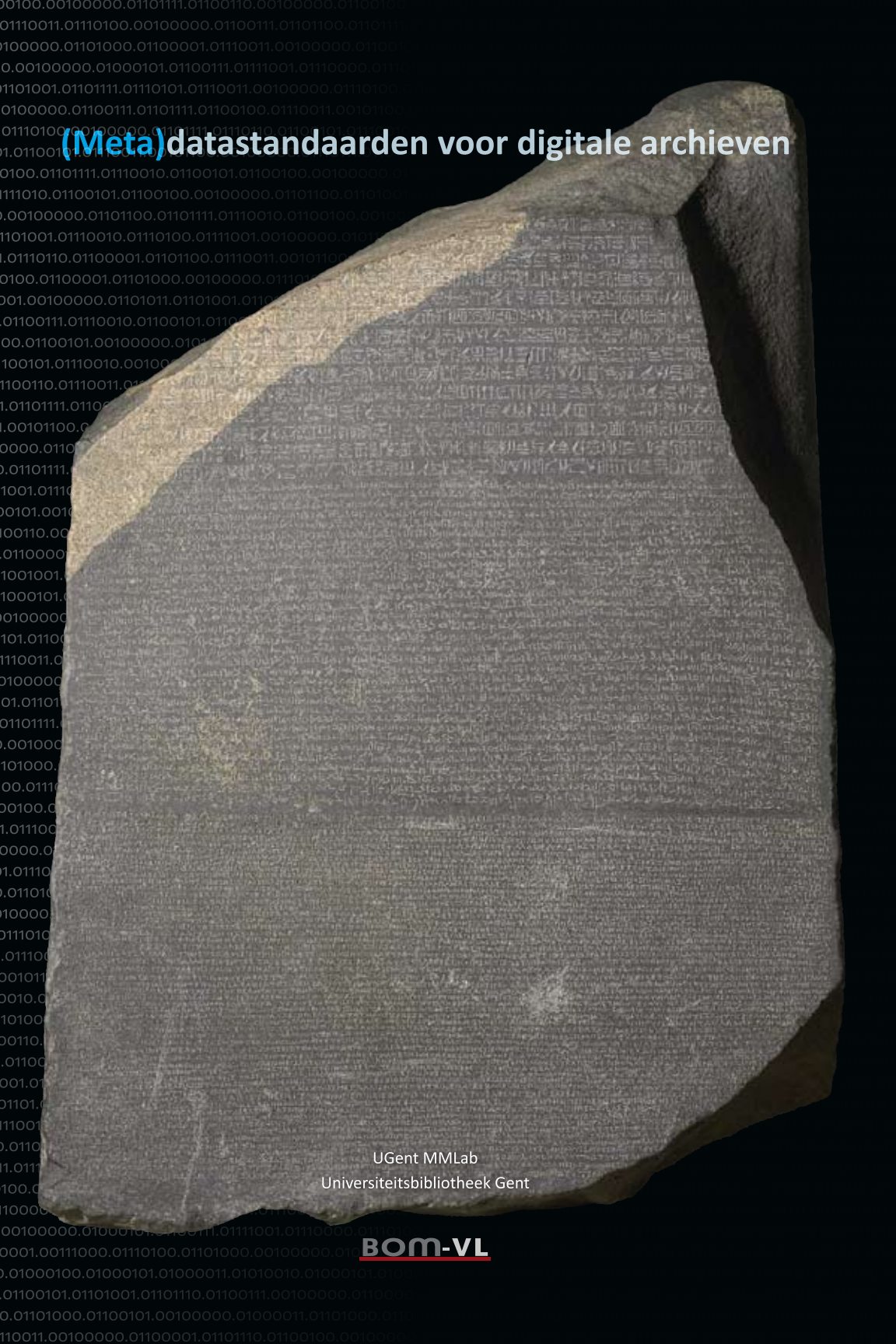
(Meta)datastandaarden voor digitale archieven



UGent MMLab
Universiteitsbibliotheek Gent

BOM-VL

(Meta)datastandaarden voor digitale archieven

BOM-Vlaanderen

(Meta)datastandaarden voor digitale archieven

IBBT
exploring the information society

MULTIMEDIA

UNIVERSITEITS BIBLIOTHEEK GENT

Met steun van de Vlaamse overheid

Onder redactie van:
Rik Van de Walle,
Sylvia Van Peteghem

Teksten:
Paul Bastijns,
Sam Coppens,
Siska Corneillie,
Patrick Hochstenbach,
Erik Mannens,
Liesbeth Van Melle

Tekstredactie:
Yana-Frauke Vandendriessche,
Liesbeth Van Melle,
Inge Van Nieuwerburgh

Cover:
The Rosetta Stone .24, no. AN16456004 © Trustees of the British Museum

Vormgeving:
Het Métier: Reinhilde Meyts

Copyright:
© 2009 Universiteitsbibliotheek Gent
Deze uitgave wordt ter beschikking gesteld overeenkomstig de bepalingen van de Creative commons Public License, Naamsvermelding-Niet-commercieel-Gelijk delen 2.0 België, http://creativecommons.org/licenses/by-nc-sa/2.0/be/deed.nl

ISBN: 9789052230009
EAN: 9789052230009
D/2009/376/1

Handle: http://hdl.handle.net/1854/LU-480734

Voorwoord

Volgens UNESCO bezit onze planeet al enkele honderden miljoenen uren aan audiovisuele archieven. In Vlaanderen bezitten niet alleen de omroepen maar ook culturele organisaties, privépersonen en overheidsinstellingen duizenden uren aan spraak- en beeldmateriaal, dat op analoge dragers is opgeslagen. Op dit ogenblik boeten die continu aan kwaliteit in, waardoor de toegankelijkheid in gevaar komt en de banden op termijn niet langer af te spelen zullen zijn. Een oplossing voor dit probleem is digitale ontsluiting en bewaring van de inhoud. Daarnaast worden dagelijks ontelbare digitale opnames gemaakt en die wil men net als een oude foto, binnen enkele tientallen jaren nog steeds kunnen raadplegen. Langetermijnbewaring van dit digitaal erfgoed staat dan ook hoog op de agenda. Om dat te kunnen verwezenlijken, zijn goede afspraken nodig over hoe de data gestructureerd moeten zijn om ze in de toekomst toegankelijk te houden en deel te laten worden van het Vlaamse erfgoed.

Het bewaren van het audiovisueel erfgoed, inclusief een moedige en onderbouwde selectiepolitiek, is een enorme uitdaging. De laatste decennia wordt digitalisering vaak als oplossing naar voren geschoven voor de conserveringsproblematiek. Bovendien schakelt al wie vandaag audiovisuele cultuurdragers creëert, van omroepen tot amateurs, over op een digitaal productieproces, wat de aangroei van digitaal erfgoed exponentieel doet toenemen. Maar gedigitaliseerd of origineel digitaal materiaal, de beheerder moet altijd rekening houden met de duurzaamheid. De snelheid waarmee IT-infrastructuur veroudert, is immers berucht. Willen we voorkomen terecht te komen in een ‘digital dark age’, waarin bestanden moeilijker af te lezen zijn dan hun analoge voorgangers, dan moeten we aansluiten bij de internationaal afgesproken normen van de digitale duurzaamheid inzake compressie, opslag, bewaring en ontsluiting. Met andere woorden, digitalisering verlost de erfgoedbeheerder niet van de klassieke vraagstukken met betrekking tot behoud en beheer. Net als analoog materiaal hebben ze een fysieke plaats en de inrichting en het onderhoud daarvan moeten de nodige aandacht krijgen. Anders zou de digitalisering wel eens heel ontnuchterend kunnen uitvallen. Vervolgens heeft ook een digitaal archief een regelmatig onderhoud nodig om de toegankelijkheid en de leesbaarheid te controleren. Bovendien verlost digitaliseren ons niet van de nood om kostbare analoge formaten te beheren.

Adequate ontsluiting van een digitaal archief is een tweede uitdaging. Voor cultuur, onderwijs, toerisme, wetenschap en het geïnteresseerde publiek moeten er verschillende modellen worden uitgewerkt om het audiovisueel

erfgoed aan te bieden. Hoe meer context daarbij kan worden meegegeven, hoe beter het systeem wordt ervaren.

Een toekomstig bewarings- en ontsluitingsmodel moet ook rekening houden met specifieke noden van enerzijds de mediasector en anderzijds de cultuur- en erfgoedsector. Iedere sector heeft immers een eigen productie, ontsluitings- en bewaarcultuur die gerespecteerd moeten worden.

Naast de culturele waarde van het audiovisuele erfgoed mag ook de aanzienlijke economische waarde niet onderschat worden. Door het digitaal beschikbaar stellen van dit materiaal wordt hergebruik gestimuleerd en kunnen innovatieve toepassingen op het gebied van nieuwe media leiden tot de ontwikkeling van hoogwaardige diensten voor het grote publiek. Ook nieuw te ontwikkelen lesmateriaal in het onderwijs zou deze multimediale toegang tot het culturele erfgoed perfect kunnen gebruiken.

Het snelle verval van het audiovisuele erfgoed enerzijds en de maatschappelijke, culturele en economische waarde ervan anderzijds vragen een oplossing op korte termijn. Het BOM-vl-project ('Bewaring en Ontsluiting van Multimediale Data in Vlaanderen') biedt een belangrijke aanzet tot een antwoord op al deze vragen en noden. Diverse actoren, actief binnen de geschetste problematiek in de Vlaamse archief-, bibliotheek-, cultuur-, erfgoed-, en mediasector, zijn partner in het project en denken mee vanuit de eigen concrete aandachtspunten.

Alleen met een degelijke langetermijnvisie kunnen we ervoor zorgen dat het audiovisueel erfgoed in Vlaanderen voor de toekomst wordt bewaard en dat de mogelijkheden van de ontsluiting, die bijdraagt aan een internationale uitstraling, ten volle worden benut.

Het succes en de impact van het BOM-vl-project zullen sterk afhangen van de mate waarin de verschillende actoren uit de Vlaamse cultuur-, erfgoed- en mediasector hun kennis en expertise met betrekking tot archivering en ontsluiting van digitale multimediale data met elkaar zullen delen en op elkaar zullen afstemmen. Hierbij wordt niet alleen gedacht aan de partners die deel uitmaken van het projectconsortium; het is de expliciete bedoeling van deze partners om alle resultaten die in het project worden behaald, publiek beschikbaar te stellen voor alle spelers in het Vlaamse (en Europese) cultuur-, erfgoed- en mediaveld.

Deze studie biedt een vergelijkend overzicht van opslag- en compressieformaten die momenteel beschikbaar zijn of gebruikt worden binnen de

cultuur- en erfgoedsector, met inbegrip van richtlijnen voor het gebruik van geschikte formaten bij ontsluiting. Het is een intersectoraal baken om het hoe en waarom van alle huidig gebruikte (metadata)formaten te begrijpen. Aan de hand van een paar praktijkvoorbeelden worden ten slotte een aantal conclusies geformuleerd die het gebruik van een overkoepelend gelaagd metadatamodel verantwoorden. We hebben immers geleerd dat een sectoroverschrijdende persistente link tussen multimediale data en de bijbehorende metadata niet alleen wenselijk, maar broodnodig is!

prof. dr. ir. Rik Van de Walle,
hoofd UGent/IBBT-MMLab

dr. Sylvia Van Peteghem,
hoofdbibliothecaris Universiteitsbibliotheek Gent (Boekentoren)

1 Inhoud

Figurenlijst

2 Inleiding en probleemstelling

Door onze fysieke multimediale archieven te digitaliseren en in elektronische vorm op te slaan en te bewaren, meent men dat ons cultureel erfgoed gevrijwaard is van gevaren zoals diefstal, verlies, beschadiging, branden en andere natuurrampen. Als het 'papieren' archief in rook zou opgaan, kan men immers steeds terugvallen op een versie van het archief dat - liefst op meer dan één server - digitaal in veiligheid is gebracht. Behalve de garantie op bewaring biedt een digitaal archief uiteraard nóg tal van voordelen. Ruimtelijke grenzen zijn opgegeven, mobiliteit vormt geen drempel meer, opzoekingen leveren een enorme tijdwinst op in vergelijking met het 'ter plaatse' raadplegen van fysieke materialen en documenten, enzovoort. Toch is een dergelijk digitaal archief niet onkwetsbaar. Meer nog, een minimale veroudering van bepaalde apparatuur bijvoorbeeld kan catastrofale gevolgen hebben. In het licht van de nauwelijks bij te houden evolutie op het gebied van informatietechnologie is de preservering van ons digitaal erfgoed een brandend actuele problematiek. Een oplossing hiervoor is stringent, wil men vermijden dat de miljoenen terrabytes aan digitaal materiaal – en de daarmee gepaard gaande investeringen aan tijd en geld – waardeloos worden.

Indien men digitale multimediale data wil preserveren, moeten de archiefomgevingen waarin ons digitaal erfgoed bewaard moet worden aan een aantal bijzondere eisen beantwoorden. Enerzijds moeten software- en hardwareoplossingen de toegang tot informatie gedurende lange tijd garanderen. Anderzijds staat ook menselijke input, in de vorm van archiefbeschrijvingen, werkprocessen en het gebruik van standaarden, er voor in dat informatie zo lang mogelijk beschikbaar en interpreteerbaar blijft voor een grote gebruikersgroep.

Digitale informatie is uiterst broos. Terwijl sommige gevaren ook analoge documenten bedreigen, zijn andere kenmerkend voor digitale data:

- Informatie in digitale vorm is een conceptueel object. Digitale multimedia kunnen bijvoorbeeld gemakkelijk gekopieerd en gewijzigd worden zonder een zichtbaar effect op de representeerbare inhoud. In tegenstelling tot analoge informatie is het daarom moeilijker de authenticiteit van digitale informatie te bewaren. Op korte termijn kunnen hardware, software

en menselijke fouten voor dataverlies zorgen. Vaak worden deze fouten onmiddellijk opgelost door specialistische correctiemethodes. In andere gevallen worden deze datacorrupties pas in een latere fase opgemerkt, op een moment waarop de data al schijnbaar correct verwerkt zijn zonder dat rekening is gehouden met de technische en visuele aspecten van de intellectuele inhoud.

- Door technologische wijzigingen kunnen dataformaten op termijn onbruikbaar worden. Ook de levensduur van opslagtechnieken is eindig. Om blijvende toegang tot de informatie te garanderen, zijn migratie- of emulatieplannen noodzakelijk. Technische metadata moeten bovendien voldoende informatie over de opgeslagen data aanreiken om tijdige ingrepen mogelijk te maken.

- Op lange termijn verandert het kennisdomein van gebruikersgroepen, verdwijnen dataspecialisten, wijzigen organisaties zich of krijgen ze een nieuwe opdracht. Het gevaar bestaat dat oudere opgeslagen data niet meer interpreteerbaar zijn voor een nieuwe gebruikersgeneratie. De opgeslagen data moeten daarom met voldoende contextuele informatie gedocumenteerd worden opdat deze nieuwe gebruikersgroepen de informatie kunnen blijven interpreteren.

Afhankelijk van de aard van de multimedia, worden de data in verschillende bestandsformaten opgeslagen, die aan de hand van voldoende technische metadata onderbouwd zijn en zo van veroudering gevrijwaard zijn. Vervolgens bepaalt het specifieke toepassingsgebied welke descriptieve metadata noodzakelijk zijn. Digitale beeldbestanden in een bibliotheek representeren bijvoorbeeld een gescand boek dat met bibliografische metadata beschreven moet worden. In een museum zullen beeldbestanden dan weer kunstwerken voorstellen, waarop andere descriptieve metadatavelden van toepassing zijn. Een videobestand in het bezit van een omroepzender bestaat mogelijk uit een aflevering van een televisie-uitzending, terwijl een videobestand dat deel uitmaakt van een installatie van een videokunstenaar weer op een totaal andere manier opgevat moet worden. In het eerste geval beschrijven de descriptieve metadata de serie en de aflevering waar de video betrekking op heeft, in het tweede geval worden de kunstenaar en de specificaties van de installatie beschreven.

Voor het beschrijven van multimedia zullen voor elke sector dus specifieke eisen gelden. Niettemin is er voor de consultatie van het archief nood aan een overkoepelend descriptief metadatamodel zodat zoekacties in de volledige dataset mogelijk zijn.

Dit rapport wil verder inzicht bieden in:

- de structurele eisen voor de beschrijving van digitale informatie,
- de gangbare bestandsformaten en compressietechnieken voor de opslag en uitwisseling van multimediale data,
- de beschikbare descriptieve metadatastandaarden die in elke sector gebruikt worden,
- de nodige metadata om de authenticiteit van de gedigitaliseerde informatie te bewaren,
- de noodzakelijke technische metadata om alle kenmerken van de individuele bestanden te beschrijven,
- bestaande praktijkvoorbeelden van gelijkaardige projecten waarin de langetermijnbewaring van multimedia centraal staat,
- het noodzakelijke datamodel voor de langetermijnbewaring van digitale informatie.

In het eerste deel van dit rapport wordt het OAIS-model (ISO-14721) toegelicht (§3). OAIS, voluit Open Archival Information System, is een conceptueel referentiemodel dat richtlijnen biedt bij de opzet van een digitaal archief voor langetermijnbewaring.

Vervolgens wordt een overzicht gegeven van gangbare compressieformaten (§4), metastandaarden (descriptief, technisch en administratief) en ontologieën uit de bibliotheek-, omroep-, culturele en erfgoedsector (§5) en containerformaten die de structurele samenhang van data en hun metadata waarborgen (§6).

Daarna worden twee praktijkvoorbeelden beschreven, namelijk de ontwikkeling van het e-Depot in de Koninklijke Bibliotheek van Nederland, en Multimatch, een project waarin de opzet van een Europese meertalige zoekmachine voor cultureel erfgoedonderzoek centraal staat (§7). De klemtoon in de twee projecten verschilt enigszins en dit verantwoordt ook de keuze voor deze praktijkvoorbeelden: in het eerste staat langetermijnbewaring van digitale objecten centraal terwijl het tweede zich voornamelijk concentreert op de ontsluiting en consultatie van het archief.

Ten slotte worden een aantal conclusies geformuleerd in verband met het gelaagd metadatamodel dat in het kader van het project BOM-Vlaanderen geconstrueerd zal worden.

3 Open Archival Information System (OAIS)

3.1 Geschiedenis

In 1982 werd het Consultative Committee for Space Data Systems (CCSDS) opgericht. Dit internationaal forum richtte zich tot ruimtevaartorganisaties die geïnteresseerd waren in de ontwikkeling van standaarden voor data-uitwisseling in de context van onderzoek. De studies wekten al gauw de aandacht van de International Organisation for Standardisation (ISO), die in 1990 een plan voorstelde om de voorschriften van CCSDS in een standaardisatieprogramma te formaliseren. Op vraag van het technisch comité van de ISO (ISO/TC20/SC13) begon het CCSDS aan een voorstudie over de langetermijnbewaring van digitale databestanden met betrekking tot ruimtevaartmissies. Dit internationale onderzoek resulteerde in 1995 in het conceptueel raamwerk OAIS (Open Archival Information System), dat als basis kon dienen voor verdere standaardisatieactiviteiten.[1]

Vanaf de start van het project bleek al dat de OAIS-studie niet enkel relevant was voor ruimtevaartorganisaties maar ook toepassingen kon vinden in andere uiteenlopende projecten. Zo toonden staatsinstellingen, bedrijven en universiteiten grote belangstelling voor de resultaten. Via workshops en debatten werd het model in een bredere context geplaatst en aangepast. Dat leidde in 1997 en 1999 tot nieuwe ontwerpen, die in 2000 door het ISO als een conceptstandaard aanvaard werden. Het OAIS-referentiemodel werd in 2002 goedgekeurd als de internationale ISO-standaard 14721.

