA haar schema in feite ontwikkeld).
Deze schema's bieden beide meer granulariteit en detail waar het gaat om bijvoorbeeld de transferapparatuur, kalibratie, de naalden die werden gebruikt bij disctransfers etc.

Deze uitgebreide herkomst-informatie, in combinatie met de link naar de technische kenmerken van de originele, analoge bron wordt door sommige deskundigen onmisbaar geacht voor het kunnen vaststellen van de authenticiteit van een object door gebruikers.[3]

Tenslotte moet er voor de Beeld en Geluid Dictionairy ook nog een uitgebreidere set technische metadata voor analoge dragers worden ontwikkeld . Ook hier bieden zowel de NARA's reVTMD als de AES 57-2011 standaard oplossingen. Een metadata -expert van de Rutgers Universiteit ziet goede mogelijkheden om de AES 57-2011 standaard zodanig uit te breiden dat alle multimedia-dragers kunnen worden opgenomen en werkt momenteel aan de implementatie hiervan.[4]

Attribute of	Name	Definition	Value type Text/Numeric/Date/Binary	Obligation	Repeatable	Data Constraint	Values allowed, or link to CV-list.
Moving Image Audio Photo	**1.3.1 preservationLevelValue**	A value indicating the set of preservation functions expected to be applied to the object.	Text	M	NR	CV	Example values for R: bit-level; full; 0; 1; 2. For File: bit-level; full; 0; fully supported with future migrations.
Moving Image Audio Photo	**1.3.2 preservationLevelRole**	A value indicating the context in which a set of preservation options is applicable. Repositories may assign preservationLevelValues in different contexts which must be differentiated, and may need to record more than one context	Text	O	NR	CV	Example values: requirement; intention; capability.
Moving Image Audio Photo	**1.3.3 preservationLevelRationale**	The reason a particular preservationLevelValue was applied to the object.	Text	O	R	None	
Moving Image Audio Photo	**1.3.4 preservationLevelDateAssigned**	The date, or date and time, when a particular preservationLevelValue was assigned to the object.	Date	O	NR	Structured form.	
Moving Image Audio Photo	**1.5.2 Fixity**	Container element for holding information used to verify whether an object has been altered in an undocumented or unauthorized way.	Text	O	R	Container	
Moving Image Audio Photo	**1.5.2.1 messageDigestAlgorithm**	The specific algorithm used to construct the message digest for the digital object.	Text	M	NR	CV	Example values for F: MD5; Adler-32; HAVAL; SHA-1; SHA-256; SHA-384; SHA-512; TIGER; WHIRLPOOL.
Moving Image Audio Photo	**1.5.2.2 messageDigest**	The output of the message digest algorithm.	Text	M	NR	None	
Moving Image Audio Photo	**1.5.2.3 messageDigestOriginator**	The agent that created the original message digest that is compared in a fixity check.	Text	O	NR	None	
Moving Image Audio Photo	**1.5.3 Size**	The size in bytes of the file or bitstream stored in the repository. to indicate the storage requirements or file size of a digital media item. As a standard, express the file size in bytes. - use attribute 'unit of measurement'	Numeric	O	NR	Integer	

Fig.8 Uittreksel uit de Beeld en Geluid Preservation Metadata Dictionary V1.0 : voorbeelden van preservation metadata gebaseerd op PREMIS.

[3] Otto, Jane Johnson(2010) 'A Sound Strategy for Preservation: Adapting Audio Engineering Society Technical Metadata for Use in Multimedia Repositories', Cataloging & Classification Quarterly, 48: 5, 403 — 422
[4] Ibid.

Belangrijke bevindingen

Synchroniteit met de praktijk

OAIS is een abstract model dat vanuit de specifieke situatie binnen het archief vorm moet krijgen. Het is geen 'alles of niets'-concept. De kwaliteitseisen zijn eisen op hoofdlijnen en bevatten veel mogelijke, uiteenlopende oplossingsrichtingen.