Anno 2009 worden OAIS-concepten wereldwijd toegepast in digitale archieven.[2] De term 'OAIS-compliant' is een handelsmerk geworden voor vele commerciële archieven (zie bijvoorbeeld IBM's DIAS, OCLC's Digital Archive Service, ExLibris' DigiTool en het huidige project DPS in samenwerking met de

1 CCSDS (2002).

2 IBM (2008).

National Library of New Zealand).[3] Informatiearchitecten in de bibliotheek- en archiefwereld werken aan dataformaten zoals METS en MPEG-21/DIDL (cf. §6), die implementaties van OAIS-informatiepakketten zijn. Metadatastandaarden zoals PREMIS, mede ontwikkeld door bibliotheken, musea, staatsinstellingen en bedrijven, vullen OAIS aan op het gebied van preserveringsmetadata (cf. §5.3.1). Projecten zoals TRAC en DRAMBORA ontwikkelden audit- en certificatiestandaarden voor zogenaamde 'trusted digital repositories' in OAIS-stijl.[4]

In WP3 van het project BOM-Vlaanderen wordt eveneens geopteerd voor het OAIS-referentiemodel bij de opstelling van een gelaagd metadatamodel dat de preservering van multimediale data moet garanderen, hoewel er alternatieve referentiemodellen, waaronder CORDRA, DLF en IMS, bestaan.[5] De bewezen doeltreffendheid en bruikbaarheid binnen de genoemde internationale projecten met gelijkaardige doelstellingen als BOM (zie ook §7), het grote toepassingsbereik binnen archiefsystemen, de efficiëntie en helderheid van het OAIS-model en de focus ervan op langetermijnbewaring verantwoorden echter de keuze voor OAIS.[6]

De toelichting van het OAIS-model is grotendeels gebaseerd op het CCSDS-rapport van 2002.[7] Ook de heldere samenvatting van Lavoie geldt als bruikbare referentie.[8]

3 Respectievelijk IBM (2008); OCLC (2008); ExLibris (2008a, 2008b). Zie ook §7 voor een bondige toelichting bij deze projecten.

4 DRAMBORA (2008); OCLC (2007).

5 Respectievelijk CORDRA (2006); Dempsey & Lavoie (2005); IMS (2001-2008) en IMS (2003).

6 Ook J. Allinson (Allinson, 2006) licht heel helder de bruikbaarheid van OAIS voor preserveringsprojecten en digitale repositories toe. Ze refereert daarbij eveneens naar alternatieve referentiemodellen en suggereert dat '[e]valuating these would be a useful follow-on exercice' (p. 14). Ten slotte wijst ze ook op enkele 'Reference Models projects' die door het JISC (Joint Information Systems Committee) zijn ingericht en die interessant kunnen zijn voor 'the development of reference models for repositories' (p. 14). Andere referenties die de keuze voor OAIS ondersteunen, zijn: Brindley, 2000 (aanhaling van OAIS-implementatie in de British Library); Thibodeau, 2007 (verantwoording van het gebruik van OAIS in ERA, de Electronic Records Challenge, een initiatief van de National Archives and Records Administration, The National Archives, 2008), waarin de mogelijkheden voor langetermijnbewaring van digitale records onderzocht worden); Beagrie, 2004 (toelichting bij de implementatie van OAIS door JISC). Voor andere voorbeelden van projecten, zie §7.

7 CCSDS (2002).

8 Lavoie (2004).

3.2 Open Archival Information System[9]

In het OAIS-referentiemodel worden drie belangrijke concepten onderscheiden. De invulling van die concepten is onontbeerlijk voor een goede werking van het digitaal archief. Ten eerste is er de *designated community* of de doelgroep die van het archief gebruik zal maken. Het archief moet namelijk zo opgevat zijn dat die doelgroep, die niet noodzakelijk het grote publiek hoeft te zijn, zonder hulp en expertise van buitenaf het archief kan hanteren en de gegevens interpreteren. Daarom moeten in het eerste deel van het OAIS de doelstellingen van een open archiefsysteem nauwkeurig beschreven worden. Daarvoor wordt een gemeenschappelijke, abstracte terminologie gehanteerd, die gebruikt kan worden binnen elke archiefomgeving om daarin alle facetten van de gebruikte systemen, procedures, informatiedragers en uitwisselingspakketten te beschrijven (§3.2.1). In het tweede deel wordt het OAIS als een functioneel model benaderd, dat uiteenvalt in zes 'taken' of werkprocessen die nodig zijn voor de langetermijnbewaring van data (§3.2.2). Ten slotte wordt in het derde deel de kern van het OAIS-model betreden. Hier gaat het om een informatiemodel voor de beschrijving van de opgeslagen digitale data (§3.2.3).

3.2.1 Verantwoordelijke actoren binnen een OAIS-archief

Een mogelijke definitie van een OAIS-archief luidt: 'een samenwerking van mensen en machines die ervoor instaan informatie te archiveren en beschikbaar te stellen voor een doelpubliek: de *designated community*'. Deze definitie beklemtoont de twee belangrijkste functies van een archief:

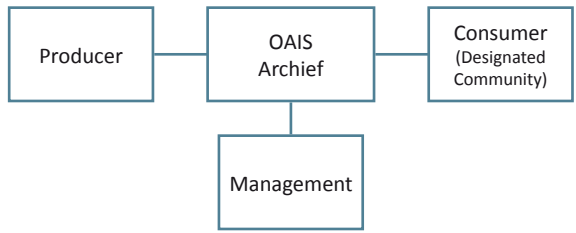
- informatie voor lange tijd archiveren,
- deze informatie beschikbaar houden voor een speciale doelgroep, de *designated community*.

De *designated community* staat centraal in het OAIS. Het is immers de doelgroep voor wie informatie gearchiveerd wordt. Een OAIS-archief moet ervoor zorgen dat deze gebruikersgroep de opgeslagen informatie op ieder moment kan raadplegen en begrijpen zonder een beroep te doen op externe expertise. Enkel door het bereik van de *designated community* te bepalen, kan een

9 In deze Nederlandstalige tekst zijn toch zoveel mogelijk Engelse termen met betrekking tot het OAIS-model behouden omdat een vertaling hiervan afbreuk zou doen aan een juist begrip van de betreffende sleutelconcepten.

OAIS garanderen dat de opgeslagen informatie op lange termijn beschikbaar en begrijpelijk blijft.

De *designated community* kan het grote publiek zijn, maar dat is zeker geen vereiste. Het begrijpelijk houden van opgeslagen informatie voor een groot publiek, zonder dat men de dataexperts hoeft te raadplegen, kan namelijk nodeloos een onoverkomelijk grote opdracht voor archieven vormen. Een voorbeeld is een OAIS-archief dat wetenschappelijke publicaties over een bepaald vakgebied bevat. De *designated community* bestaat dan mogelijk uit alle experts binnen deze discipline voor wie het archief als basis voor verder onderzoek dient. Alle tabellen, meetresultaten, enzovoort moeten interpreteerbaar blijven voor deze *designated community*, zonder dat die een beroep moet doen op *producers* (cf. infra). Daarnaast kan het OAIS-archief de informatie ook beschikbaar maken voor het grote publiek, zonder aan de eis van volledige begrijpelijkheid te voldoen.



Figuur 1: Actoren binnen het OAIS-archief[10]

Informatie wordt via een proces van *ingest* door een *producer* aan een OAIS-archief geleverd. De interactie tussen een *producer* en een OAIS-archief is vaak geformaliseerd in de vorm van een *submission agreement*, dat de specifieke details van aanlevering bevat: welke datatypes worden aanvaard, welke metadatavelden moeten voorzien zijn, welke protocols en logistieke ondersteuning zijn noodzakelijk, enzovoort.

De beleidslijnen voor een OAIS-archief worden bepaald door het *management*. Die bepaalt de aangewezen archiefstandaarden, de strategische planning, het bereik en is verantwoordelijk voor de beveiligde langetermijnbewaring van de aangeleverde informatie. Deze groep kan door middel van certificatie een betrouwbaar OAIS-archief uitbouwen.

10 Figuren 1 t.e.m. 4 zijn gebaseerd op figuren in CCSDS (2002).

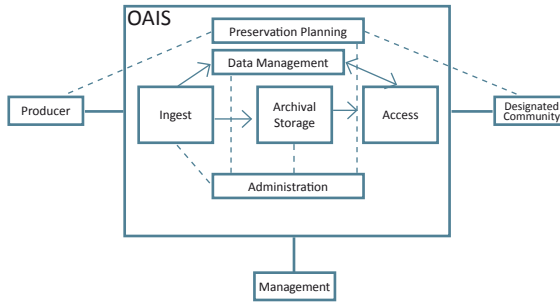
Een OAIS-archief stelt de opgeslagen informatie aan een groep *consumers* ter beschikking, waarvan de *designated community* een bijzonder type *consumer* is. Het archief moet de opgeslagen data zo archiveren dat de *designated community* onafhankelijk van de *producer* de informatie kan interpreteren. Andere mogelijke *consumers* zijn het grote publiek, interne werkprocessen of externe OAIS-archieven, eventueel in samenwerking met elkaar.

Samengevat is een archief in de eerste plaats voor de volgende taken verantwoordelijk, wil het aan de richtlijnen van het OAIS-referentiemodel voldoen:

- onderhandelen met de *producers* om alle nodige informatie te verkrijgen voor de archivering van hun data,
- voldoende rechten krijgen op de gearchiveerde informatie om de langetermijnbewaring te garanderen,
- het bereik van de *designated community* definiëren,
- garanderen dat de opgeslagen informatie begrijpelijk blijft voor de *designated community* zonder dat deze de hulp van de *producers* moet inroepen, gedocumenteerde procedures en richtlijnen volgen die garanderen dat de opgeslagen informatie gevrijwaard is van alle mogelijke risico's die beschikbaarheid of begrijpelijkheid van de data onmogelijk zouden maken, op elk moment toegang kunnen verlenen tot de authentieke kopieën van de gearchiveerde informatie in originele vorm of de weg naar de originele vorm aanwijzen, de gearchiveerde informatie beschikbaar stellen aan de *designated community*.

3.2.2 Het functioneel model van OAIS

Behalve een definitie van actoren en gebruikersgroepen vormt het OAIS ook een functioneel model voor de werking van een archief. Hierin zijn zes taken te onderscheiden. In §3.2.1 werd al het proces van *ingest* aangehaald, waarmee *producers* informatie aan een archief leveren. De *ingest*-module is de externe toegang tot het OAIS-archief, die zichtbaar is voor *producers*. Specifieke functies zorgen vervolgens voor de aanlevering van data, de bevestiging van ontvangst, de validatie van alle datacomponenten, de transformatie van de informatie in een vorm die geschikt is voor de opslag en het beheer binnen het systeem, de extractie of aanmaak van descriptieve metadata om de zoekinterfaces van het OAIS-archief te ondersteunen en het transport van de aangeleverde data naar de uiteindelijke archiefomgeving.



Figuur 2: Het functioneel model van OAIS

De *archival storage*, het hart van het digitaal archief, garandeert de langetermijnbewaring van de gedigitaliseerde informatie. Deze component zorgt ervoor dat de geleverde data in geschikte vorm op online-, nearline- of offlinesystemen opgeslagen worden. Individuele *bitstreams* worden zodanig opgeslagen dat alle bits voortdurend beschikbaar blijven zonder risico op bit-rot. De weergave van de bits in een leesbare presentatievorm moet ook worden gewaarborgd. Om aan beide eisen te voldoen, zullen regelmatig datamigraties naar nieuwe opslagmedia plaatsvinden en zal het archief *error checking-procedures* en *disaster recovery* moeten voorzien. Dataformaten zullen mogelijk naar nieuwe formaten geconverteerd worden als er technische wijzingen plaatsvinden die ervoor zorgen dat de *bitstreams* niet meer met standaardtools afgespeeld kunnen worden. Door software-emulatie kunnen *bitstreams* behouden blijven indien dataconversie onmogelijk blijkt wegens het risico op informatieverlies of wegens een gebrek aan voldoende *representation information* (cf. infra).

Een derde component, *data management*, verzorgt de catalogusomgeving waarin de gearchiveerde data geïdentificeerd en beschreven worden. De catalogus bevat naast de administratieve metadata, die de interne werking van het OAIS-archief ondersteunen, beschrijvingen van alle versies van opgeslagen bestanden en systemen, de geschiedenis van de datamigraties en formaatconversies. *Data management* verzorgt ook eventuele meldingen en aanvragen van OAIS-componenten.

Het echte onderhoud van de opgeslagen data en hun bescherming tegen veroudering gebeurt door de *preservation planning*. Ze controleert de staat van de opgeslagen informatie: is die nog leesbaar met behulp van de huidige technologieën en is ze nog begrijpelijk voor de *designated community*? De *preservation planning* schat ook de impact van veranderende technologieën

op de gearchiveerde informatie in en stelt een planning op om het OAIS-archief aan zijn verplichtingen tegenover de *designated community* te laten voldoen. Er worden strategieën ontworpen voor eventuele migratie, conversie of emulatie van de informatie en er wordt gezorgd voor de implementatie van deze strategieën in het OAIS-archief.

Access biedt vanzelfsprekend de toegang tot het OAIS-archief aan: in enge zin voor de *designated community*, in ruime zin voor de gebruikersgroep *consumers*. De *access*-component stuurt zoekvragen van de *consumer* door naar het *data management* en presenteert de daaruit resulterende metadata, eventueel naar een vereenvoudigde vorm geconverteerd. *Access* is ook verantwoordelijk voor de authenticatie en autorisatie van eindgebruikers en toegang tot het gearchiveerde materiaal. Gearchiveerde data uit de *archival storage* worden (eventueel na een interne conversie) in een presenteerbare vorm doorgegeven aan de eindgebruikers. De kerntaak van *access* bestaat er dus in om de gearchiveerde data toegankelijk te maken voor de *designated community*.

De laatste component, *administration*, staat in voor de dagelijkse werking van het OAIS-archief. *Administration* verzorgt niet alleen de contacten met de *producers* en *consumers* maar is ook verantwoordelijk voor de archief- en toegangssystemen. Ze verifieert de nodige functionaliteiten zoals *monitoring*, *system performance* en *updates* en staat in direct contact met de andere OAIS-componenten.

3.2.3 Het informatiemodel van OAIS

3.2.3.1 Inleiding

Om langetermijnbewaring van informatie mogelijk te maken, is een duidelijke definitie van informatie in het kader van OAIS noodzakelijk.

Personen of systemen hebben een *knowledge base* die hen toelaat om een set aangeleverde informatie te begrijpen. Zo kan iemand die het hiëroglyfenschrift en de oud-Egyptische taal kent, oud-Egyptische teksten lezen en interpreteren.

De definitie van *information* luidt: ‘elk type kennis dat in de vorm van data uitgewisseld kan worden’. Data vormen een representatie van de informatie. Op de Rosettasteen (zie afbeelding cover) wordt een oud-Egyptische tekst (informatie) in de vorm van hiëroglyfen (data) gerepresenteerd. De combinatie van de hiëroglyfen en de kennis van de oud-Egyptische taal en haar

schrift vormt betekenisvolle begrijpelijke informatie. Zonder kennis van deze taal of dit schrift kunnen de data niet geïnterpreteerd worden. De data moeten begeleid worden van een beschrijving van het Egyptische schrift en van de wijze waarop men de tekens moet omzetten naar een taal die wel tot de *knowledge base* behoort, bijvoorbeeld het oud-Grieks. Deze begeleidende data zijn de *representation information*. Iemand met een *knowledge base* die het oud-Grieks bevat, kan de Rosettasteen begrijpen. Dat was inderdaad het geval aan het begin van de 19de eeuw toen Jean-François Champollion met behulp van zijn kennis van het oud-Grieks het oud-Egyptische schrift op de steen kon ontcijferen.

De definities van *information*, *data* en *representation information* zijn ook toepasbaar op digitale informatie. Een voorbeeld: een file van 50MB (data) wordt aan een OAIS-archief geleverd. Deze *data* zijn een bitstroom die zonder *representation information* niet interpreteerbaar is. Begeleidende data moeten deze bitstroom beschrijven als een TIFF 6.0-bestand, dat een afbeelding in de Adobe RGB 1998 kleurenruimte bevat en dat een scan betreft van de afbeelding van de Rosettasteen met oud-Griekse en oud-Egyptische teksten. Om de bitstroom op lange termijn toegankelijk te houden, moet een OAIS-archief informatie opslaan over de TIFF 6.0-standaard, over de Adobe RGB 1998 kleurenruimte en in extremis ook over de taal waarin het object is opgesteld. Deze laatste eis kan van toepassing zijn in bepaalde vakgebieden waarin het gebruikte jargon aan verandering onderhevig is. In zulke gevallen is de archivering van een woordenlijst noodzakelijk om de opgeslagen informatie binnen een wetenschappelijk discours te bewaren.

De recursieve aard van dergelijke pakketten vormt een bijkomende moeilijkheid. *Representation information* zoals de TIFF 6.0-standaard, ZIP, Adobe RGB 1998 of het Grieks, is namelijk ook een vorm van *data* die gearchiveerd en beschreven moeten worden met eigen *representation information*. Dat leidt tot een netwerk van beschrijvingen. Een OAIS-archief moet de *knowledge base* van een *designated community* kennen om de archivering van *representation information* tot een minimum te beperken. Bovendien kan de *knowledge base* van de *designated community* evolueren zodat een aanpassing van de nodige *representation information* noodzakelijk is.

In de praktijk mag men er niet van uitgaan dat de bewaring van *representation information* niet nodig is omdat er altijd sofware beschikbaar zou zijn om de dataobjecten te renderen. Dat is namelijk een illusie. De archivering van een werkende IT-infrastructuur (software en hardware), bijvoorbeeld door

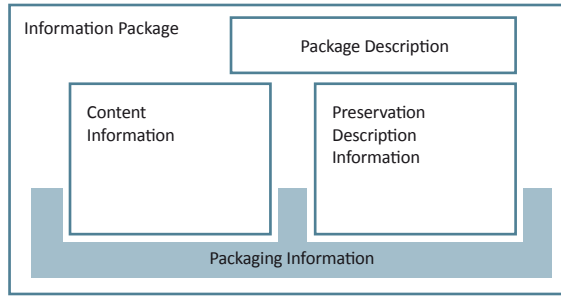
middel van emulatie, is complexer dan de opslag van *representation information* in digitale of papieren vorm.[11]

De geschetste methode van databeschrijving tot op bitniveau lijkt enigszins tegengesteld aan objectgeoriënteerde beschrijvingen waarbij dergelijke implementatiedetails verborgen worden. Het is echter een typisch vereiste bij de archivering van digitale data om minutieus *representation information* bij te houden. Digitale informatie is een conceptueel object dat altijd geïnterpreteerd moet worden binnen een specifieke IT-infrastructuur, die gevoelig is voor allerlei technologische veranderingen. Bovendien wordt deze informatie door mensen en organisaties geraadpleegd, van wie het kennisdomein en de organisatiestructuur zich eveneens op lange termijn zullen wijzigen. Dat zorgt ervoor dat digitale informatie geen statisch gegeven is maar voortdurend aan de gangbare praktijk getoetst moet worden.

3.2.3.2 Information Packages

Voor de uitwisseling van data tussen *producers*, het *archive* en de *consumers* voorziet OAIS *information packages*. Deze pakketten zijn containers die twee types informatie bevatten:

- *content information*, met de te archiveren informatie,
- *preservation description information*, met metadata voor de langetermijnarchivering van *content information*.



Figuur 3: Information Package

11 De auteurs van het rapport CCSDS (2002), p. 2-4 verklaren: 'it is harder to preserve working software than to preserve information in digital or hardcopy forms'.

Zoals aangegeven bestaat *content information* uit twee delen:

- *data*: de representatie van informatie die gearchiveerd moet worden (bijvoorbeeld een beeldbestand),
- *representation information*: de informatie die *data* naar interpreteerbare concepten omvormt (bijvoorbeeld de TIFF-standaard beschrijft hoe een reeks bytes naar een beeld omgevormd kan worden).

Representation information bevat twee types data:

- *structure information* beschrijft hoe bytes omgezet kunnen worden naar interpreteerbare concepten zoals letters, pixels, geluid, tabellen,...
- *semantic information* is een uitbreiding van *structure information*. Zelden volstaat informatie over een gebruikte standaard (bijvoorbeeld TIFF) om de data te interpreteren. *Semantic information* beschrijft de informatie zelf: bijvoorbeeld de taal van een document, de beschrijving van een beeld, de relaties tussen de dataobjecten,...

Voor de beschrijving van *content information* wordt een *preservation description information*–pakket (PDI) gevormd dat de volgende metadata bevat:

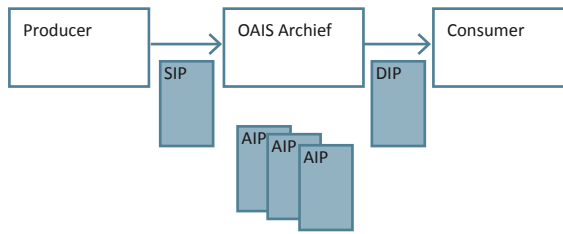
- *provenance*-metadata beschrijven de ontstaansgeschiedenis van de *content information*: de oorspronkelijke eigenaar(s) van de data, de processen die de huidige vorm van de data bepaald hebben en de beschikbare versies. Enkele voorbeelden van *provenance*-metadata zijn beschrijvingen van de gebruikte scanapparatuur, metadata over het scanproces, de gebruikte software en de processen, verwijzingen naar het analoge origineel, de versiegeschiedenis van alle files, *copyright statements* en de beschrijving van licentiehouders.
- *context*-metadata beschrijven de relaties van de *content information* tot informatie die niet in het *information package* zit. Voorbeelden van *context*-metadata zijn verwante datasets, verwijzingen naar documenten in de originele omgeving op het moment van publicatie, helpfiles en de taal.
- *reference*-metadata beschrijven de externe en interne *identifiers* waarmee *content information* op een unieke wijze geïdentificeerd kan worden.

Voorbeelden van *reference*-metadata zijn *object identifiers*, bibliografische beschrijvingen, ISBN's, ISSN's, DOI's, handlers, versienummers, namen en titels.

- *fixity*-metadata bevatten *checksums* en andere beveiligingen die testen of *content information* op een ongedocumenteerde wijze werd aangepast. Voorbeelden van *fixity*-metadata zijn *checksums, digital signatures*, certificaten, encryptie en CRC's.

Packaging information betreft de metadata die alle onderdelen van de *content information* en *preservation description information* op logische of fysieke wijze met elkaar verbinden. Wanneer de *content information* en *preservation description information* bijvoorbeeld in een ZIP-file aangeleverd worden, dan is de *packaging information* de *manifest file* met de namen van de files en hun beschrijvingen in het ZIP-pakket.

Package description information bevat de metadata waarmee bepaalde *content information packages* teruggevonden kunnen worden in grote collecties, bijvoorbeeld via een Dublin Core-record.



Figuur 4: Information packages

De bovenstaande *information packages* worden tussen de *producer*, het *archive* en de *consumer* uitgewisseld. Over het algemeen ontbreken er in elk pakket archiefmetadata, waardoor niet volledig aan de OAIS-standaard voldaan kan worden. Zo bevatten pakketten die door *producers* aangeleverd worden doorgaans onvoldoende *preservation description information* en de opbouw van de pakketten is mogelijk niet onmiddellijk geschikt voor opslag in het langetermijnarchief. *Information packages* die aan *consumers* geleverd worden, zullen dan weer een ander type data bevatten dan de data die het archief bevat.

OAIS onderscheidt daarom drie types *information packages*:

- het *submission information package* (SIP) is het *information package* dat door de *producer* aan het OAIS-archief geleverd wordt.
- één of meerdere SIP's vormen samen, na de nodige transformaties en metadataverrijking, een *archival information package* (AIP), dat in het archief opgeslagen wordt.
- Als antwoord op vragen van *consumers* zal het archief een AIP of delen ervan, al dan niet getransformeerd, als een *dissemination information package* (DIP) vrijgeven.

3.2.4 Overzicht

Uitgaand van het OAIS-referentiemodel zijn er verschillende metadatatypes nodig om digitale data volledig te beschrijven. In de volgende hoofdstukken worden de meest courante formaten, metadatastandaarden, conceptuele modellen, thesauri en containers die voor dit rapport relevant zijn, toegelicht.

4 Dataformaten

4.1 Inleiding

De opslag of verzending van multimediale data in hun ruwe vorm (beeld, geluid, video) is tegenwoordig vaak een onhaalbare en ondankbare opgave geworden. De noodzakelijke opslagruimte en bandbreedte zijn immers immens. Een uitkomst voor dit probleem wordt geboden door broncodering. Het doel van deze technologisch revolutionaire oplossing is de efficiënte representatie van informatie, zodat men optimaal gebruik kan maken van de schaarser wordende opslag- en transportcapaciteit. Bij broncodering probeert men maximaal gebruik te maken van eventuele redundantie in de data die men wil bewaren of transporteren. In sommige vormen van broncodering wordt bovendien a priori uitgegaan van tekortkomingen bij de ontvanger van de data. Zo zijn bepaalde subtiele schakeringen in de data niet waarneembaar voor het menselijk oog of oor. Die hoeven dan ook niet mee opgeslagen of verzonden te worden. De betreffende algoritmes die ontwikkeld zijn om de ruwe informatie te coderen/comprimeren en daarna te decoderen/decomprimeren, worden een codec genoemd.

In het volgende overzicht worden de gangbare compressiestandaarden (§4.2) en fysieke containerformaten (§4.3) voor de opslag van de verschillende multimediale datasoorten toegelicht. Een korte samenvatting dient hier als aanzet voor een beter begrip van de exhaustieve lijst.

Het meest gangbare compressieformaat voor videomateriaal is nog steeds MPEG-2 (§4.2.1). Door de opkomst van meer performante eindgebruikertoestellen voor het opnemen en afspelen van videomateriaal, het bestaan van breedbandverbindingen en de intrede van *high-definition television* en DVD (HD-DVD en Blu-Ray) zal H.264/MPEG-4 AVC (§4.2.1.2) binnenkort ongetwijfeld MPEG-2 als dominante standaard verdringen. Voor digitale cinema en professionele postproductie blijft men echter trouw aan het formaat Motion JPEG 2000 (§4.2.18). De *open source*-wereld heeft het dan weer meer begrepen op het patentvrije Theora-formaat (§4.2.6), dat het moet opnemen tegen H.264/MPEG-4 AVC en Motion JPEG 2000.

De populairste compressietechniek voor audiomateriaal is momenteel ongetwijfeld MP3 (§4.2.10). Toch zal die op korte termijn door AAC (§4.2.11) vervangen worden omdat dit formaat een betere kwaliteit bij eenzelfde *bitrate*

garandeert. Voor langdurige preservering van audiodata wordt er echter doorgaans gekozen voor verliesloze compressie (vb. FLAC, §4.2.11.2) of zelfs ongecomprimeerde opslag (d.m.v. zuivere PCM-*samples*). Dit laatste wordt nu soms al toegepast op de nieuwe Blu-Ray-schijfjes omdat de hoeveelheid informatie die opgeslagen moet worden voor audio enkele grootteordes lager ligt dan voor video. Terwijl de opgeslagen data bij video exponentieel stijgen naarmate grotere resoluties bewaard worden, is de hoeveelheid informatie bij audio enkel afhankelijk van het aantal kanalen (stereo, 5.1,…) dat men wil bewaren. Digitale cinema is op dat terrein dan ook dominant, aangezien de Dolby TrueHD-standaard tot veertien kanalen (13.1) aankan.[12] De huidige Blu-Ray en HD-DVD-standaarden ondersteunen slechts acht kanalen (7.1).

JPEG (§4.2.16) is veruit de meest courante compressiestandaard voor afbeeldingen. JPEG werd specifiek ontworpen voor de compressie van digitale beelden en geldt ondertussen als de standaard voor afbeeldingen op het internet en voor foto's in digitale camera's. Op het gebied van de digitale cinema is zijn superieure opvolger JPEG2000 (§4.2.18) de feitelijke standaard. Voor verliesloze compressie zijn PNG (§4.2.20) en TIFF (§4.2.21) de toonaangevende standaarden. Mogelijk zal PNG de rol van JPEG in de internetwereld meer en meer overnemen. TIFF garandeert dan weer de beste kwaliteit, bijvoorbeeld in archiveringstoepassingen, en bovendien wordt dit formaat door alle platformen het best ondersteund.

Het dominante containerformaat voor de uitwisseling van A/V-materiaal met bijhorende data en metadata in de wereld van digitale cinema en omroepen is ongetwijfeld MXF (§4.3.7). Het is ontworpen om data tijdens het productieproces probleemloos te kunnen uitwisselen. Vanwege de dominantie van Microsoft op het gebied van besturingssystemen, zullen AVI (§4.3.12) en WMA (§4.2.15) nog een poos als A/V-containerformaat in de internetwereld voorkomen. De superieure codec H.264, die in het internet en de mobiele wereld opgang maakt, heeft echter ook een eigen containerformaat MP4 (§4.3.8), waardoor zijn belang als A/V-container zeker zal toenemen. Voor de archivering van A/V-materiaal zullen vooral AVI en in de toekomst zeker ook MXF als containerformaten gebruikt worden, aangezien die ook ruwe data, dus zonder compressie, kunnen bevatten.

12 Voor digitale cinema zie o.m. DCI (2008).

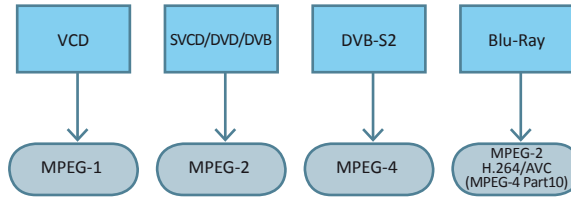
4.2 Compressieformaten

4.2.1 MPEG

4.2.1.1 Achtergrond en doelstelling

MPEG, voluit Motion Pictures Experts Group, is een werkgroep, bestaande uit academici en mensen uit de bedrijfswereld, die zich bezighoudt met de ontwikkeling van standaarden voor compressie van audio- en videosignalen.[13] De werkgroep is verantwoordelijk voor de MPEG-1, MPEG-2 en MPEG-4-standaarden en ook de MP3- en AAC-standaarden werden door hen vastgelegd. Behalve bij het ontstaan van audio- en videostandaarden was ze ook betrokken bij de ontwikkeling van standaarden voor metadata en *digital rights management.*

De codecs die door MPEG ontwikkeld zijn, zijn hoofdzakelijk *lossy* codecs, hoewel sommige standaarden ook *lossless* compressie ondersteunen. Bij de compressie van videobeelden en audiofragmenten door *lossy* codecs gaan bepaalde gegevens verloren. Hierdoor neemt de kwaliteit af maar is er een veel hogere compressie mogelijk. De MPEG-codecs hebben ondertussen een ruime ingang gevonden.



Figuur 5: Toepassingen MPEG

Een typische MPEG-videocodec is een zogenaamde hybride codec. De codec splitst het beeld op in macroblokken. Die macroblokken worden gecodeerd aan de hand van verschillende mogelijke methodes. Bij 'intra-codering' wordt gebruik gemaakt van de informatie in het beeld zelf. Aan de hand van de reeds gecodeerde macroblokken in het beeld wordt een schatting gemaakt. Dat is de zogenaamde 'spatiale' voorspelling, die spatiale redundantie uitbuit. Bij 'inter-codering' wordt in vorige beelden naar een onderdeel van het beeld

13 MPEG (2008).

gezocht dat het meest op het macroblok lijkt en er wordt een bewegingsvector opgeslagen die naar dit macroblok verwijst. Dat is de temporele voorspelling, die dus temporele redundantie uitbuit. Aangezien deze voorspellingen niet exact zijn, wordt het residu of het verschil tussen het originele beeld en het geschatte beeld, opgeslagen. Dit residu wordt achtereenvolgens getransformeerd naar het frequentiedomein en vervolgens gekwantificeerd. Hierna wordt de gecreëerde informatie met een entropiecodering gecomprimeerd.

MPEG standaardiseert echter enkel het bitstroomformaat en dus de decoder. Dat laat fabrikanten van encoders toe om aan de encoderzijde eigen optimalisaties door te voeren. Hierdoor kan de kwaliteit van een compressieformaat in de loop van de jaren verbeterd worden, terwijl alle bestanden door alle compatibele spelers afspeelbaar blijven.

4.2.1.2 MPEG-standaarden

MPEG-1[14]

MPEG-1 werd in 1991 door de Moving Pictures Experts Group ontwikkeld. Het gaat om de eerste standaard die door MPEG werd ontwikkeld en die later ook als standaard voor het video-cd formaat (VCD) gebruikt werd. De belangrijkste doelstelling van de codec bestond er in om VHS-kwaliteit op CD-formaat te leveren. Ook de populaire MP3-standaard behoort tot de MPEG-1-familie. De codec ondersteunt enkel progressieve beelden. Naast de video- en audiostandaarden beschrijft MPEG-1 ook de multiplexing en synchronisatie van de video- en audiostromen.

MPEG-2[15]

MPEG-2 werd ontwikkeld als de opvolger van MPEG-1. De standaard moest de tekortkomingen van MPEG-1 wegwerken. Zo werd een betere ondersteuning voor hogere resoluties toegevoegd. Verder werd de audiocompressie van MPEG-1 vervangen door de AAC-compressie. De standaard biedt hogere kwaliteit aan een lagere bitrate. Verder werden er ook nieuwe encryptie-systemen toegevoegd. MPEG-2 wordt onder andere gebruikt voor DVD's, digitale televisie-uitzendingen (DVB-T, DVB-C,...) maar ook Blu-Ray ondersteunt nog altijd het MPEG-2-formaat.

14 Chiariglione (1996).

15 Chiariglione (2000).

Op coderingsvlak werd het MPEG-2-formaat met zogenaamde B-beelden uitgebreid. Terwijl MPEG-1 enkel gebruik maakte van spatiale voorspelling (I-beelden) en voorspellingen aan de hand van het vorige beeld (P-beelden), laat MPEG-2 toe dat de temporele voorspelling op basis van het vorige én het volgende beeld wordt uitgevoerd (B-beelden).

MPEG-3

Oorspronkelijk werd MPEG-3 ontworpen voor *High-definition television* (HDTV) en dus voor hoge *bitrates*. Kleine wijzigingen aan MPEG-2 leken echter voldoende om hetzelfde resultaat te bekomen. De verdere ontwikkeling van MPEG-3 werd daarom beëindigd.

MPEG-4[16]

Net als de MPEG-1 en MPEG-2-standaarden bevat de MPEG-4-standaard beschrijvingen van compressietechnieken voor audio en video. De MPEG-4-standaard beschrijft echter ook enkele uitbreidingen, zoals de ondersteuning voor objecten, 3D-inhoud en interactiviteit. MPEG-4 was aanvankelijk vooral gericht op lage-bitrate-communicatie, maar die focus werd naderhand uitgebreid naar een compleet codeerformaat voor multimedia.

MPEG-4 biedt een hogere compressiegraad dan zijn voorgangers. Dankzij dit formaat passen speelfilms die DVD-kwaliteit benaderen, op één CD. Hierdoor werden implementaties van de standaard, zoals Divx en Xvid, heel populair in de piraterij. MPEG-4 wordt ook vaak gebruikt in omroepomgevingen, zoals bij Belgacom TV.

De MPEG-4-standaard is nog in volle ontwikkeling. Zo is de H.264/AVC-standaard recent afgewerkt. Die standaard vormt ondertussen Part 10 van de MPEG-4-standaard.

H.264/MPEG-4 AVC/MPEG-4 Part 10[17]

H.264/AVC, MPEG-4 Part 10 of MPEG-4 AVC (Advanced Video Coding), is de meest recente videocodec die door het MPEG-comité werd ontwikkeld in samenwerking met het ITU-T VCEG-comité. Deze standaard kent ook een uitbreiding voor schaalbaarheid (H.264/SVC), die in 2008 voltooid werd. De H.264/AVC codec streeft heel hoge compressie van videobeelden na. Tijdens

16 Koenen (2002).

17 ITU (2008); Wiegand, Sullivan, et al. (2003).

de ontwikkeling van de standaard werd als doel voorop gesteld de nodige *bitrate* tegenover de MPEG-2-standaard te halveren, wat vervolgens met succes gerealiseerd werd.

Door deze eisen is de H.264/AVC-standaard heel complex en rekenintensief. De mogelijkheden voor voorspellingen werden sterk uitgebreid. Zo kan er bij voorspelling bijvoorbeeld verwezen worden naar maximum zestien andere beelden. Die complexiteit en het rekenintensieve karakter typeerden echter ook de vorige MPEG-standaarden op het moment waarop ze geïntroduceerd werden. De H.264/AVC-standaard wordt momenteel al door YouTube en voor Blu-Ray gebruikt. Ook bij hardware voor beveiligingsdoeleinden, zoals bewakingscamera's, wint deze codec aan populariteit door de verminderde *bitrate*.

Profielen

De H.264/AVC moet eigenlijk als een toolkit beschouwd worden. De standaard stelt een grote hoeveelheid tools ter beschikking die elk afzonderlijk een invloed op de compressie en complexiteit uitoefenen. Het gebruik van die tools kan men vrij bepalen. Zo kan men er bijvoorbeeld voor opteren om voor mobiele toestellen complexe tools te vermijden. Om onderlinge compatibiliteit te voorkomen, werden voor verschillende toepassingen verschillende profielen opgesteld die vastleggen welke tools voor welke toepassingen geschikt zijn.

H.264/MPEG-4 AVC kent een zevental profielen:

- *Baseline Profile* (BP): voornamelijk bedoeld voor lagekosttoepassingen op toestellen met een beperking in beschikbare *resources*. Videoconferencing en mobiele toepassingen zijn hier typische voorbeelden van.

- *Main Profile* (MP): initieel bedoeld als belangrijkste gebruikersprofiel voor omroeptoepassingen en voor opslag. Dit profiel boet echter aan belang in sinds de uiteindelijke ontwikkeling van het *High Profile*, dat eenzelfde soort applicaties ondersteunt.

- *Extended Profile* (XP): bedoeld als *streaming* videoprofiel met de mogelijkheid tot een relatief hoge compressie, en dit bovenop een paar extra *robustness features* zodat omgegaan kan worden met dataverlies en met veranderingen van *streaming servers*.

- *High Profile (HiP)*: het belangrijkste profiel voor *broadcast*- en opslagapplicaties, vooral gebruikt voor Hoge-Definitie televisie (HD). Het profiel wordt al ondersteund voor Blu-ray, het nieuwe DVD-formaat.

- *High 10 Profile (Hi10P), High 4:2:2 Profile (Hi422P), High 4:4:4 Predictive Profile (Hi444PP):* profielen die gericht zijn op specifieke eisen van professionelen, zoals 10-bit- ondersteuning en *full-sampled* chrominatiekanalen.

Streaming

Het MPEG-formaat is uiterst geschikt voor *streaming media*. De Moving Picture Experts Group heeft daar bij de ontwikkeling van de formaten en bij de implementatie van codecs altijd rekening mee gehouden. De industrie volgde die redenering echter niet, waardoor MPEG op commercieel vlak nooit als *streaming*-formaat doorgebroken is. Dit had voornamelijk te maken met het feit dat concurrerende formaten zoals RealMedia en Windows Media een eigen *streamingserver* op de markt brachten terwijl MPEG dat niet deed. MPEG is immers een standaard die niet van een commercieel bedrijf uitgaat en dus afhankelijk is van implementaties en ontwikkelingen door derden.