Dit maakt het model aan de ene kant genuanceerd en flexibel maar zorgt er anderzijds voor dat op sommige vragen niet eenduidig geantwoord kan worden. Hetgeen haalbaar en wenselijk is binnen de eigen archiefsituatie moet uiteindelijk bepalen hoe de eisen worden uitgewerkt tot specifieke oplossingen.

Bij Beeld en Geluid liep tijdens het TDR-project gelijktijdig een traject voor de aanschaf van een nieuw MAM (Media Asset Management) systeem. Om in dit traject zoveel mogelijk eisen voor OAIS compliancy te kunnen verwerken , zijn de workflowstappen zoals die in het Informatiemodel waren vormgegeven in versnelde vaart uitgewerkt tot zeer gedetailleerde en concrete eisen die aan een leverancier van MAM-systemen voorgelegd zouden kunnen worden. Het domein van een MAM – systeem is breder dan alleen preservering en dus moesten ook in kwaliteitseisen van buiten de IP-workflow in kaart gebracht worden.

Tijdens dit proces werd duidelijk dat de kwaliteitseisen die opgesteld zijn vanuit het OAIS-model weinig houvast bieden voor concrete en gedetailleerde technische oplossingen. Ook is uit deze eisen niet altijd duidelijk te halen of iets in een workflow opgelost moet worden of door middel van een technische eis aan een systeem. Uiteindelijk is er voor gekozen om per eis een keuze te maken tussen de verschillende opties. Daarmee ontstond een eerste, beperkte lijst met OAIS compliant life cycle requirements aan een MAM-systeem.

Preservering in de Enterprise architectuur

Bij het stroomlijnen van de OAIS-eisen met de requirementstrajecten die liepen in de praktijk, is duidelijk geworden dat niet alle kwaliteitseisen voor duurzame preservering vervuld moeten of kunnen worden binnen één systeem. Het is realistischer is om de invulling van bepaalde eisen op te delen naar verschillende onderdelen van de totale Enterprisearchitectuur.

Voor dit doel is meer inzicht vereist in zowel onderscheid tussen, als verwevenheid van de soorten (meta) data , alsmede van de samenhang tussen de verschillende systemen, workflows en functies binnen die Enterprise architectuur in relatie tot OAIS-compliant *digital life cycle management*. Voor Beeld en Geluid zijn hier nog wel wat stappen te zetten.

Voor de transformatie van de analoge AV-archiefwereld naar digitale preservering binnen een IT domein is het van belang de processen te bekijken vanuit een Enterprise IT architectuur perspectief. Een dergelijk perspectief richt zich op het complete life cycle management van de data/content in het archief, i.c. op alle business activiteiten die op de data moeten worden toegepast. In de TDR context zijn dit dan specifiek die processen die lange termijn preservering en duurzame toegankelijkheid van het object zeker moeten stellen.