Voor H.264/MPEG-4 AVC lijkt hier ondertussen verandering in te komen. Een belangrijk voorbeeld hiervan is de Darwin Open Source Streamingserver van Apple, die *streaming* van H.264/MPEG-4 AVC ondersteunt. Alles zal echter afhangen van de ondersteuning door de *players*. Inmiddels is de QuickTime-speler, dankzij het succes van iPod en iTunes, vrijwel even bekend als de Windows Media-speler. Vooral in het mobiele segment en de Set-top-box (iDTV & IPTV) zijn de kansen voor MPEG-*streaming* aanzienlijk.

4.2.2 VC-1[18]

VC-1 is een *lossy* videocodec, ontworpen door Microsoft. De videocodec werd na zijn ontwerp door SMPTE gestandaardiseerd. SMPTE is een organisatie van film- en video-experts, met leden in 85 landen. De standaarden die SMPTE invoert, worden wereldwijd overgenomen door professionals op het gebied van video, bewegend beeld en digitale cinema. De VC-1-standaard werd door Microsoft in Windows Media Video 9 geïmplementeerd en wordt onder andere in de Blu-Ray-standaard ondersteund. Net als bij de MPEG-standaarden wordt enkel de bitstreamsyntaxis vastgelegd. Er zijn ook twee

18 Cf. de documenten van SMPTE (2009); Goldman (z.j.).

documenten gepubliceerd die het transport en de gelijkvormigheid van VC-1 behandelen.[19]

De redenen waarom Microsoft voor de standaardisatie van VC-1 koos, waren de toegankelijkheid en de interoperabiliteit. Standaardisering stimuleert immers onafhankelijke ontwikkeling en garandeert dat verschillende implementaties interoperabel zijn. De standaardisatie bevordert zo de implementatie van de technologie binnen de sector.

De VC-1-codec is ontwikkeld als een tegenhanger van H.264/AVC, waarbij gestreefd werd naar minder complexiteit zonder veel aan kwaliteit in te boeten. Zo werd de ondersteuning voor meerdere referentiebeelden weggelaten en is de entropiecodering eenvoudiger. De VC-1-codec werd ontworpen om voor een groot bereik van *bitrates* compressie te bieden, van HD-resolutiebeelden bij 6-30 Mbits/s tot sub-CIF (< 352x288 pixels) bij 10 Kbits/s.

VC-1 betekent, net zoals de MPEG-standaarden, een evolutie van de conventionele DCT-gebaseerde videocodec. Bijgevolg is VC-1 heel gelijkaardig aan H.264. Zoals AVC bevat VC-1 een breed gamma aan geavanceerde coderingstools. Veel van die tools lijken op H.264/AVC. Sommige verschillen dan weer in kleine details, bijvoorbeeld in het gebruik van filters.

4.2.3 DivX/XviD[20]

DivX is een populaire implementatie van het MPEG-4 Visual format. De eerste versie van DivX bestond uit een gehackte versie van de Microsoft MPEG-4-codec. De DivX-codec werd heel populair in de videopiraterij. De mogelijkheid om DVD's aan een heel hoge kwaliteit op één CD op te slaan, maakte het delen van films immers heel eenvoudig. De implementatie van Microsoft week echter op enkele punten af van de officiële MPEG-4-standaard. Om die problemen op te lossen, werd het bedrijf DivX Networks opgericht, dat een eigen MPEG-4-codec ontwierp. Aanvankelijk ging het om een *open source* formaat (DivX 4) maar later werd de code gesloten (DivX 5). De codec werd gaandeweg geoptimaliseerd op het gebied van kwaliteit en snelheid, onder meer door de ondersteuning voor SSE-eenheden op *processors*. Verder ontwikkelde DivX Networks ook profielen en een certificering voor mediaspelers die het DivX-formaat ondersteunen.

19 Respectievelijk SMPTE (2005), RP227 en SMPTE (2006), RP228.

20 DivX (2009) en Xvid (2006).

Aangezien DivX gesloten is, werd het *open source*-project XviD opgericht. Het project baseerde zich op de code die DivX Networks onder de naam DivX4 had opengesteld. De XviD-programmeurs wisten de DivX4-codec te verbeteren en evenaarden ondertussen de beeldkwaliteit van DivX.

4.2.4 DIRAC[21]

Dirac is een *open source* videocodec die door de BBC in 2004 ontwikkeld werd. De codec is ontworpen voor een breed spectrum aan toepassingen, van webcontent in lage resolutie en HD broadcasting tot *near-lossless studio editing*. In tegenstelling tot de MPEG-codecs maakt Dirac gebruik van zogenaamde *wavelet*-codering.

De Dirac-codec werd om uiteenlopende redenen ontwikkeld. Volgens haar mandaat moet de BBC in internetdistributie voorzien en een eigen codec zou de licentiekosten hiervoor kunnen drukken. Verder wordt de BBC als het ware verplicht om open technologie te ondersteunen.

Dirac is opgebouwd uit twee verschillende codecs: Dirac en Dirac Pro. Dirac is gericht op lage *bitrate* distributie en ondersteunt dan ook lange GOP's (*groups of pictures*) en een zware aritmetische codering. Dirac Pro is gericht op de productie en postprocessing van media en voorziet daarom enkel in I-beelden (die aanpassingen tot op het beeld toelaten).

Dirac is veruit de meest flexibele codec in dit overzicht. Dirac ondersteunt beelden met resoluties van 176x144 tot 4096x3112 pixels, verschillende chromaformaten (4.4.4, 4.2.2, 4.2.0), 8 tot 16 bit bitdiepte, *interlacing* van metadata,...

De Dirac-codec is volledig vrij in gebruik. De BBC heeft de nodige patenten aangevraagd zodat de codec vrij kan blijven.

4.2.5 MJPEG/Motion JPEG/Motion JPEG2000[22]

MotionJPEG en MotionJPEG2000 zijn extensies van respectievelijk de JPEG-standaard en de JPEG2000-standaard. Bij die extensies wordt voor de encodering van video gebruik gemaakt van de beeldcompressietechnieken JPEG en JPEG2000, die beide ontwikkeld zijn door de Joint Photographic Experts Group (JPEG). De codering gebeurt enkel op *intraframe*-basis. MotionJPEG

21 DIRAC (2008). Cf. ook Onthriar, Loo, et al. (2006); Eeckhaut, Schrauwen, et al. (2005).

22 ISO/IEC (2002); ISO/IEC (2000); Marpe, George, et al. (2004).

gaat niet uit van temporele redundantie en is dus in dat opzicht vergelijkbaar met H.264/AVC-codering, waarbij ook enkel van I-beelden gebruik gemaakt wordt. Door het wegvallen van de temporele voorspelling is de compressie beduidend lager maar worden *editing*-toepassingen eenvoudiger omdat elk beeld een afzonderlijk geheel vormt. MJPEG werd, vóór de opkomst van MPEG-4, wegens zijn eenvoud vaak gebruikt in digitale camera's voor de encodering van filmpjes.

Bij een vergelijking van MotionJPEG2000 met de AVC-intra-codering blijkt dat voor lagere resoluties AVC hogere kwaliteit biedt bij gelijkaardige *bitrate*. Bij HD-resoluties is de situatie precies omgekeerd. MotionJPEG2000 werd dan ook als de standaard voor digitale cinema gekozen.

4.2.6 Theora[23]

Theora is een *open source* codec voor videobestanden, ontwikkeld door de Xiph.org Foundation. De codec bouwt voort op de VP3-codec van On2, waarvan de code in september 2001 werd vrijgegeven. De codec moet een *open source* alternatief vormen voor de MPEG-familie. Naast Theora wordt ook de audiocodec Ogg Vorbis ontwikkeld en een eigen containerformaat, allemaal *open source* en dus rechtenvrij.

Theora is, net zoals de meeste andere codecs die hier aan bod komen, een *lossy* codec. Hij maakt eveneens gebruik van gelijkaardige compressietechnieken zoals blokgebaseerde bewegingscompensatie, DCT's, intra-codering en temporele voorspelling. Hierbij worden echter geen B-beelden ondersteund.

Recent werd bekend dat de Theora-standaard ondersteund zal worden door de browser Mozilla Firefox, maar ook spelers en codecs zoals VLC, RealPlayer, Mplayer, Quicktime en FFmpeg ondersteunen de standaard.

4.2.7 DV[24]

Digitale Video (DV) is een digitaal videoformaat voor de opslag op tape. DV werd door Sony ontwikkeld en in 1995 gelanceerd. Sinds zijn ontstaan, is DV uitgegroeid tot de standaard voor videoproductie door amateurs of semi-professionelen. De DV-specificatie definieert zowel de codec als het tapeformaat. Er bestaat ook een verwant digitaal videoformaat, miniDV, voor de opslag op

23 Theora.org (2008); Xiph.org Foundation (2008); Giles (2004).

24 Digital Video (1994-2008).

kleinere tapes. DV levert een videokwaliteit die superieur is aan de analoge varianten zoals Video8, Hi8 en VHS-C.

DV gebruikt DCT *intraframe* compressie aan een vaste bitsnelheid van 25 Mbps, wat samen met de data voor het geluid, de foutendetectie en -correctie resulteert in een bitsnelheid van ongeveer 36 Mbps. Aan dezelfde bitsnelheid presteert DV iets beter dan de oudere MJPEG-codec en is het vergelijkbaar met *intraframe* MPEG-2. Bij deze laatste zijn enkel de I-frames intragecodeerd. DCT-compressie is *lossy*, waardoor soms artefacten rond kleine, complexe objecten zoals tekst optreden. De DCT-transformatie is speciaal aangepast voor de opslag op tape. Het beeld wordt verdeeld in macroblokken die bestaan uit vier 'luminantie-DCT-blokken' en één 'chrominantie-DCT-blok'. Zes macroblokken, geselecteerd op posities die voldoende ver uiteen liggen, worden gecodeerd in een vast aantal bits. Uiteindelijk wordt de informatie van ieder gecomprimeerd macroblok zoveel mogelijk in één 'sync-blok' op de tape opgeslagen. Hierdoor kan video op tape aan een hoge snelheid doorzocht worden en dit zowel in een voorwaartse als achterwaartse richting. Zo kunnen ook foute 'sync-blokken' opgespoord en gecorrigeerd worden.

Het DV-formaat gebruikt 'L-size-cassettes', terwijl de MiniDV zogenaamde 'S-size-cassettes' gebruikt. Beide cassettes bezitten een ingebed geheugen, van 4 Kbit voor MiniDV-cassettes tot 16Kbit. Dit geheugen kan gebruikt worden om uiteenlopende soorten data op te slaan, zoals een inhoudsopgave, de tijden en data van de opnames en van de camerasettings. Het is een EEPROM-geheugen dat gebruik maakt van het PC-protocol. Dit geheugen wordt echter zelden aangewend op gebruikersniveau. De meeste cassettes voor gebruikers bezitten zelfs geen chip, wat de prijs van een cassette aanmerkelijk verhoogt. De gebruikersapparatuur bevat meestal wel de nodige elektronica om gegevens van de chip in te lezen en weg te schrijven, hoewel hier nauwelijks gebruik van gemaakt wordt.

Er bestaan verschillende varianten van de DV-standaard. De meest bekende zijn Sony's DVCAM en Panasonic's DVCPRO voor professioneel gebruik. Sony's gebruikersformaat Digital8 is een andere variant die lijkt op het DV-formaat maar die aangepast is voor de opslag op Hi8-tape.

DVCAM van Sony is een professionele variant van de DV-standaard en gebruikt dezelfde codec en cassettes als DV en MiniDV, maar transporteert de tape 33% sneller. Hierdoor is DVCAM veel nauwkeuriger aan te passen. Een andere eigenschap van DVCAM wordt *locked audio* genoemd. Als DV gekopieerd wordt, valt de audio na een aantal kopieën niet meer met het beeld te synchroniseren. DVCAM kent dit probleem niet.

Panasonic ontwikkelde de DVCPRO-codec om een betere lineaire *editing* van het DV-formaat te verkrijgen. De tape is naast een controlespoor voor een beter *editing*, ook voorzien van een longitudinaal analoog audiospoor. De audio is beschikbaar in de 16 bit/48 kHz variant. DVCPRO gebruikt ook steeds 4:1:1-kleursubsampling (zelfs bij PAL). Op het niveau van de bitstroom is de standaard DVCPRO (DVCPRO25) identiek aan de standaard DV. DVCPRO werd door Panasonic aangeprezen als de uitgelezen DV-variant voor professionele *high-end*-applicaties. DVCPRO50 is in feite de combinatie van twee DVCPRO-codecs in parallel. De bitsnelheid wordt dus verdubbeld tot 50Mbps en gebruikt 4:2:2 *chroma subsampling* in plaats van 4:1:1. De verkregen videokwaliteit is vergelijkbaar met die van zijn tegenhanger, Digital Betacam. DVCPRO HD, ook wel gekend als DVCPRO100, gebruikt vier parallelle DVCPRO-codecs, resulterend in een bitsnelheid van 100Mbps. DVCPRO HD maakt gebruik van 4:2:2 kleurensampling. Deze codec is dus geschikt voor de opslag van HD-materiaal op tape.

4.2.8 Betacam[25]

Betacam is een familie van professionele videotapeproducten en werd door Sony ontwikkeld. Betacam heeft betrekking op de camcorder, de tape, de videorecorder of het formaat. Het originele Betacam-formaat, een analoog videoformaat, werd in augustus 1982 gelanceerd. De luminantie, Y, werd opgeslagen op één spoor en de chrominantie op een ander spoor die dienst deden als afwisselende segmenten van de R-Y en B-Y-componenten. Dit leverde een kwaliteit van 300 lijnen horizontale lumaresolutie en 120 lijnen chromaresolutie op. Een nadeel van dit formaat was de beperkte opnametijd. De cassettes konden maar een half uur video opslaan. In 1986 werd vervolgens het Betacam SP-formaat ontwikkeld, waarvan de horizontale resolutie tot 340 lijnen verhoogd werd. De kwaliteitsverbetering hiervan was klein, maar in combinatie met een nieuwe cassette die 90 minuten kon opnemen, werd Betacam SP de industriestandaard voor de meeste TV-stations en *high-end* productiehuizen eind jaren negentig.

Het digitale formaat, Digital Betacam (digibeta of d-beta), werd in 1993 gelanceerd. Het verving het Betacam en Betacam SP-formaat en de cassettes hadden een opnametijd van 40 minuten of 124 minuten. Het digitale Betacam-formaat neemt een DCT-gecomprimeerd signaal op met 10 bit YUV 4:2:2 *sampling* in NTSC- (720x486) of PAL- (720x576) resolutie. Daarnaast worden vier kanalen opgenomen met 48 kHz 16 bit PCM audio. Een vijfde analoog audiospoor is ook beschikbaar voor *cueing*. Digitale Betacam gebruikt temporele

25 Betacam PALsite (2000).

compressie, waarbij een sequentie van I- en B-beelden wordt opgenomen. Het is een populair digitaal videoformaat in de omroepwereld.

Betacam SX is een digitale versie van Betacam SP en werd in 1996 geïntroduceerd als een goedkoper alternatief voor de digitale Betacam. Het slaat de video op door gebruik te maken van MPEG-2 4:2:2 compressie, samen met vier kanalen voor de audio. Dit levert een betere chromaresolutie op en laat bepaalde postproductieprocessen toe. De cassettes hebben een opnametijd van 62 of 194 minuten.

MPEG IMX is een ontwikkeling van het digitale Betacam-formaat van 2001. Het gebruikt de MPEG-compressie zoals Betacam SX, maar aan een hogere bitsnelheid. De toegepaste compressie heeft drie formaten: 30 (6:1 compressie), 40 (4:1 compressie) of 50Mbit/s (3.3:1 compressie). Dit laat toe om de video op te slaan met verschillende ratio's aan kwaliteit en opslagefficiëntie. De video is opgenomen met het MPEG-2 4:2:2-profiel.

HDCAM, dat werd geïntroduceerd in 1997, is een *High Definition*-versie van het digitale Betacam-formaat. Het gebruikt een 8-bit DCT-compressie en het 3:1:1-profiel. De resolutie is 1440x1080 pixels, waardoor het compatibel is met 1080i. De bitsnelheid is 144Mbit/s. Voor de audio worden vier kanalen gebruikt van 20 bit/48 kHz digitale audio. HDCAM SR, de opvolger in 2003 van HDCAM, kan opnemen in 10 bits 4:2:2 of 4:4:4 RGB aan een bitsnelheid van 440 Mbit/s. Voor de compressie gebruikt HDCAM SR het nieuwe MPEG-4 Part 2 Studio profiel en het aantal audiokanalen wordt verhoogd tot 12 24 bit/48kHz kanalen.

4.2.9 MP2[26]

MPEG-1 Audio Layer II (MP2, ook wel Musicam genoemd) is een audiocodec, die door de ISO/IEC 11172-3 standaard gedefinieerd is. Terwijl MP3 veel populairder is voor PC en internetapplicaties, blijft MP2 de dominante standaard voor zenders.

De ontwikkeling van de MP2-standaard werd gestart aan het eind van de jaren tachtig door ISO's Moving Picture Expert Group (MPEG). MP2 is een psycho-akoestisch compressiealgoritme. Dat wil zeggen dat het informatie verwijdert die voor het menselijk gehoor quasi niet waarneembaar is. Om te bepalen welke signalen niet waarneembaar zijn, wordt het audiosignaal

26 ISO/IEC (1993).

geanalyseerd volgens een psycho-akoestisch model, dat de karakteristieken van het menselijk gehoor aanneemt.

Uit studies is gebleken dat bij een sterk signaal op een bepaalde frequentie de zwakkere signalen op naburige frequenties niet meer waarneembaar zijn. Hiervan maakt MP2 gebruik. MP2 deelt het audiosignaal in 32 frequentiesubbanden onder. Wanneer de audio van een subband niet waarneembaar is, wordt deze subband weggelaten. MP3 bijvoorbeeld verdeelt het audiosignaal over 576 frequentiesubbanden, waardoor MP3 een hogere frequentieresolutie heeft. MP3 gebruikt daarenboven ook nog entropiecodering, wat verklaart waarom MP3 lagere bitsnelheden dan MP2 nodig heeft om een vergelijkbare audiokwaliteit te bekomen.

MP2 is minder rekenintensief dan MP3, wat de codec efficiënter maakt voor hoog kwalitatieve percussieve geluiden (impulsen) en dit dankzij de goede eigenschappen van de filterbank op het gebied van timing. Een bijkomend voordeel hiervan is dat MP2 beter bestand is tegen digitale transmissiefouten. Ook daarom wordt MP2 nog steeds gebruikt voor applicaties in *broadcast*-omgevingen.

MP2 maakt deel uit van de DAB digitale radio en DVB digitale televisiestandaarden. Ook de meeste DVD-spelers bezitten een MP2-decoder, waardoor MP2 in die markt een concurrent is van Dolby Digital.

4.2.10 MP3[27]

MP3 is een afkorting van MPEG-1 Layer III en behoort tot de MPEG-1-standaard. Het is het derde onderdeel of de derde laag binnen de MPEG-1-audiocodering, na Layer I en Layer II. Iedere laag voegde nieuwe compressiemogelijkheden, nieuw ondersteunde *bitrates* en/of audiofrequenties toe.

MP3 is een *lossy* codec die gebruikt wordt om geluid te comprimeren. Bij MP3 worden, net zoals bij andere audiocompressietechnieken, onhoorbare geluidselementen uit het geluid verwijderd. Verder wordt een kwantisatiestap uitgevoerd die de compressie nog verder verhoogt. MP3 ondersteunt ook *joint stereo*. Bij deze modus wordt informatie die in beide geluidskanalen voorkomt slechts eenmalig opgeslagen. Die techniek is echter enkel efficiënt bij lage *bitrates*.

27 Hacker, Hayes (2000); Fraunhofer IIS (2008a).

MP3 was de standaard die leidde tot de doorbraak van digitale muziek in bestandsvorm. MP3 maakte muziek klein genoeg om van het internet te downloaden en ze op flashgeheugen te bewaren. Uiteindelijk deed ze een nieuw product ontstaan: de MP3-speler. MP3 wordt echter ook vaak in verband gebracht met piraterij, aangezien het mogelijk werd om audiofragmenten zodanig klein te maken dat ze eenvoudig via het internet gedeeld konden worden.

Uit subjectieve luistertesten blijkt dat gebruikers MP3's aan 192 kbits/s al een aanvaardbare kwaliteit toeschrijven. Dat komt neer op een compressie met factor zeven. Bij de maximale *bitrate* van 320 kbits/s is het kwaliteitsverschil niet meer hoorbaar, tenzij voor een getraind oor of met *high-end* geluidsapparatuur.

Mettertijd werden enkele extensies voor MP3 geïntroduceerd. Een voorbeeld hiervan is mp3PRO. Hierbij bestaat een MP3-bestand van een 44 kHz audiobestand aan 128 kbits/s uit een 22kHz 64 kbits/s MP3-stroom die door elke speler die de MP3-standaard ondersteunt, afgespeeld kan worden en het MP3-bestand is aangevuld met een zogenaamde SBR-extensie (Spectral Band Replication) die de hogere frequenties efficiënter codeert.

De compressie- en decompressiealgoritmes van MP3 zijn gepatenteerd door de eigenaar, het Fraunhofer instituut, en dus niet vrij beschikbaar voor commerciële producten of voor commercieel gebruik van de technologie. Persoonlijk gebruik van de MP3-software is toegestaan. *Open source* encoders en decoders, zoals Lame, worden toegelaten.

4.2.11 AAC/MPEG-2 Part 7/MPEG-4 Part 3[28]

AAC, voluit Advanced Audio Coding, is de gedoodverfde opvolger van het MP3-formaat. Het behoort tot de MPEG-2 en MPEG-4-standaarden en ondersteunt, in tegenstelling tot het basisformaat van MP3, *multichannel* audio. Uit geluidstesten blijkt dat AAC een gelijkaardige kwaliteit biedt bij 128 kbits/s van een 192 kbits/s MP3-compressie. Het formaat is echter complexer en minder populair dan MP3, wat zijn doorbraak zeker niet ten goede kwam. Momenteel wordt de standaard, ondermeer dankzij Apples' iTunes Store, meer en meer ondersteund. Zo ondersteunen onder andere Apple's iPods, Creative's Zens en Sony's Walkmans het AAC-formaat.

28 Fraunhofer IIS (2008b); DOLBY (2008); EBU (2003).

4.2.11.1 AAC+ of HE-AAC

Net zoals MP3 MP3pro-extensies kent, die hogere compressie garanderen, kent AAC HE-AAC. HE-AAC heeft twee versies, v1 en v2. Beide extensies gebruiken, net zoals MP3Pro, SBR (Spectral Band Replication). V2 voegt hier ook Parametric Stereo (PS) aan toe. Met die techniek wordt een monokanaal gecodeerd en wordt informatie aan de stroom toegevoegd, die toelaat het originele stereosignaal te berekenen. Die informatie neemt typisch 2 à 3 kbits/s in beslag. Deze techniek is dan ook vooral geschikt voor lage *bitrates*. Luistertesten wijzen uit dat HE-AAC v2 bij *bitrates* tot 32 kbits/s een beduidend hogere kwaliteit bereiken, een voorsprong die bij 48 kbits/s al helemaal weg is of zelfs omgezet is in een achterstand door de imperfecte reconstructie van de stereokarakteristieken.

Op *bitrates* van 32-64 kbps heeft HE-AAC doorgaans de beste geluidskwaliteit, in vergelijking met andere codecs. Bij 48 kbits/s wordt HE-AAC bijvoorbeeld bijna dubbel zo hoge kwaliteitswaarden toegeschreven in vergelijking met WMA en MP3. Bij hogere *bitrates* is HE-AAC nog altijd beter dan MP3. Een HE-AAC van 80/96 kbps klinkt ongeveer zoals een 128/140 kbps MP3 maar is even groot als de 80/96 kbps MP3. Vanaf 128 kbps en meer zal HE-AAC hetzelfde als MP3 klinken. HE-AAC wordt vooral gebruikt voor internetradio en mediaspelers, waarvan de opslagcapaciteit niet al te hoog is.

HE-AAC wordt echter nog maar weinig ondersteund. De *open source* encoder FAAC ondersteunt het formaat. Voorts kunnen ook Winamp, Foobar2000 en iTunes het formaat weergeven.

4.2.11.2 FLAC[29]

FLAC (Free Lossless Audio Codec) behoort tot een andere categorie codecs. FLAC is immers een verliesloze codec. FLAC is snel en *open source* en kent relatief hoge compressieratio's. Het is één van de best ondersteunde *lossless*-formaten. Bijna alle besturingssystemen (Mac OS X, Windows, Linux) ondersteunen FLAC en ook veel mediaspelers zoals Winamp, Media Player Classic en VLC.

Aangezien het om een *lossless* formaat gaat, hangt de behaalde compressie van heel wat factoren af. Sommige muziekgenres zijn eenvoudiger te comprimeren dan andere. Algemeen is er sprake van een compressie van 30-50%, tegenover een typische compressie van 80% bij MP3.

29 Coalson (2008); Malvar (2007).

4.2.12 Ogg Vorbis[30]

Ogg Vorbis is een *open source* compressiemethode voor geluidsbestanden. Het formaat werd ontwikkeld door Xiph.org als patentvrij alternatief voor formaten zoals MP3, AAC en WMA. 'Vorbis' heeft betrekking op het gebruikte compressiealgoritme en 'Ogg' is het containerformaat waarin de data bewaard worden. Net zoals bij MP3 wordt het geluid gecomprimeerd door irrelevante geluidsinformatie, die nauwelijks hoorbaar is, weg te filteren.

In subjectieve audiotests scoort Vorbis voor 128 kbits/s ongeveer tussen MP3, WMA en HE-AAC, MP3Pro in. Bij lagere *bitrates* zoals 32 kbits/s komt Vorbis na WMA9, HE-AAC en MP3Pro.

Het formaat wordt onder meer gebruikt door Epic Games (van de makers van Unreal Tournament) en EA Games. Verder zijn er meer en meer MP3-spelers die Vorbis ondersteunen, zoals de Bang & Olufsen BeoSound 6, Cowon-spelers, iRiver-spelers en de Sandisk Sansa. Voorts zijn er spelers voor zowel Mac OS X, Linux, Windows als voor BeOS beschikbaar.

4.2.13 AC-3/Dolby Digital[31]

AC-3, of Dolby Digital, is een *multichannel* audiocodec die door Dolby Laboratories ontwikkeld werd. Het formaat wordt wijdverbreid gebruikt voor DVD's en in digitale cinema. Het geluid bestaat doorgaans uit zes audiokanalen (5.1 genaamd), twee kanalen achteraan (links en rechts), drie kanalen vooraan (links, midden, rechts) en één kanaal voor de lage frequenties. De opvolger van DVD, Blu-Ray, maakt gebruik van een opvolger van Dolby Digital, namelijk Dolby TrueHD. Deze standaard ondersteunt tot veertien kanalen van elke 24bits/96khz audio.

4.2.14 TTA[32]

TTA of True Audio is een *open source* audio encoder. Net zoals FLAC is TTA *lossless*, wat betekent dat er geen geluidsinformatie weggelaten wordt. Uit testen blijkt dat TTA een iets hogere compressie dan FLAC kent en dit bovendien aan een hogere snelheid verkrijgt. Er zijn *plug-ins* beschikbaar voor Winamp, dBpowerAMP en Foobar maar er bestaan ook directShow-filters. Op het gebied van hardware is de ondersteuning voor het formaat echter

30 Xiph.org (2008); Svítek (2006).

31 ATSC (2005); DOLBY (2009a); DOLBY (2009b).

32 Djuric en Oler (ed.) (2006); Heijden (2005).

minimaal. Momenteel wordt het formaat enkel door de Neuston Maestro DVD-speler ondersteund.

4.2.15 Windows Media Audio[33]

WMA, voluit Windows Media Audio, is een *lossy* audiocodec, ontwikkeld door Microsoft. Het gaat om een gesloten formaat en wordt standaard gebruikt in het Windows besturingssysteem. De werking van het formaat is vergelijkbaar met MP3 maar levert in luistertests een iets betere score op. Bovendien ondersteunt WMA ook DRM, *digital rights management*, dat muziek beschermt tegen kopiëren of onrechtmatig afspelen. WMA wordt door veel MP3-spelers ondersteund. Naast MP3 is het de meest populaire standaard bij deze spelers. De iPod is één van de weinige MP3-spelers die de standaard niet ondersteunt. Op het gebied van software wordt WMA zoals gezegd ondersteund door de Windows besturingssystemen. Voor Mac OS X bestaat het Flip4Mac-programma, dat niet door Microsoft ontwikkeld werd maar dat wel WMA ondersteunt. Voor Linux is ook ondersteuning mogelijk, maar ook hiervoor is Microsoft zelf niet verantwoordelijk.

WMA kent naast de gewone compressie ook enkele specifieke uitbreidingen. De Professional versie gebruikt een geüpdatet algoritme en ondersteunt geluid tot 96 kHz, 24 bits en met 8 discrete kanalen. Voorts is er ook *lossless* compressie. Uit testen blijkt dat dit formaat beter comprimeert dan TTA maar significant trager is, hoewel het nog altijd relatief snel is. Ten slotte bestaat de Voice-tak, die specifiek ontwikkeld werd voor de compressie van spraak en die dan ook slechts geluid tot 22 kHz en mono ondersteunt. BBC World Service gebruikt deze standaard voor internet *radio streaming*.

4.2.16 JPEG[34]

JPEG is een beeldcompressieformaat dat ontwikkeld is door de Joint Photographic Experts Group (JPEG). Het formaat is de meest courante standaard voor afbeeldingen op het internet en in digitale camera's. Het bestandsformaat is *lossy*. Kenmerkend voor een te grote compressie zijn, net zoals bij de MPEG-codecs, blokartefacten en wazige contouren. Een typische compressie gebeurt met factor tien en dit zonder een opvallend kwaliteitsverlies. De compressie hangt echter hoofdzakelijk af van de afbeeldingen zelf. Beelden die zeer gedetailleerd zijn, zijn moeilijker te comprimeren dan beelden met weinig details en contrast. Het formaat is dan ook beter geschikt voor natuur-

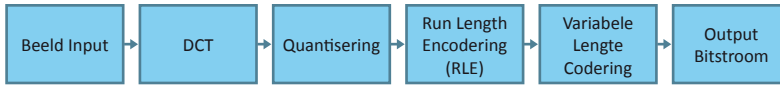
33 Heijden (2005); Windows Media (2009).

34 ITU (2004); JPEG en JBIG (2007); Wallace (1991); Wolfgang (z.j.).

lijke beelden dan voor de compressie van bijvoorbeeld grafieken en teksten. Hiervoor zijn bestandsformaten zoals GIF en PNG beter geschikt.

4.2.16.1 Werking

De JPEG-indeling is complex. In tegenstelling tot indelingen zoals van PNG of GIF wordt meer dan één mechanisme gebruikt. Er worden namelijk een aantal stappen na elkaar toegepast om tot het uiteindelijke JPEG-bestand te komen.

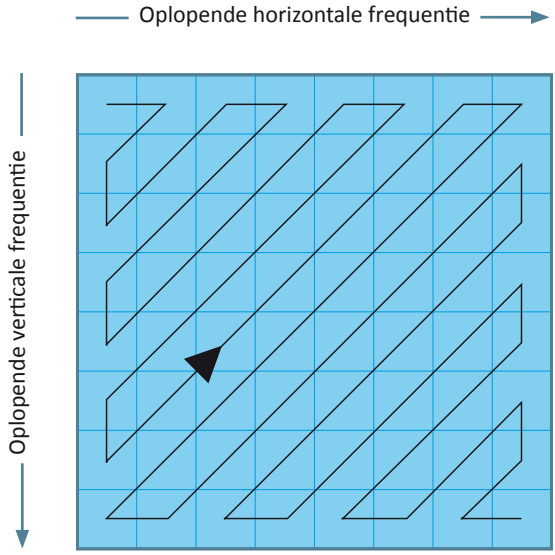


Figuur 6: JPEG - stadia van de compressie

Allereerst wordt het beeld opgesplitst in blokken van 8x8 pixels. Ieder blok wordt vervolgens verwerkt. De opdeling in en aparte verwerking van de blokken resulteren in de blokartefacten bij hoge compressie.

Na de verdeling in blokken wordt gebruik gemaakt van *intra-frame* codering. Hierbij worden enkel de gegevens binnen een macroblok gebruikt. Eerst wordt een DCT-transformatie op ieder blok toegepast, waarbij de spatiale pixelwaarden naar het frequentiedomein vertaald worden. Zo ontstaat een nieuw blok met frequentiecomponenten die gesorteerd zijn volgens stijgende detailinformatie. De component linksboven beschrijft de gemiddelde waarde van het volledige blok, de componenten rechts of onder dit blok zullen de details beschrijven van steeds kleinere gebieden. Op de frequentiecomponenten kan een kwantisatie worden toegepast. Hierbij wordt bepaald hoeveel informatie behouden zal worden en wat de grootte van de compressie zal zijn. Voor de kwantisatie wordt een matrix gebruikt met een waarde voor elke frequentiecomponent. Alle frequentiecomponenten zullen door hun overeenkomstig getal uit de kwantisatiematrix gedeeld worden en tot een geheel getal afgerond worden. Hierbij verdwijnt dus informatie. Het verhogen van de waarden van de kwantisatiematrix zal er voor zorgen dat meer en meer waarden nul worden. Binnen één kwantisatiematrix zullen de waarden voor de details (de frequentiecomponenten rechtsonder) doorgaans groter zijn om zo in verhouding meer informatie van de details te kunnen verwijderen en dus een hogere compressie te bereiken. Vervolgens wordt een *run-length* codering uitgevoerd. Voor die codering worden de frequentiecomponenten ingelezen in een zig-zag-patroon (zie figuur 7). Hierdoor worden eerst de niet-nulcoëfficiënten ingelezen, gevolgd door de nulcoëfficiënten. Hoe meer nulcoëfficiënten, hoe minder plaats het blok

zal innemen. Na de *run-length* codering wordt een variabele lengte met de Huffman-codering uitgevoerd. De verkregen waarden vormen ten slotte het JPEG-bestand.



Figuur 7: JPEG – run-length codering

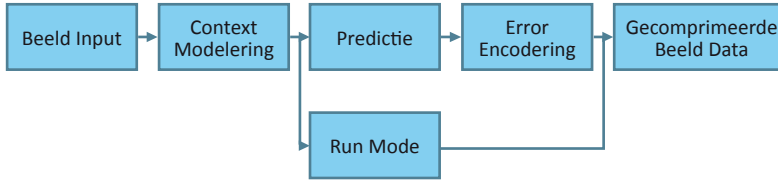
Behalve beeldcompressie ondersteunt een JPEG-bestand ook de toevoeging van extra informatie. Zo kan een JPEG-bestand vrije commentaar bevatten, maar ook EXIF-data (een standaard voor de beschrijving van informatie die in digitale camera's of scanners kan worden toegevoegd) of IPTC-data (de gestandaardiseerde indeling voor gegevens van de afbeelding).

Tot slot laat JPEG ook bepaalde verliesloze bewerkingen toe. Dit is mogelijk doordat de compressie in een horizontale of verticale richting werkt. Hierdoor kan een beeld verliesloos gedraaid worden over hoeken die een veelvoud van 90º bedragen. Ook horizontaal en verticaal spiegelen is mogelijk zonder informatieverlies.

4.2.17 JPEG-LS[35]

JPEG-LS of JPEG Lossless is een verliesloze uitbreiding van de JPEG-standaard, eveneens door de Joint Photographic Experts Group (JPEG) ontwikkeld. Behalve verliesloos kan JPEG-LS ook *near-lossless* comprimeren, waarbij een maximaal aantal afwijkingen toegelaten is. De standaard is ontwikkeld met het oog op de compressie van bijvoorbeeld medische beelden, waarbij fouten in het beeld niet wenselijk zijn. De latere JPEG-2000-standaard ondersteunt ook *lossless* codering maar is veel complexer dan het JPEG-LS algoritme.

Bij de ontwikkeling van de standaard werden verschillende mogelijke algoritmes onderzocht. Uiteindelijk werd geopteerd voor het LOCO-I algoritme (LOw COmplexity LOssless COmpression for Images) van HP Labs. Het algoritme is ontstaan als een zogenaamde 'lage-complexiteit projectie'. Het algoritme combineert eenvoud met het compressiepotentieel van contextuele modellen. Een toenemende complexiteit van een algoritme leidt immers in verhouding vaak tot een kleine compressietoename.



Figuur 8: JPEG-LS

4.2.17.1 Context modeling

LOCO-I is een zogenaamd *context modeling*-algoritme. Hierbij wordt tijdens het comprimeren een context aangemaakt die bijhoudt wat de kans is dat een bepaalde pixel door een andere pixel gevolgd wordt. Die context wordt gebruikt als bijkomende input van de compressie. Op die manier is een compressie mogelijk die minder bits nodig heeft dan een entropie van de 0-de orde. Bij LOCO-I bestaat de context uit de pixels links, linksboven, boven en rechtsboven.

4.2.17.2 Voorspelling

In deze stap wordt de waarde van de volgende pixel voorspeld. Dat gebeurt door het uitvoeren van steeds dezelfde primitieve test, de *predictor*. Die test is vrij eenvoudig en bestaat uit drie mogelijke berekeningen naargelang aan

35 Weinberger, Seroussi, et al. (1998); Weinberger en Seroussi (1999); ISO/IEC (1997).

bepaalde voorwaarden voldaan is. De nodige berekeningen hiervoor zijn minimum, maximum en optellen en aftrekken.

De *predictor* is zo opgesteld dat verticale en horizontale kleurovergangen gedetecteerd worden. Wanneer er links naast de huidige pixel een verticale overgang is, zal de output de pixel bovenaan zijn. Wordt er een horizontale rand gedetecteerd boven de huidige pixel, dan wordt de output de linkse pixel. Indien er geen overgang blijkt te zijn, dan worden de linkse en de bovenste pixel opgeteld en verminderd met de pixel linksboven.

4.2.17.3 Contextbepaling

Aangezien het slechts om een voorspelling gaat, zal er steeds sprake zijn van een mogelijke fout. Deze wordt de voorspellingsfout of het residu genoemd. Het 'context model' dat dit residu bepaalt, wordt aangeduid door de contextvector Q=(q1,q2,q3), waarbij q1, q2 en q3 de lokale overgangen (verschillen) of gelijkenissen voorstellen: q1=rechtsboven-links, q2=links-linksboven, q3=linksboven-boven.

4.2.17.4 Residu codering

Ten slotte wordt het residu gecodeerd met behulp van Golomb-codes, die ideaal zijn voor de codering van tweezijdige geometrische verdelingen.