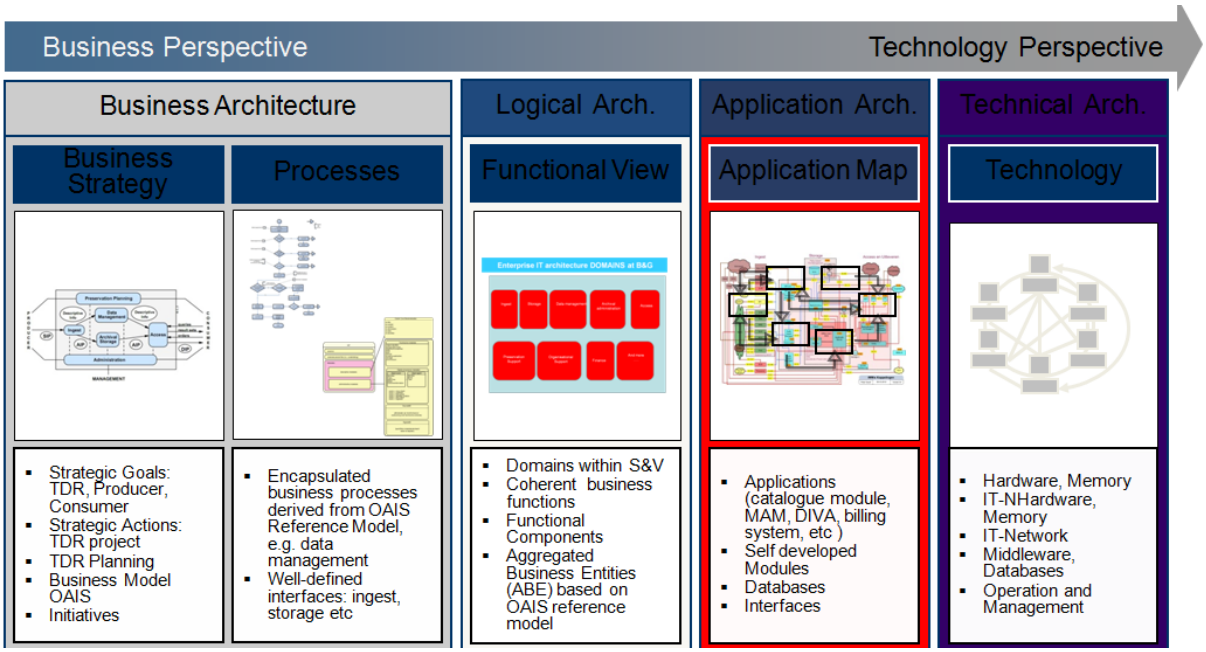

Fig. 9 De onderscheiden niveaus binnen de Enterprise-architectuur van Beeld en Geluid en hun corresponderende functies en domeinen.

Een Enterprise architectuurperspectief op een digitaal archief vereist een duidelijk begrip van het onderscheid tussen IT processen en de processen die samenhangen met de "business" van preservering. Het kan daarbij verwarrend werken dat er een overlap bestaat tussen bepaalde functionele domeinen binnen OAIS en IT processen. Als voorbeeld kan de OAIS-functie 'Administration' (i.c. de verzamelde beheersfuncties van het digitale archief) dienen waaronder issues als beveiliging en risicomanagement vallen, beiden ook vast onderdeel van 'regulier' IT procesmanagement. Vragen die hier vanuit beide domeinen worden gesteld zijn : is de beveiliging in orde? Is de data integer? Wordt de opslagcapaciteit adequaat gemanaged ? etc. IT procesmanagement richt zich hier echter uitsluitend op de juiste werking van specifieke applicaties binnen de infrastructuur en beschouwt deze zaken niet vanuit het oogpunt van duurzaam behoud van de collecties.

De nauwe relatie tussen IT en preserveringsfuncties kan leiden tot het gelijkstellen van 'opslag' aan 'preservering'. Maar in feite is 'opslag' niet meer dan een set van IT applicaties binnen een IT architectuur, waarin zaken aan de orde komen als het schrijven en registreren van de data, het behouden van de data voor een bepaalde periode en het lezen van en toegang verschaffen tot de data. Geavanceerde opslag bestaat kan nog bestaan uit het dupliceren van de files en technieken omvatten als RAID, erasure coding, het genereren van meerdere kopieën op verschillende media en/of verschillende locaties (georeplication) en de optimalisering van het 'fysiek' plaatsen van data zoals in hierarchical storage management (HSM) systemen. Een preserveringsysteem echter, omvat ook *alle* processen die plaatsvinden gedurende de instroom (ingest) en de beschikbaarstelling (access) van het materiaal, nl. de acties die op de data/content worden uitgevoerd met als doel de uiteindelijke preservering te ondersteunen. Preservering derhalve is een proces dat weliswaar ook bestaat uit opslag, maar dat daartoe geenszins beperkt is. [5]

Vanuit het niveau van de Enterprise IT architectuur moet een IT-strategie worden ontwikkeld, die is gedacht vanuit de businessprocessen van een organisatie (in dit geval preservering) en die vervolgens wordt vertaald naar functionele IT domeinen.

[5] PP D2.3.1
Preservering van digitale AV-Collecties volgens de OAIS Standaard, A. de Jong, B. Delaney. D. Steinmeier
Mei 2014

Daarna moeten er applicaties worden ingericht waarbinnen deze businessprocessen gegarandeerd worden uitgevoerd . In het TDR project werd een eerste aanvang gemaakt met het beschrijven van deze processen.

Kennisopbouw, communicatie en samenwerking

Al bij de start van het TDR project werd duidelijk dat digitale preservering niet ophoudt bij de IT – afdeling van Beeld en Geluid; bij het implementeren van op OAIS gebaseerd lifecycle management is het hele archiefinstituut betrokken: bij de instroom, gedurende de opslag, tijdens metadataverrijking en bij het aanbieden aan gebruikers. Dit houdt in dat iedere afdeling van de organisatie feitelijk een eigen rol en verantwoordelijkheid heeft in het lange termijn behoud en het garanderen van toegang tot de collecties: de acquisitie- en selectiemedewerkers, de catalogiseerders even goed als de mensen van de klantenservice. OAIS als uitgangspunt maakt ook de belangrijke relatie tussen archief en depotgevers duidelijk en benadrukt de band met de gebruikers, de Designated Communities: alle instroom en disseminatie-activiteiten van beeld en geluid spelen immers een integrale rol in de archivale levenscyclus van de materialen.

Een belangrijk nevendoel van het TDR project was dan ook het verspreiden van kennis over digitale preservering in de organisatie. Er werd gekozen voor een projectvorm voor iets wat in wezen beleidsvorming is. Het opzetten van betrekken van medewerkers uit verschillende afdelingen binnen het archief zou de eigen rol en verantwoordelijkheid in digitale preservering en life cycle management kunnen verduidelijken en versterken.

Dit doel is maar ten dele gehaald. Voor de meeste projectgroepleden bleek het niet mogelijk zich naast hun dagelijks werk intensief te verdiepen in de literatuur en de theorie teneinde deze te kunnen vertalen naar normatieve beleidsdocumenten die konden worden toegepast op de eigen organisatie. Het werk is feitelijk geheel uitgevoerd door de afdeling Informatiebeleid van Beeld en Geluid, die hier toch al mee bezig was.

De projectvorm heeft wel geleid tot een bredere bewustwording van de diverse aspecten van preservering en digital lifecycle management. Men is zich er inmiddels van bewust wat dit voor begrippen zijn en waarom ze belangrijk zijn voor het behoud van de collecties. Ook is de OAIS standaard en de bijbehorende terminologie voor gegevens en processen niet meer voor iedereen vreemd. Daarnaast zijn bepaalde belangrijke technische concepten, zoals validatie, fixity checking en checksums, meer gaan leven.

Het is nu van groot belang dat de kennis en het inzicht opgedaan tijdens het OAIS-traject gaan dienen als referentie voor de IT afdeling van Beeld en Geluid. Deze afdeling moet immers de preserverings-businessprocessen die in het Informatiemodel zijn uitgewerkt, daadwerkelijk in de systemen gaan ondersteunen. De vereisten voor integraal life cycle management van digitale AV-data zoals die nu zijn opgesteld, maken immers duidelijk dat het noodzakelijk is om eerst alle processen en workflows te mappen aan functionele gebieden binnen de IT systeem architectuur, alvorens te bepalen wat de beste applicaties zijn om deze functies te gaan uitvoeren. Zo wordt voorkomen dat teveel wordt gewerkt vanuit een geïsoleerd applicatieperspectief in plaats van te beginnen vanuit een hoger abstractieniveau waarin eerst de business, de processen en objecten zijn benoemd. Op die manier kunnen zowel ontwerp, bouw als implementatie van een archiefomgeving die is gebaseerd op OAIS plaatsvinden op basis van samenwerking en gezamenlijke inzichten van alle betrokken afdelingen binnen het instituut.

Tenslotte

Na het afronden van de fase van het opstellen van requirements voor OAIS compliant AV-archivering resteren nog essentiële vragen en issues. Behalve het uitwerken van de Enterprise architectuur, waarbij moet worden bepaald waar in de systemen welke preserveringsfuncties zullen landen, moet ook worden begonnen aan het modelleren van de kosten van digitale preservering. Hoe duur is het nu geschetste workflow- en informatiemanagementscenario eigenlijk ? En wie gaat er betalen voor welke preserveringsvoorziening in de digital life cycle management omgeving? Het archief zelf, als Trusted Digital Repository? De producer/depotgever als aanleveraar van de digitale collecties of misschien wel de Designated Communities, voor wie alle files immers duurzaam toegankelijk en dus permanent technisch up-to-date gehouden moeten worden?