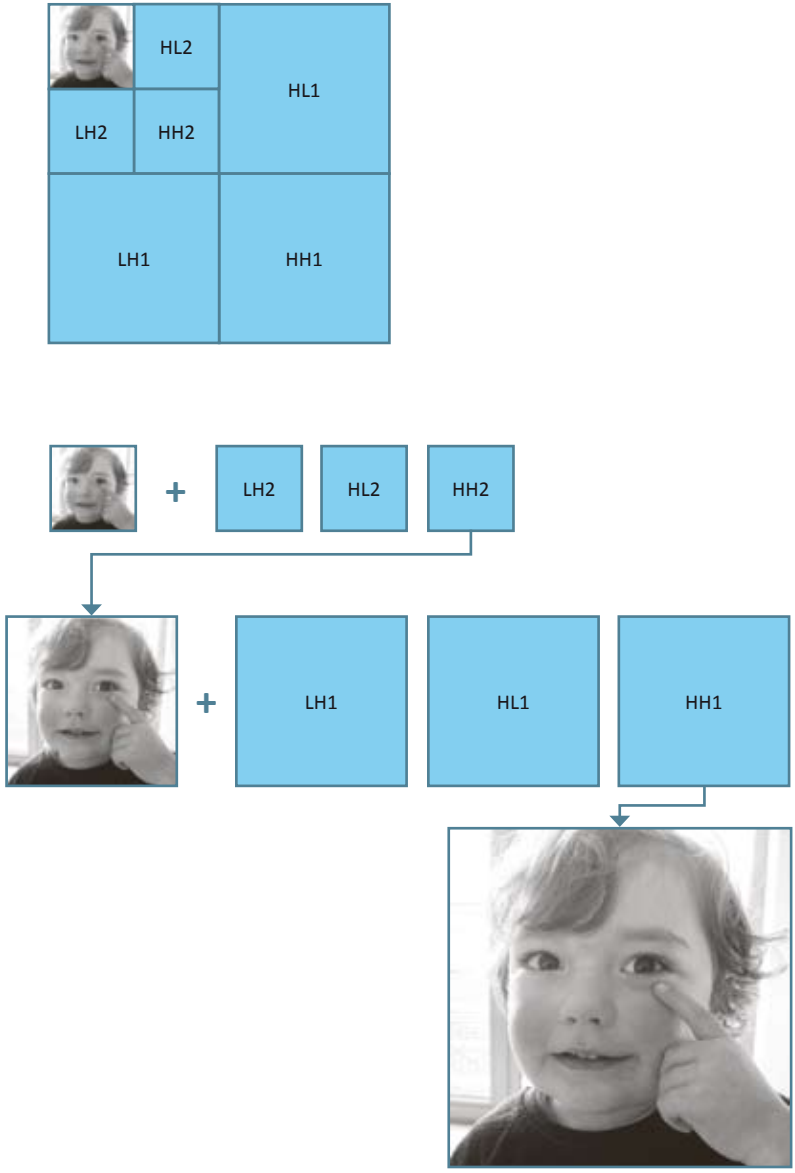
4.2.18 JPEG 2000[36]

JPEG2000 is de nieuwste standaard voor beeldcompressie van de Joint Photographic Experts Group (JPEG).

De discrete cosinustransformatie (DCT) is bij JPEG2000 vervangen door de *wavelet* transformatie, een overgang van een lokale, blokgebaseerde transformatie naar een globale beeldtransformatie. Hierdoor worden de storende blokartefacten bij een laag bitdebiet vermeden ten koste van meer uitgesmeerde beelden. De *wavelet* transformatie leent zich bovendien erg goed voor schaalbare of ingebedde codering, waarbij de geleverde prestaties door de hiërarchische en progressieve DCT-mode van de oude JPEG-standaard overschaduwd worden. Om de complexiteit te verlagen, kan het beeld alsnog opgesplitst worden in verschillende delen die afzonderlijk gecomprimeerd worden. Dit verlaagt het geheugengebruik bij de verwerking.

36 JPEG (2007); Santa Cruz, Ebrahimi, et al. (2000).

In vergelijking met de JPEG-standaard kent JPEG2000 een superieure beeldkwaliteit. Vooral bij lage bitdebieten vallen de artefacten van JPEG2000 minder op aangezien dan geen blokvorming optreedt. In vergelijking met andere *lossless* standaarden zoals PNG en JPEG-LS heeft JPEG2000 een iets lagere compressieratio maar kent ze wel het voordeel van de schaalbaarheid. Doorgaans wordt immers een diadische *wavelet*-transformatie gebruikt. Hierbij wordt het beeld bij de encodering gehalveerd in een verticale en een horizontale richting en wordt de nodige informatie voor de reconstructie van de volledige resolutie samen met het verkleinde beeld opgeslagen (de eerste afbeelding van figuur 9). Bij de decodering van een beeld wordt van het verkleinde beeld uitgegaan en wordt de rest van de informatie gebruikt om het beeld terug op te schalen (de tweede afbeelding van figuur 9).



Figuur 9: JPEG2000

JPEG2000 heeft nog niet een zodanig grote ingang gevonden als zijn voorganger JPEG. Dat is enerzijds te wijten aan licentieclaims met betrekking tot de oorspronkelijke JPEG-standaard en anderzijds aan de hogere complexiteit. Ondertussen zijn die problemen van de baan en begint JPEG2000 zowel voor *low-end* als *high-end devices* zijn intrede te maken. JPEG2000 is ondertussen wel goed doorgebroken op *high-end*-markten, zoals videobewaking en medische beeldvorming, waar zowel kwaliteit als schaalbaarheid van belang is. Ook voor digitale cinema werd MJPEG2000, die de JPEG2000-standaard gebruikt, als codeerstandaard geselecteerd.

4.2.19 GIF[37]

GIF, of Graphics Interchange Format, is een formaat ontwikkeld door CompuServe. De naam geeft al aan dat het formaat minder geschikt is voor natuurlijke beelden maar vooral gericht is op *graphics*, getekende beelden, met continue kleuren. Het formaat ondersteunt tot 8 bits per pixel of dus 256 kleuren. Deze kleuren worden per beeld afzonderlijk gekozen en kunnen dus per beeld verschillen. Verder biedt het formaat ook ondersteuning voor animaties en transparante achtergronden.

De compressie van GIF gebeurt door het aantal kleuren in een beeld te beperken, waardoor de kleuren aan de hand van minder bits voorgesteld kunnen worden. Afbeeldingen met weinig kleuren kunnen dus sterk gecomprimeerd worden. Om nog verdere compressie mogelijk te maken, wordt gebruik gemaakt van de Lempel-Ziv-Welch-compressie. Deze technologie is echter gepatenteerd door Unisys. Dat was een van de belangrijkste redenen voor de ontwikkeling van de rechtenvrije grafische bestandsindeling PNG. In de loop van 2003 en 2004 zijn de patenten op het LZW-algoritme verlopen en volgens de Free Software Foundation is het laatste patent in verband met GIF-compressie in augustus 2006 verlopen.

4.2.19.1 Dithering

Doordat een GIF-bestand maximaal 256 kleuren kan bevatten, is het niet erg geschikt voor kleurenfoto's. Toch kan men aan de hand van enkele technieken, waaronder *dithering*, proberen de gevolgen van het beperkte kleurenpalet te beperken. Het is echter moeilijk de geschikte kleuren voor het palet te kiezen. Elk programma volgt hiervoor een eigen methode.

37 Unisys (2009); CompuServe Inc. (1990); Blackstock (z.j.).

Bij *dithering* wordt de kleur van een pixel niet enkel bepaald door de waarde van de originele pixel maar ook door de afwijkende kleuren van de omliggende pixels. Op die manier ontstaat een vrij korrelig patroon dat gemiddeld precies de juiste kleuren heeft. Als men de afbeelding op een afstand bekijkt, zodat individuele pixels niet meer zichtbaar zijn, ziet men nauwelijks dat het aantal kleuren beperkt is. Deze techniek wordt ook bij printen toegepast.

4.2.20 PNG[38]

PNG, voluit Portable Network Graphic, is een bestandsformaat voor de opslag van afbeeldingen met verliesloze compressie. Het formaat werd ontwikkeld als patentvrij alternatief voor het GIF-formaat. Mettertijd werd PNG steeds populairder. Momenteel kan bijna elk beeldverwerkingprogramma het formaat aan en sinds Internet Explorer 7 ondersteunt ook Microsoft het formaat. *Open succes*programma's gebruikten het formaat al langer.

In vergelijking met BMP en TGA levert PNG dezelfde kwaliteit (verliesloos) en neemt het toch relatief weinig ruimte in. Tegenover GIF ondersteunt PNG een groter kleurenpallet, namelijk zestien miljoen kleuren in plaats van 256. Net zoals een GIF-afbeelding kunnen PNG-afbeeldingen ook een pallet van 256 kleuren hebben. Hierdoor kan de compressie nog aanzienlijk verhoogd worden en kan PNG gebruikt worden om zowel geheugenruimte voor eenvoudige afbeeldingen te sparen als om verliesloze opslag van afbeeldingen te voorzien.

Net als GIF ondersteunt PNG ook een alfa-kanaal waarin zogenaamde transparantiewaarden worden opgeslagen. Die bepalen de mate van transparantie van een pixel. PNG kent ook een extensie voor de ondersteuning van animaties, namelijk APNG of Animated PNG.

4.2.21 TIFF[39]

TIFF staat voor Tagged Image File Format. Momenteel is TIFF, naast JPEG en PNG, het meest courante formaat voor de opslag van beelden. Het is tevens het meest universele en ondersteunde formaat voor alle platformen: MAC, Windows en UNIX.

TIFF was medio jaren tachtig oorspronkelijk ontworpen als een gemeenschappelijk beeldformaat voor *desktop*-scanners. Aanvankelijk was TIFF enkel

38 Adler, Boutell, et al. (2003).

39 Adobe (2008a); Adobe Developers Association (1992).

een binair beeldformaat, met slechts twee mogelijke waarden voor een pixel. Naarmate de scanners evolueerden en de opslagruimte toenam, begon TIFF grijswaarden te ondersteunen en uiteindelijk ook kleurwaarden. TIFF ondersteunt nu de meeste kleurenruimtes, zoals RGB, CMYK en YcbCr.

TIFF onderscheidt zich van de meeste beeldcompressieformaten door de flexibele beeld-*header*, die men bovendien zelf kan definiëren. De *header* kan samengesteld worden door een set van informatievelden of tags. Die tags kunnen de meest elementaire informatie bevatten, zoals beeldgrootte of bitvolgorde, maar ze kunnen ook bijvoorbeeld de rechten beschrijven. Het is zelfs mogelijk om 'private' tags te gebruiken die volgens de betreffende applicatie specifieke informatie bevatten. Het voordeel hiervan is dat de data vergezeld kunnen worden van om het even welke informatie. TIFF kan uiteindelijk beschouwd worden als een containerformaat voor beelden. Ze biedt ondersteuning voor *multipage*, meerdere beelden binnen één bestand, en *multilayer*, meerdere lagen binnen één beeld.

Aan de hand van twee tags kan ook de gebruikte compressie en de kleurenruimte worden gedefinieerd. TIFF laat dus toe om het even welke compressie te gebruiken in combinatie met om het even welke kleurenruimte, als de draagbaarheid van het bestand buiten beschouwing wordt gelaten. TIFF kan met of zonder compressie gebruikt worden. Zo wordt G3-compressie gebruikt als de standaard voor fax en *multi-page* bestanden. Het is ook mogelijk om de verliesloze compressie LZW toe te passen. De compressie zal dan herhalende identieke *strings* detecteren en die vervangen door één instantie, zodat die zonder informatieverlies terug gedecodeerd kunnen worden. Het gebruik van deze compressie veroorzaakt wel een vertraging in het openen en opslaan van de bestanden. LZW is meest effectief wanneer *solid indexed colors* gecomprimeerd worden en is minder effectief voor 24bit *continuous* fotoformaten. LZW is effectiever voor grijswaarden dan voor kleurenbeelden. 48bit-beelden leveren nauwelijks een compressie op. Zoals eerder vermeld, ondersteunt TIFF beelden die uit verschillende pagina's of lagen bestaan, waardoor een TIFF-bestand bijvoorbeeld een vector-gebaseerde *clipping path* kan bevatten.

Een ander krachtig mechanisme van TIFF is zijn ondersteuning van verschillende datatypes, van *signed* of *unsigned integers*, *floating point*-waarden tot zelfs complexe datastructuren. Samen met de mogelijkheid om verschillende beeldkanalen op te slaan, maakt van TIFF een heel handig formaat voor wetenschappelijke data. Zonder compressie wordt TIFF vooral gebruikt voor het archiveren van beelden waarvan de kwaliteit zeer belangrijk is.

Aan deze flexibiliteit zijn ook nadelen verbonden. Nieuwe types zijn gemakkelijk te vormen, maar dit kan dan weer leiden tot incompatibiliteit. Dat kan vermeden worden door de standaard TIFF-types te gebruiken die door de meeste applicaties ondersteund worden. Een ander nadeel van TIFF is de beperking van de beeldgrootte, die maximum 4 GB bedraagt. Momenteel zoekt men naar een oplossing voor deze beperking, namelijk het BigTIFF-bestandsformaat als opvolger van het TIFF-formaat.

4.3 Fysieke containers

4.3.1 WAV[40]

WAV, of Waveform Audio Format, is een bestandsformaat voor de opslag van audio op PC. WAV slaat de audiodata ruw op. Door het verliesloze karakter van ruwe audio kunnen deze WAV-bestanden echter heel groot worden.

Een WAV-bestand is opgebouwd uit zogenaamde *chuncks*. Deze *chuncks* geven informatie over het geluid of bevatten het geluid zelf. Naast deze *chuncks* bevat een WAV-bestand een *header* met onder meer informatie over de gebruikte formaatstructuur.

De maximale grootte van een WAV-bestand bedraagt 4GB, wat overeenkomt met ongeveer 405 minuten geluid in CD-kwaliteit (44.1kHz, 16 bit, stereo) en 62 minuten in DVD-audiokwaliteit (tot 192kHz, tot 24 bit, stereo). Om die beperkingen weg te werken, werd later het W64-formaat ontworpen dat de grootte van het bestand in 64 bits in plaats van in 32 bits beschrijft. De EBU heeft om dezelfde reden het RF64-formaat ontwikkeld. Dat formaat biedt onder meer ook ondersteuning voor maximaal 18 *surround*-kanalen. Naast ruwe audio ondersteunt de WAV-container ook andere codecs, zoals GSM, ADPCM en MPEG Layer-3.

4.3.2 AIFF[41]

AIFF, of Audio Interchange File Format, is de Apple Macintosh-tegenhanger van WAV. Het formaat komt grotendeels overeen met het WAV-formaat van Microsoft. Maar terwijl bij WAV de *samples* in een *little-endian-byte*-volgorde

40 IBM Corporation en Microsoft Corporation (1991); Windows Hardware Developer (2007).

41 Kabal (2005); Apple Computer (1989); Apple Computer (1991).

worden opgeslagen, gebeurt dit bij AIFF in *big-endian-byte*-volgorde. Sinds Mac OS X heeft Apple echter een nieuw type AIFF gecreëerd dat in *little-endian-byte*-volgorde wordt opgeslagen. Dat had te maken met de overschakeling naar het gebruik van Intel-processoren, die *little-endian-byte*-volgorde gebruiken.

AIFF is ook opgebouwd uit een *header* en zogenaamde *chuncks*, die zowel de informatie over het geluid als het geluid zelf kunnen bevatten.

4.3.3 XMF[42]

XMF, of eXtensible Music Format, is een familie van muziekgerelateerde formaten, ontworpen door de MIDI Manufacturer's Association. XMF heeft tot doel één of meer bestanden in bestaande formaten, zoals MIDI en WAV, samen te voegen.

XMF bestaat uit twee delen: het XMF Meta-File Format en een reeks XMF File Types, die het XMF Meta-File Format gebruiken. Tot nu toe zijn XMF Type 0, XMF Type 1 en Mobile XMF gedefinieerd, die allemaal op MIDI gericht zijn.

Een XMF Meta-bestand bestaat uit verschillende nodes die hiërarchisch gegroepeerd zijn, vergelijkbaar met een bestandssysteem met folders en bestanden. Bij XMF worden hiervoor de respectieve begrippen *containers* en *resources* gebruikt. Een node is ofwel een *container*, ofwel een *resource*. Een *resource node* bevat dan een verwijzing naar een intern bestand of een URL die verwijst naar een extern bestand.

4.3.4 MPEG-21Part 9 (File Format)[43]

Binnen de MPEG-4 standaard (ISO/IEC 14496) (zie ook §4.3.8 MP4) zijn er verschillende onderdelen die bestandsformaten definiëren voor de opslag van tijdsgebaseerde media, zoals audio en video. Die zijn echter allemaal gebaseerd op en afgeleid van het ISO Basis Media File Format (ISO/IEC 14496-12), dat een hiërarchisch gestructureerde, mediaonafhankelijke definitie omvat, die ook gepubliceerd is als deel van de JPEG2000-familie. Het bevat onder meer een basis containerstructuur en een definitie van tijdssequenties van multimedia binnen een dergelijk containergestructureerd bestand.

42 MIDI (2004).

43 Bormans en Hill (ed.) (2002).

Het bestandsformaat MPEG-21 (ISO/IEC 21000-9) gebruikt de structurele definitie van een containergebaseerd bestand, zoals gedefinieerd in het ISO Basis Media File Format, maar zonder de extra definities voor tijdsgebaseerde media. Het definieert de opslag van een MPEG-21 Digital Item (zie ook §6.5 MPEG-21/DIDL) èn alle eventueel bijkomstige (meta)data in datzelfde bestand, zoals foto's, filmpjes of andere niet-XML-data. Containergebaseerde bestandsformaten laten immers toe om flexibele bestanden te creëren die meerdere containers, met mogelijk verschillende specificaties, bevatten. Binnen MPEG-21 wordt een generieke meta-container op bestandsniveau gebruikt om de beschrijving (MPEG-21 DID) van de *resource* weer te geven en voorts ook een lijst met alle verwante *resources*, al dan niet in een subcontainer ingebed of als extra verwijzing naar een ander bestand. De volledige flexibiliteit en kracht van een dergelijke URL-gebaseerde meta-container kan aan de hand van het volgende voorbeeld aangetoond worden:

- *Items* (andere bestanden) die nodig zijn voor het te beschrijven Digitale Items kunnen geïntegreerd worden in ditzelfde MPEG-21-bestand of in een ander bestand (al dan niet ook een MPEG-21-bestand).

- *Items* (bestanden) geïntegreerd in dit MPEG-21-bestand of in andere bestanden kunnen gefragmenteerd zijn en ook de fragmenten zelf kunnen zeer fragmentarisch zijn.

- *Items* (bestanden) kunnen beschermd zijn en er kan binnen het bestand aangegeven worden hoe daarmee omgegaan moet worden.

- *Items* (bestanden) kunnen een naam krijgen, waardoor er gemakkelijk kan naar verwezen worden in een MPEG-21-bestand of zelfs vanuit een extern bestand.

4.3.5 OGM/OGG[44]

OGM, of OGG Media, is een containerformaat dat een uitbreiding vormt op het OGG-containerformaat van Xiph.org.[45] OGM voegt aan OGG onder meer de ondersteuning voor andere codecs toe dan diegene die ontworpen zijn door Xiph.org (Speex, Theora en Ogg Vorbis).[46] OGM biedt namelijk ook ondersteuning aan videocodecs die gebruik maken van VfW en audiocodecs

44 Zie ook de tekst in Mannens, Paridaens, et al. (2007), p. 74.

45 Xiph.org (1994-2008b).

46 Respectievelijk Xiph.org (1994-2008a); Xiph.org Foundation (2008) en Xiph.org (2008b).

die ACM gebruiken. Net zoals bij videobitstromen wordt bij de audiobitstromen ondersteuning geboden voor een variabele *bitrate*. Algemeen wordt OGM als een tussenfase beschouwd ten opzichte van de andere containerformaten, zoals Matroska, die volgroeid zullen zijn en dezelfde mogelijkheden zullen bieden. Tot die mogelijkheden behoren onder meer de ondersteuning voor hoofdstukken, meerdere ondertitels en meerdere audiokanalen.