Verder is er nog de lastige kwestie van de eerder ingestroomde files en metadata : hoe gaat het archief er bijvoorbeeld voor zorgen dat de technische metadata die automatisch zijn gegenereerd tijdens digitaliseringsprojecten in voorgaande jaren (de zgn. 'dark' metadata) deel uit gaan maken van de officiële verzameling preserveringsmetadata, opdat ze gemanaged kunnen worden? Wat te doen met het checken en valideren van alle eerder *niet* OAIS-compliant ingestroomde materialen , samen nu zeker 400.000 uur ? Ook de preserveringsbusiness processen zèlf zullen nader moeten worden uitgewerkt. Gaan we de normatieve workflow en metadata toepassen op *alle* collecties die binnenkomen of doen we dat gedifferentieerd ? En *hoe* dan: wordt er bijvoorbeeld onderscheid gemaakt tussen preserveringsniveau's voor omroepproductiemateriaal en voor cultuurhistorisch erfgoed van nationaal belang? En gaan we op *alle* ge-ingeste soorten media typen (metadata, foto's, contextdocumenten etc.) hetzelfde hoge preseveringsniveau toepassen of geldt dit alleen de kerncollecties bewegend beeld en geluid?

Value	**Fixity Check**		
Definition	**The process of verifying that an object has not been changed in a given period.**		Note: This is the comparison of two message digest calculations.
Semantic unit	Semantic component	Sample value(s)	Notes
eventIdentifier	eventIdentifierType	NIBGPS0.3	domain within which id is unique
eventIdentifier	eventIdentifierValue	E634721158	
eventType	none	fixity check	
eventDateTime	none	2013-01-16T020:20:32+01:00	
eventDetail	none		
eventOutcomeInformation	eventOutcome	{pass;fail}	
eventOutcomeDetail	eventOutcomeDetailNote		link to report if failed
linkingAgentIdentifier	linkingAgentIdentifierType	NIBGPS 0.3	domain within which id is unique
linkingAgentIdentifier	linkingAgentIdentifierValue	A282179	id of agent entity which gives name, version and type of agent performing fixity check; agentName="MD5Deep", version=3.6 ; agentType=sw program
linkingAgentRole	none	Executing Program	
linkingObjectIdentifier	linkingObjectIdentifierType	NIBGPS0.3	domain within which id is unique
linkingObjectIdentifier	linkingObjectIdentifierValue	549-274io0	id of file being verified

Fig. 10 Een brandende vraag: moeten alle materialen die in voorgaande jaren zijn ge-ingest in het archief van Beeld en Geluid alsnog het complete OAIS-compliant proces van fixity checking ondergaan?

Voor een rationale en kosteneffectieve archivale bedrijfsvoering zullen deze vragen op termijn exact beantwoord moeten worden en moeten worden vertaald naar beleid . Met het Informatiemodel en de andere normatieve documenten van het TDR project is nu al een belangrijk referentiekader opgetrokken. Aan deze documenten kan Beeld en Geluid afmeten hoe het als AV- archief opereert waar het gaat om het bewust omgaan met de ingest, de opslag en de beschikbaarstelling van de collecties die het onder zijn hoede heeft , of het nu gaat om omroeproductiemateriaal of om cultureel erfgoed. De gedocumenteerde TDR projectresultaten leggen een solide , theoretische basis voor de technische en organisatorische inrichting van een OAIS-compliant preserveringsomgeving in het AV domein. De rollen van alle betrokkenen in het AV-preserveringsproces zijn vastgelegd en afgebakend.

Beeld en Geluid weet nu hoe het volgens de standaarden zou moeten en kan dus - bij inrichting en implementatie van de systemen- aangeven waar van deze standaarden wordt afgeweken en waarom. Door dit alles is de organisatie beter in staat verantwoording af te leggen tegenover depotgevers/producers en tegenover de gebruikersgroepen, een van de basisvoorwaarden van het duurzaam en betrouwbaar werken .