4.3.6 Matroska (MKV/MKA)[47]

Matroska is een open standaard multimediacontainerformaat dat gebaseerd is op EBML (Extensible Binary Meta Language).[48] Dat is een binair *byte*-gebonden formaat dat gebaseerd is op de principes van XML.[49]

Een Matroskabestand bestaat uit een *header* met informatie over de gebruikte EBML-versie en het bestandstype, in dit geval dus een Matroskabestand. De *header* wordt gevolgd door de *Metaseek*-sectie die de plaats aanduidt van de verschillende andere secties in het bestand. Dat is noodzakelijk omdat iedere sectie in principe overal in het bestand kan voorkomen en men dus het hele bestand zou moeten *parsen* om de informatie te zoeken. Er zijn secties voorzien voor onder andere kanaalinformatie, hoofdstukinformatie en *tags*.[50]

Matroska kent twee onderverdelingen: MKV, dat zowel video als audio kan bevatten, en MKA, dat enkel bedoeld is voor audio. Ondersteuning is mogelijk voor welhaast alle video- en audioformaten, zoals MPEG-1, MPEG-2, MPEG-4, Quicktime, Real, Theora voor video en MP1, MP2, MP3, PCM, AC3, FLAC, AAC voor audio. Hierbij worden zowel variabele audio *bitrate* als variabele *framerate* ondersteund. Verder maakt Matroska het ook mogelijk bestanden van om het even welk type toe te voegen. Zo kunnen bijvoorbeeld transcripties aan het bestand toegevoegd worden.

Matroska kan een onbeperkt aantal videostromen, audiostromen, afbeeldingen en ondertitels bevatten en laat ook toe lettertypes toe te voegen voor bijvoorbeeld de ondertitels. Matroska biedt verder ook een robuuste ondersteuning voor *streaming*, hoofdstukken en DVD-achtige menu's.

47 Zie ook de tekst in Mannens, Paridaens, et al. (2007), p. 74-75.

48 Corecodec (2005-2009a).

49 EBML (z.j.).

50 Voor een voorbeeld van audiotags, cf. Corecodec (2005-2009b).

4.3.7 MXF[51]

MXF, of Material eXchange Format, is een standaard containerformaat voor professionele video en audio. Het formaat wordt gevormd door een set SMPTE-standaarden. MXF is een open bestandsformaat dat specifiek ontworpen werd om A/V-materiaal samen met de geassocieerde data en metadata tijdens de productiefase uit te wisselen. De ontwikkeling van MXF gebeurt door een samenwerking van verschillende fabrikanten en de organisaties Pro-MPEG, EBU en de AAF Association.

MXF is een veelzijdig bestandsformaat dat voor de volgende taken kan instaan:

- bewaren van eenvoudig afgewerkt materiaal en bijhorende metadata (*tape replacement*),
- bewaren van materiaal in een *streamable* formaat, dat bekeken kan worden terwijl het doorgestuurd wordt,
- verpakken en bewaren van een *playlist* met bestanden en hun bijhorende informatie voor synchronisatie,
- verpakken van om het even welk compressieformaat,
- bewaren van *cuts-only* EDL's (een Editing Decision List bevat de gegevens die gebruikt zijn bij audiovisuele *content editing*-systemen en doet dienst als een soort tijdslijn) en van het eigenlijke materiaal waarop het van toepassing is.

Zowel *real-time streaming* (eindgebruikers kijken 'live') als bestandstransfers (tussen computersystemen onderling) zijn belangrijk in een veralgemeende A/V-constellatie. Daarom is het nodig dat beiden compatibel en uitwisselbaar zijn. MXF is daarom ook zo ontworpen dat het een *streaming*-formaat is, waardoor het naadloos een brug kan vormen tussen de beide transfertypes. MXF ondersteunt alle mogelijke video- en audioformaten en laat ook toe dat willekeurige bestanden worden toegevoegd. Zo kunnen ook transcripties, beelden, enzovoort worden toegevoegd.

51 Zie deels ook de tekst in Mannens, Paridaens, et al. (2007), p. 75; SMPTE (z.j.).

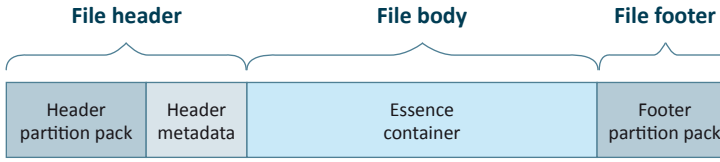
Zoals vermeld, was interoperabiliteit de belangrijkste doelstelling tijdens de ontwikkeling van MXF. Het formaat is bijgevolg:

- *Cross-platform*: het is volledig onafhankelijk van een netwerkprotocol of besturingssysteem.
- Compressieonafhankelijk: er worden geen converties tussen verschillende codecs uitgevoerd, maar het is gemakkelijk om verschillende codecformaten, alsook ongecomprimeerde data, in eenzelfde omgeving te beheren.
- Een brug tussen *streaming* en *transfers*: er is een volledig transparante uitwisseling in twee richtingen mogelijk tussen al dan niet *streamable* bestanden.

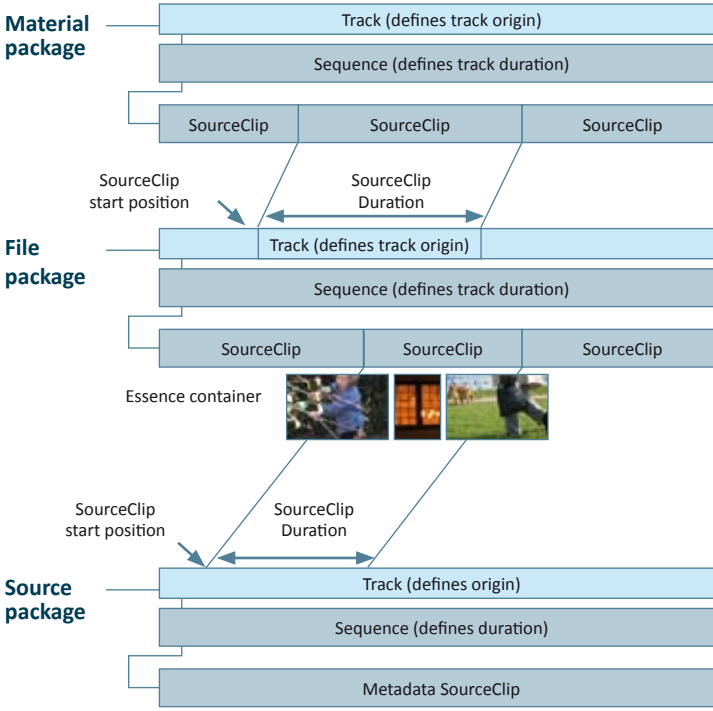
Een MXF-containerbestand bestaat uit een *header*, een *footer* en een *body* (met het eigenlijke A/V-materiaal). Elk *item* in een MXF-bestand is KLV-geëncodeerd (Key Length Value), waardoor het uniek geïdentificeerd kan worden door een 16-*byte*-sleutel en de lengte. Het kennen van de lengte van elk *item* in het MXF-bestand laat immers toe om eenvoudige *decoders* te implementeren en stukken 'onbekende' data links te laten liggen tot er een (volgende) *decoder* gebruikt wordt die het betreffende stuk data wel kan interpreteren. In de handige *header* van het MXF-bestand worden metadata, tijdsparameters en de synchronisatie bijgehouden. Informatie over de synchronisatie en de beschrijving van het materiaal worden op drie verschillende niveaus bewaard:

- *Material Package* (MP): hier wordt de tijdslijn van het bestand bewaard.
- *File Package* (FP): het eigenlijke materiaal wordt hierin beschreven.
- *Source Package* (SP): bevat een beschrijving van de afgeleiden van dat materiaal (bijvoorbeeld EDL's).

Ieder *package* (MP, FP of SP) kan een eigen hoeveelheid *tracks* (audio, video en/of metadata) bevatten. Elke afzonderlijke *track* kan dan weer een sequentie van *SourceClips* bevatten.

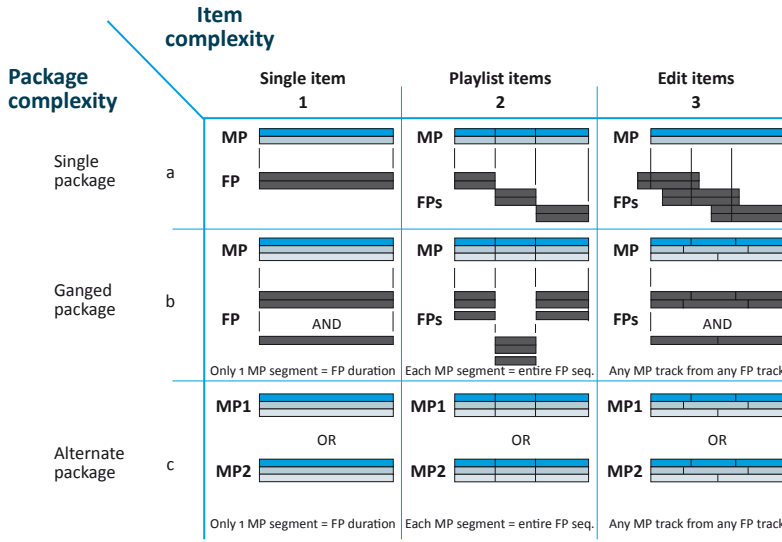


Simple MXF file structure



Figuur 10: MXF-structuur

Om de complexiteit en de vele 'vrije' mogelijkheden van MXF te beheren, bestaan bovendien een aantal 'operationele patronen' (*Operational Patterns*). Hieronder wordt de *grid* weergegeven die verticaal onderverdeeld wordt naargelang de complexiteit van de tijdslijn in het MXF-bestand en horizontaal onderverdeeld is naargelang het aantal *packages* in datzelfde MXF-bestand.



Figuur 11: MXF-bestand

MXF ondersteunt ook de toevoeging van metadata en enkele professionele functies, zoals een volledige *timecode* platformonafhankelijkheid. MXF-metadata kunnen de volgende informatie bevatten:

- bestandsstructuur
- titel en trefwoorden
- ondertitels
- referentienummers
- annotaties
- versienummer
- locatie, tijd en datum

4.3.8 MP4[52]

MP4 of MPEG-4 part 14 is een multimedia containerformaat dat een onderdeel vormt van de MPEG-4-standaard. MP4 kan zowel audio- als videostromen bevatten. Hierbij ondersteunt MP4 de standaard videoformaten MPEG-1, MPEG-2, MPEG-4 en MPEG-4 AVC. Voor audio worden de standaarden (HE)-AAC, MP3, MP2, MP1, CELP, TwinVQ, Vorbis en Apple Lossless ondersteund.

52 Zie ook de tekst in Mannens, Paridaens, et al. (2007), p. 75-76; Digital Formats (2007b).

Wanneer een MP4-container enkel audio bevat, krijgt die vaak de extensie M4A. Die extensie wordt onder andere voor podcasts gebruikt.

Naast de gewone audio- en videostromen kent MP4 ook zogenaamde private stromen. Deze private stromen kunnen om het even welke gegevens bevatten. Zo gebruikt Nero de stromen om ondertitels in Dvd-formaat toe te voegen.

MP4 ondersteunt voorts ook afbeeldingen, hyperlinks, ondertitels, hoofdstukken, variabele audio *bitrate* en variabele *framerate*.

4.3.9 3GP[53]

3GP, of 3G Protocol, is een multimedia containerformaat, ontworpen door de Third Generation Partnership Project (3GPP), voor het gebruik met 3G mobiele telefoons. 3GP is een vereenvoudigde versie van het MP4-containerformaat en is ontworpen met als doel een vermindering van opslag- en bandbreedtevereisten te bereiken. 3GP ondersteunt zowel MPEG-4 Part 2, H.264/AVC als H.263 voor video en AMR-NB, AMR-WB, AMR-WB+ en (HE)-AAC-LC voor audio. 3GP biedt ook ondersteuning aan variabele audio *bitrates*, variabele *framerates* en ondertitels. 3GP-bestanden kan men zowel streamen als downloaden (zie bijvoorbeeld MMS-berichten).

4.3.10 ASF[54]

ASF, of Advanced Systems Format, is een containerformaat ontworpen door Microsoft als onderdeel van het Windows Media Framework. De vroegere naam Advanced Streaming Format geeft het hoofddoel weer van het containerformaat, namelijk *streaming*. Van ASF bestaan twee versies. Versie 1.0 is veruit de meest gebruikte maar is gesloten; de opbouw is dus op enkele details na niet gekend. Versie 2.0 is open maar wordt nauwelijks gebruikt.

ASF ondersteunt bijna alle video- en audioformaten die werken via VfW en ACM, maar het formaat wordt meestal gebruikt in combinatie met de eigen formaten van Microsoft. Verder ondersteunt ASF ook metadata, zoals artiest en titel, variabele audio *bitrate*, variabele *framerate*, hoofdstukken en ondertitels. ASF biedt ook foutcorrigerende technieken en een *digital rights management framework* aan.

53 Zie ook de tekst in Mannens, Paridaens, et al. (2007), p. 76; 3GPP (2009).

54 Zie ook de tekst in Mannens, Paridaens, et al. (2007), p. 76; Windows Media (2004).

4.3.11 MOV[55]

MOV is een multimedia containerformaat dat door Apple ontworpen is als basis voor het MP4-containerformaat. De container kan zowel video, audio als hoofdstukken bevatten en ondersteunt variabele audio *bitrate* en variabele *framerate*. MOV ondersteunt alle formaten die de Quicktime codecmanager ondersteunen, zoals MPEG-4 en de Sorensen codec, en alle audioformaten die de soundmanager en core-audio ondersteunen, zoals AIFF, WAV en MP3.

In een MOV-container kan elk kanaal voorgesteld worden door de mediastroom zelf of door een referentie naar de mediastroom in een ander bestand. In de MOV-container worden de kanalen in een hiërarchische structuur van atomen geplaatst. Die atomen kunnen ofwel 'ouder' zijn dan andere atomen ofwel bevatten ze zelf media of data.

MOV-containers bevatten een tijdslijn die losstaat van de mediastromen. Hierdoor kunnen MOV-containers eenvoudig worden aangepast zonder dat de mediastromen gekopieerd moeten worden.

4.3.12 AVI[56]

AVI, of Audio-Video Interleaved, is een multimedia containerformaat dat ontworpen is door Microsoft. AVI-containers kunnen meerdere audio- en videokanalen bevatten. Een AVI-container bestaat uit een *header* met informatie over de video, zoals breedte, hoogte en *framerate*, en de eigenlijke data. Verder kan een container ook een index bevatten, die toelaat te navigeren binnen de container. AVI-containers ondersteunen nagenoeg alle audio- en videoformaten die beschikbaar zijn via DMO, ACM en VfW. AVI ondersteunt variabele audio *bitrates*, mits enkele beperkingen (namelijk niet via ACM), en variabele *framerates*. Ondertitels en hoofdstukken worden ook ondersteund via modificaties, maar dan buiten Microsoft.

4.3.13 FLV[57]

FLV, of Flash Video, is een containerformaat van en ontworpen door Adobe, dat onder meer door Google Video en YouTube gebruikt wordt. FLV kan slechts één video en één audiostroom per bestand bevatten. Verder kan

55 Zie ook de tekst in Mannens, Paridaens, et al. (2007), p. 76-77. Voor MOV specificaties, cf. Apple (2005).

56 Zie ook de tekst in Mannens, Paridaens, et al. (2007), p. 77; McGowan (1996-2004).

57 Zie ook de tekst in Mannens, Paridaens, et al. (2007), p. 77; Digital Formats (2007a) en de officiële specificatie, Adobe (2008b).

een container ook Flash-content bevatten. FLV ondersteunt de videoformaten Sorensen, VP6 en Screen Video en de videoformaten MP3, Nellymoser, ADPCM en PCM. Een FLV-container kan op verschillende manieren bij de eindgebruiker terechtkomen: via download, in een flash-animatie of door *streaming* via het RTMP-protocol.

In de nieuwe versie van FLV wordt ook ondersteuning voor H.264/AVC en HE-AAC geboden.

4.3.14 RealMedia[58]

RealMedia is een multimedia containerformaat dat door RealNetworks ontworpen is.[59] Realmedia is een populair formaat voor het streamen van audio en video via het internet. Het formaat ondersteunt de videoformaten RealVideo 8-9-10 en de audioformaten HE-AAC, Cook, Vorbis en RealAudio Lossless. Voorts biedt Realmedia ook ondersteuning aan variabele *framerate*, ondertitels en met behulp van de RMVB-extensie ook aan variabele *bitrates*.

58 Zie ook de tekst in Mannens, Paridaens, et al. (2007), p. 77.

59 Realnetworks (2009).

5 Informatie over de data

5.1 Inleiding

Doorgaans worden metadata omschreven als 'data over data'. Metadata bieden (gestructureerde) informatie over een bron of *resource*. Een bron is ieder mogelijk object of subject waarover informatie opgeslagen kan worden. Dit kan een tekst, een fysiek object of software zijn, maar evenzeer een persoon, gebeurtenis of dienst.

Naargelang de soort informatie die de metadata bevatten, kan men verschillende types onderscheiden: administratieve metadata (rechten, plaats,...), beschrijvende metadata (auteur, titel,...), bewaringsinformatie (staat, verhuizingen,...), technische metadata (formaat, encryptie, softwareversie,...) en informatie over het gebruik van de data.

Geavanceerdere metadata geven ook onderlinge betekenisvolle relaties aan: zo worden bijvoorbeeld de persoon 'interviewer' en de bron 'mondelinge historische bron' door de relatie 'creatie' met elkaar verbonden.

Metadatastandaarden kunnen worden ingedeeld volgens zoekmogelijkheden. Naast 'gewone' metadatastandaarden (zoals MARC/MARC21) bestaan er namelijk ook semantische standaarden die 'intelligente' zoekmethoden ondersteunen en daarbij rekening houden met de betekenis van zoektermen of gebruik maken van thesauri. Een bekend probleem is bijvoorbeeld dat zoekmachines gegevens zoals 'H. Claus' en 'Hugo Claus' als twee verschillende personen beschouwen. Om dergelijke beperkingen te ondervangen, worden vaak woordenboeken met afgesproken termen gebruikt, zogenaamde ontologieën. Ook thesauri kunnen dus helpen door bij zoekacties termen met een gelijkaardige betekenis aan de query toe te voegen.

Metadata hebben ook verschillende functies: in de eerste plaats zorgen metadata ervoor dat men relevante informatie snel en gemakkelijk terugvindt. Maar ze helpen ook bij de organisatie van elektronische bronnen en verzekeren hun interoperabiliteit. Metadata kunnen data van een digitale identificatie voorzien en de archivering en preservering ondersteunen.

Tegenwoordig bestaan er heel veel metadatastandaarden, waardoor de keuze voor een enkele standaard allerminst voor de hand ligt. Afhankelijk van het toepassingsgebied verschillen de metadatastandaarden immers wat de semantiek en de specificiteit betreft.

De meest courante en eenvoudige standaard is Dublin Core (§5.2.1), als het ware de lingua franca van de metadatastandaarden. De kracht van deze standaard wordt ongetwijfeld bepaald door zijn eenvoud en algemene toepasbaarheid. Dublin Core bestaat uit vijftien velden, waarmee elke bron beschreven kan worden. Niettemin zijn deze beschrijvingen meestal te beperkt. Daarom wordt Dublin Core vaak gebruikt als een bijkomende metadatastandaard naast een standaard die de bronnen nauwkeuriger beschrijft. Aangezien de meeste systemen met deze algemene standaard kunnen omgaan, zorgt een mapping van een meer specifieke standaard naar Dublin Core voor de nodige interoperabiliteit. De vijftien velden van Dublin Core zijn facultatief en herhaalbaar, waardoor bijna iedere metadatastandaard naar Dublin Core gemapt kan worden. Hierbij zal echter vaak dataverlies optreden aangezien niet alle velden inhoudelijk naar de vijftien velden van Dublin Core gemapt kunnen worden.