De eerste stap op weg naar de status van Trusted Digital Repository is hiermee gezet. Op basis van het verrichtte beleids- en modelleringswerk binnen het TDR project komt Beeld en Geluid nu al in aanmerking voor een bepaald type certificering voor digitale archieven, te weten het zgn. Data Seal of Approval (objectief Europees keurmerk voor data-archieven). De normatieve documenten die zijn opgeleverd kunnen dienen als goede inhoudelijke bewijsstukken voor de sommige eisen die de DSA stelt. Het DSA zal dan ook binnenkort door Beeld en Geluid worden aangevraagd.

Referenties

1. Kwaliteitseisen Digitaal Archief Beeld en Geluid V1.0
2. Preservation Metadata Dictionary Beeld en Geluid V 1.2.
3. Designated Communities Beeld en Geluid, typering en uitlevereisen V1.0
4. Handleiding voor het maken van Submission Agreement en Order Agreement V1.0
5. Informatiemodel Digitaal Archief Beeld en Geluid V1.0
6. Handleiding ontwikkeling preserveringsstrategieën en preserveringplan V 1.0
7. Storage binnen OAIS: normatief model en gapanalysis V1. 0
8. Risk Analysis Mechanismen V1.0
9. Information Security Digitaal Archief Beeld en Geluid V0 .9
10. Disaster Recovery Digitaal Archief Beeld en Geluid V0.9

Over de auteurs

Annemieke de Jong werkt als sr. Policy Advisor Digital Preservation bij het Nederlands Instituut voor Beeld en Geluid. Zij is verantwoordelijk voor het opzetten van strategisch beleid voor digitaal AV-collectiemanagement en AV-preservering in het media- en erfgoeddomein. De Jong houdt zich ook bezig met het ontwikkelen en (inter)nationaal aanbieden van kennis en expertise op bovengenoemde gebieden en heeft meerdere publicaties op haar naam staan. Annemieke de Jong is lid van de Media Management Commission van FIAT/IFTA, de internationale vakfederatie voor AV-archieven . Zij is afgestudeerd in Nieuwe Media en Digitale Cultuur en Informatiemanagement met een specialisatie in digitale archieven. **Beth Delaney** houdt zich al zo'n 25 jaar bezig met eisen aan collectie-management systemen, het implementeren van metadatastandaarden en het ontwikkelen van beleid in Amerikaanse en Europese AV-archieven . Zij heeft ervaring in zowel analoog als digitaal collectiemanagement bij omroeparchieven en in erfgoeinstellingen. Als consultant heeft Delaney zich gespecialiseerd in digitale preserveringsprocessen, preserveringsbusiness-requirements en preserveringsmetadata in het AV-domein. Delaney heeft een master's titel in Bibliotheek- en Informatiewetenschappen, met een specialisatie in archieven. **Daniel Steinmeier** is als technisch specialist werkzaam bij de afdeling Applicatiebeheer van Beeld en Geluid. Hij heeft veel ervaring in het via diverse webplatforms toegankelijk maken van AV-materiaal voor educatieve doeleinden. Metadatastandaarden, waaronder IEEE-LOM en uitwisselingsprotocollen zoals OAI-PMH- hebben van aanvang een belangrijke rol gespeeld in zijn werk. Momenteel is Steinmeier bij Beeld en Geluid o.m. belast met het coördineren van digitale instroomprocessen. Steinmeier heeft een master's titel in Taal- en Cultuurstudies met media als specialisatie.

De Jong, Delaney en Steinmeier vormden gedurende 2012, 2013 en 2014 de kern van het **Informatiemanagementeam** van Beeld en Geluid dat verantwoordelijk was voor de ontwikkeling van alle criteria, modellen en normatieve beleidsdocumenten, noodzakelijk voor de implementatie van OAIS-compliant business processen bij het instituut. Momenteel wordt gewerkt aan de ontwikkeling van een Preservation Metadatabase, voor het managen van de preservation metadata die wordt gegenereerd tijdens de verschillende stadia in het archivale proces.