In wat volgt wordt een overzicht gegeven van de gangbare metadatastandaarden in het bibliotheekwezen, de media, de culturele sector en de archiefsector. Die indeling kan men echter niet altijd strikt aanhouden. Afhankelijk van het materiaal zijn er tussen de sectoren nogal wat overeenkomsten en sommige metadatastandaarden zijn dan ook in meerdere sectoren toepasbaar. Daarom worden in deze inleiding vrij summier de courante standaarden per sector overlopen, terwijl in het uitvoerige overzicht de indeling per sector achterwege wordt gelaten. De toelichtingen worden meestal geïllustreerd met een voorbeeldrecord in XML of RDF/XML, aangezien die de betreffende semantiek zichtbaar maken.[60]

In de omroepsector worden momenteel twee metadatastandaarden vaak gebruikt: P/META en SMEF-DM. P/META (§5.2.3) wordt binnen de Vlaamse omroepsector meest toegepast. De betreffende velden voorzien audiovisueel materiaal van typische informatie die de uitwisseling van programmagegevens mogelijk maakt. In de praktijk wordt zowel door de commerciële als publieke omroepen gebruik gemaakt van het IPEA-model, dat een subset van P/META is. Het IPEA-model is ontwikkeld in samenwerking met de Vlaamse

60 Voor de toelichting van de standaarden in de volgende paragrafen kon grotendeels uitgegaan worden van het overzicht dat in het rapport *Haalbaarheidsstudie naar een innovatieve applicatie voor de ontsluiting van mondelinge bronnen* in het kader van het project *Van horen zeggen* werd samengesteld. Cf. Mannens, Paridaens, et al. (2007).

omroepen (cf. §5.2.3.7). De BBC leverde een gelijkaardige inspanning, wat resulteerde in de standaard SMEF-DM (§5.2.4).

Binnen het bibliotheekwezen zijn vooral MARC/MARC21 en het FRBR-model courant. MARC (§5.2.5) is een standaard voor de representatie en de uitwisseling van bibliografische informatie. De belangrijkste functie van de standaard bestond er dan ook in boeken in een bibliotheek snel en eenvoudig achterhaalbaar te maken. MARC is heel verfijnd, waardoor de standaard complex is. De gedetailleerdheid zorgt ervoor dat men de bron zeer nauwkeurig kan beschrijven en toch is het formaat vrij compact. Dat laatste komt voort uit het feit dat de veldnamen, zoals 'plaats van publicatie', door een korte numerieke code vervangen zijn, wat de leesbaarheid van de standaard dan weer bemoeilijkt. FRBR (§5.4.1) is een conceptueel model in de bibliografische wereld, dat gericht is op de eindgebruiker. Het model is ontwikkeld om bepaalde gebruikersactiviteiten te vergemakkelijken, zoals de *retrieval* van records. De bibliografische entiteiten die in het model gedefinieerd worden, kan men in groepen opsplitsen, die op hun beurt weer kunnen worden onderverdeeld. Deze standaard is dus ook zeer granulair en laat een nauwkeurige beschrijving van de entiteiten toe. Het gebruik van een dergelijke granulaire standaard, en dat geldt ook voor MARC, brengt echter een zekere implementatiekost met zich mee. Naargelang de diepte van beschrijven zijn meer of minder inspanningen noodzakelijk.

Aan de hand van CDWA (§5.2.7) worden data uit kunstdatabanken beschreven. Het conceptueel kader van CDWA biedt richtlijnen aan voor de beschrijving en *retrieval* van informatie over kunstwerken, architectuur en ander cultureel materiaal. De standaard wordt dan ook voornamelijk in de cultuursector gehanteerd. CDWA bevat 512 categorieën en subcategorieën, waarvan een kleine subset de zogenaamde *core* vormt. Deze stelt de minimale informatie voor die nodig is om een werk te beschrijven en te identificeren. Een XML-schema van de *core*, namelijk CDWA Lite, draagt bij tot de implementatie van het schema. CDWA concordeert bovendien met OAI-PMH, dat instaat voor een vlotte uitwisseling van gegevens tussen verschillende bibliotheken. CIDOC-CRM (§5.4.2) is een ander gestandaardiseerd conceptueel model binnen de cultuursector. Het CIDOC Conceptual Reference Model (CRM) definieert en structureert concepten en relaties die toepasbaar zijn op de documentatie en beschrijving van cultureel erfgoed. CIDOC-CRM richt zich voornamelijk op de beschrijving van contextuele informatie. Het gaat dan in het bijzonder over de historische, geografische of theoretische achtergrond van de tentoongestelde items, waardoor hun waarde en betekenis toenemen. Net zoals Dublin Core geldt deze standaard vaak als een zogenaamde metadataspil om de interoperabiliteit van het systeem te vergroten. Terwijl

een beschrijving in Dublin Core echter behoorlijk voor wat gegevensverlies zorgt, is CIDOC-CRM voldoende uitgebreid om de uitwisseling zonder veel informatieverlies te laten verlopen.

In de archiefsector is ISAD(G) (§5.2.11) de voornaamste standaard voor de beschrijving van collecties en objecten. De standaard bevat verschillende regels maar voorziet niet in een eigen codering, waardoor ze veeleer als 'handleiding' voor de beschrijving van collecties geldt. Het gaat bijvoorbeeld om richtlijnen voor 'gelaagde' beschrijvingen (van collectie tot één object), het gebruik van referenties, titels, dateringen, enzovoort.

5.2 Descriptieve metadatastandaarden

5.2.1 Dublin Core

Dublin Core is een sectoroverschrijdende metadatastandaard.[61] Bij de ontwikkeling van de standaard werd zeker geen verfijning of complexiteit van een standaard als MARC vooropgezet. Met Dublin Core wordt namelijk een grootste gemene deler van metadatastandaarden uit verschillende sectoren beoogd, met het doel de onderlinge informatie-uitwisseling en zoekopdrachten te vereenvoudigen. Daarom moet men er rekening mee houden dat de semantiek van de elementen verschilt naargelang de sector.[62]

In Dublin Core is sprake van *resources*, *elements*, *qualifiers* en *schemes*. *Resources* zijn de objecten die beschreven worden aan de hand van vijftien elementen, waaronder *creator* en *rights*. Zo geeft het element *type* de aard van het object weer, bijvoorbeeld *sound,* waartoe ook mondelinge historische bronnen behoren. Deze vijftien elementen vormen het zogenaamde Dublin Core Simple-profiel. Ze zijn facultatief en herhaalbaar, waardoor bovendien bijna iedere metadatastandaard naar Dublin Core gemapt kan worden.[63] Dublin Core Qualified voegt drie extra elementen (waaronder het doelpubliek)

61 DCMI (2009b).

62 Zie ook DCMI (2009a) (documenten van DCMI), IANA (2007) (een thesaurus met mogelijke Media Types), DCMI (2006) (een forum voor wie Dublin Core metadata wil implementeren in een context van digitale langetermijnbewaring), DEN (2008) (toelichting van Digitaal Erfgoed Nederland bij Dublin Core), LOC (2003) (o.m. een afweging van Dublin Core en MODS in het kader van het project American Memory van het LOC).

63 Mogelijke mappingvoorstellen tussen Dublin Core en andere standaarden: Baca, Clarke, et al. (2009); LOC (2006b); LOC (2008f); LOC (2008a).

toe aan Simple Dublin Core en vult het profiel ook met *qualifiers* en *schemes* aan. *Qualifiers* worden gebruikt om elementen te specificeren, bijvoorbeeld dat het element *creator* een fotograaf of een auteur aangeeft. De *qualifiers* zijn niet aan regels gebonden, waardoor niet alle software die facultatieve metadata zal begrijpen. Software die de term 'fotograaf' niet herkent, interpreteert dit dan als *creator*. *Qualifiers* bezorgen aan de software die de gegevens verwerkt extra informatie, zonder daarbij aan compatibiliteit in te boeten.

De toevoeging *scheme* laat vervolgens toe aan te geven hoe elementen moeten worden ingevuld. Zo kan worden aangegeven dat het *subject* een trefwoord betreft uit een bepaalde thesaurus en geen vrij sleutelwoord is. In het attribuut *scheme* wordt dan de gevolgde thesaurus vermeld.

Behalve de standaardset van elementen kunnen ook andere elementen worden toegevoegd. Het is echter raadzaam elementen te gebruiken die van andere metadatastandaarden afkomstig zijn. Een nieuwe set elementen vormt dan een *application profile*.

Voordelen

- Dublin Core vereenvoudigt aanzienlijk het samenvoegen van metadata van instellingen die deze standaard gebruiken.
- Het nadeel van het beperkt aantal elementen van Dublin Core Simple kan verholpen worden door het gebruik van *qualifiers*.
- De standaard ondersteunt RDF-gebaseerde opslag.

Nadelen

- Dublin Core beperkt zich tot de beschrijving van *resources*, zoals boeken en geluidsfragmenten, en laat de beschrijving van personen en instellingen moeilijk toe.
- Dublin Core beschrijft vooral het object zelf (formeel) en slechts in beperkte mate het uitgebeelde/beschreven onderwerp (inhoudelijk).
- Verschillende interpretaties van hetzelfde element kunnen leiden tot 'vertaalproblemen'[64], hoewel dit nagenoeg alle metadatastandaarden typeert.

64 Deze semantische misinterpretaties zijn echter onvermijdelijk, voor welke sectoroverschrijdende metadatastandaard men ook zou opteren.

Voorbeeld van een Dublin Core-entry in RDF/XML-syntaxis:

```
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.1/">
    <rdf:Description rdf:about="http://media.example.com/audio/guide.ra">
        <dc:creator>X</dc:creator>
        <dc:title>Interview met een oudstrijder</dc:title>
        <dc:description>Interview met een oudstrijder 1940-1944</dc:description>
        <dc:date>1999</dc:date>
    </rdf:Description>
</rdf:RDF>
```

5.2.2 MPEG-7

MPEG-7 werd ontworpen door de Motion Pictures Expert Group (MPEG).[65] Hoewel deze werkgroep vooral bekend is om zijn coderingsstandaarden voor video en audio, richt MPEG-7 zich op de representatie van informatie over de data in plaats van op de data zelf.[66] MPEG-7 reikt een grote verzameling gestandaardiseerde hulpmiddelen aan voor de beschrijving van multimediale data. Die beschrijvingen moeten mogelijk zijn zonder dat rekening wordt gehouden met de opslagwijze, de codering, de technologie, enzovoort. Zo kan een beschrijving zowel een geprinte foto als een interview in een digitaal audioformaat betreffen.[67]

MPEG-7 bestaat uit *descriptors*, *multimedia description schemes*, *description definition language* en hulpmiddelen die de binaire notatie, de synchronisatie, het transport en de opslag van de *descriptors* voor hun rekening nemen.

Een *descriptor* is de voorstelling van een kenmerk. Die voorstelling ligt zowel syntactisch als semantisch vast. Een object heeft uiteraard meerdere kenmerken, waardoor één object door verschillende *descriptors* beschreven moet worden. *Multimedia description schemes* beschrijven de structuur en de semantiek van de relaties tussen de verschillende *descriptors* en tussen andere *description schemes*.

65 MPEG (2008).

66 Martinez (ed.) (2004).

67 Zie bijvoorbeeld Hunter (2002) over het gebruik van MPEG-7 in combinatie met CIDOC-CRM voor de beschrijving van multimediaal materiaal in musea.

Voor de definiëring van de structurele relaties tussen *descriptors* wordt gebruik gemaakt van een XML-gebaseerde taal: de *description definition language*. Hiermee kan men *description schemes* creëren en aanpassen.

Beschrijvingen in MPEG-7 kunnen op verschillende niveaus plaatsvinden en al dan niet gedetailleerd zijn. Men kan met andere woorden informatie weglaten of verder verfijnen. Die verfijningen verschillen per toepassingsgebied. Voor historische audiobronnen is een mogelijke beschrijving op een hoog niveau: 'Interview met een oudstrijder'. Op een lager niveau wordt eventueel meer detaillistische informatie verschaft. Zo kan de algemene beschrijving uitgebreid worden met de naam van de oudstrijder, informatie over de oorlog, plaats, enzovoort.

Behalve beschrijvingen over de inhoud van een object kan men nog extra informatie toevoegen over de creatie- en productieprocessen van de data, het gebruik van de data (copyright, raadplegingen in het verleden,…), het opslagformaat, de collecties, de interactie tussen gebruiker en data,… Dit zijn dan de administratieve en technische metadata.

Voordelen

- Algemeen erkende en gebruikte MPEG-standaard.

Nadelen

- Wegens de complexiteit is er voorlopig weinig industriële interesse voor de standaard. Er worden echter pogingen ondernomen om de 1182 elementen, 417 attributen en 377 complexe types aan de hand van gereduceerde profielen het hoofd te bieden.

- Ook wegens de te grote flexibiliteit, wat de interoperabiliteit niet altijd ten goede komt, bestaat er voorlopig weinig belangstelling voor MPEG-7. Het is bijvoorbeeld mogelijk om op verschillende abstractieniveaus dezelfde modulaire beschrijvingen te geven, *descriptors* kunnen aan een arbitrair segment toegevoegd worden met om het even welk detailniveau en het huidige schema kan zelfs onbeperkt uitgebreid worden.

- De standaard is nog steeds in ontwikkeling aangezien men wijzigingen blijft voorstellen, vooral aan het *query format*.

Voorbeeld van een MPEG-7-record in XML-syntaxis:

```
<Mpeg7>
    <Description xsi:type="CreationDescriptionType">
        <CreationInformation id="track4">
            <Creation>
                <Title type="songTitle">Interview met een oudstrijder</Title>
                <Abstract>
                    <FreeTextAnnotation>Interview over het leven van een
                     oudstrijder</FreeTextAnnotation>
                </Abstract>
                <Creator>
                    <Agent xsi:type="PersonType">
                        <Name>
                            <FamilyName>De Smedt</FamilyName>
                            <GivenName>Jan</GivenName>
                        </Name>
                    </Agent>
                    <CreationCoordinates>
                        <Date><TimePoint>1999</TimePoint></Date>
                    </CreationCoordinates>
                </Creator>
            </Creation>
        </CreationInformation>
    </Description>
</Mpeg7>
```

5.2.3 P/META

5.2.3.1 Toepassingsgebied en opzet

Bij de uitwisseling van programma-inhoud spelen metadata een cruciale rol. Omdat de behoefte aan interoperabiliteit bij moderne systemen toeneemt, ontstaat ook de nood aan projecten die de standaardisatie van metadata bevorderen. Het EBU P/META-project, dat in 1999 van start ging, neemt op dat gebied een vooraanstaande plaats in. Eigen aan dit metadataproject is bovendien het perspectief en de medewerking van de omroepen zelf, wat een uitgelezen uitgangspunt betekent.

EBU staat voor European Broadcasting Union en verenigt wereldwijd nationale omroepen[68] met het oog op een uniforme omgang met audiovisueel materiaal. Met de ontwikkeling van een metadatastandaard wil EBU een uitwisselingskader creëren in de *business-to-business*-context van elektronische informatie van en over programma's. Dit moet bovendien mogelijk zijn zonder de interne structuur, de werkmethode of het algemene concept van de participerende organisaties te wijzigen. Specifieke opslagschema's van de instellingen kunnen op het P/Meta-schema overgezet worden zodat de uitwisseling van metadata tussen verschillende organisaties mogelijk is, onafhankelijk van de onderliggende technologische infrastructuur voor datatransport.

Voor de volgende toelichting en documentatie van P/Meta is hoofdzakelijk uitgegaan van de rapporten die het EBU in haar *EBU Technical Review* publiceerde. Het laatste rapport dateert van juli 2007 en vat de doelstellingen van de standaard goed samen.[69]

5.2.3.2 Wat is P_META?

Het P/Meta-schema is een verzameling van definities die fungeert als een semantisch kader voor de uitwisseling van informatie met betrekking tot audiovisueel (omroep)materiaal. Het schema bevat in de eerste plaats elementen voor de identificatie van concepten en subjecten die van belang zijn tijdens een eerste analyse van de data. Hiernaar wordt gerefereerd met *identifiers* en *names*. Die identificaties zijn noodzakelijk met het oog op een maximale nauwkeurigheid in de beschrijvingen en een maximale flexibiliteit in het (her)gebruik van de (meta)data en in de definitie van basiselementen en datastructuren.

De uitwisseling van metadata is een proces dat op drie niveaus plaatsvindt: op de *definition layer* (1), de *technology layer* (2) en de *data interchange layer* (3). De *definition layer* (1), die ook als *descriptive metadata layer* of als *semantic layer* aangeduid wordt, bevat de semantiek van de informatie-elementen en wordt door P/Meta ingevuld. Concreet gaat het over de exacte definitie en betekenis van elk beschrijvend element dat van belang is bij een productieproces. Die definities berusten op de professionele betekenis en interpretatie van de concepten en worden daarom uitgedrukt in menselijke taal. De invulling van de *technology layer* (2) heeft betrekking op de gehan-

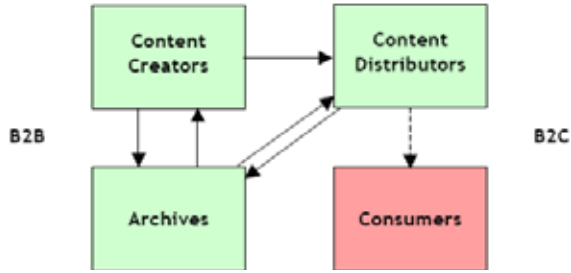
68 Voor België zijn dat RTBF en VRT.

69 EBU (2007). Zie ook de documenten Hopper (2002) en Hopper (2000) voor meer documentatie en achtergrond. De 'metadata library' van 2007, EBU (2007), bevat o.m. ook schema's van de standaard en definities van de P/Meta-elementen.

teerde technologie bij de uitwisseling van informatie. Die technologie mag tijdens de uitwisseling van informatie de originele betekenis van de informatie-elementen niet wijzigen. Mogelijke technologieën zijn XML, KLV of platte tekstdocumenten. De *data interchange layer* (3) geeft aan op welke wijze en via welk medium de gecodeerde informatie werkelijk uitgewisseld wordt.

5.2.3.3 Context

P/Meta is hoofdzakelijk van toepassing in een *business-to-business*-omgeving (B2B). Maar ondanks de focus op B2B wordt de uitwisseling van metadata in een context van *business-to-consumer* (B2C) als een elementaire eigenschap van P/Meta beschouwd. Figuur 12 illustreert het gangbare procesmodel, waarin drie verschillende *business*-actoren en een groep *consumers* geïdentificeerd worden. De pijlen duiden alle mogelijke onderlinge interfaces voor de uitwisseling van informatie aan.



Figuur 12: P/Meta – contexten[70]

Content creators, ook wel *producers* genoemd, verzorgen de productie van programma's en andere media en zorgen ervoor dat nieuw materiaal voor publicatie beschikbaar is. *Archives* staat in voor de zorgvuldige bewaring en bescherming van bestaand materiaal. Het maakt integraal hergebruik van materiaal mogelijk en geldt als eventuele bijkomende bron voor de creatie van nieuwe programma's. *Content distributors* zorgen voor de publicatie en de levering van materiaal aan de eindgebruikers. Zij nemen ook deel aan uitwisselingen in een B2C-context. *Content distributors* worden ook als *broadcasters* of *content aggregators* aangeduid. De *consumers* ten slotte bevinden zich aan het eind van de mediadistributieketen en zijn de uiteindelijke gebruikers.

70 Figuur is ontleend aan Hopper (2002), p. 3.

5.2.3.4 Doelstellingen

Met betrekking tot een B2B-context staan vier doelstellingen voorop: de identificatie en herkenning van materiaal mogelijk maken, een beschrijving van materiaal geven die begrijpelijk en bruikbaar is, de rechten met betrekking tot het materiaal bepalen en minimale technische informatie meegeven zodat de uitwisseling en het correct gebruik van materiaal mogelijk zijn.

5.2.3.5 Het model van P/Meta

De P/Meta-standaard stelt een gelaagd hiërarchisch model voorop dat uit vijf *exchange concepts* bestaat: een *Programme Group*, een *Programme*, een *Item Group*, een *Item* of *Programme Item* en een *Media Object* of *MOB*. De EBU definieert de vijf concepten als volgt:

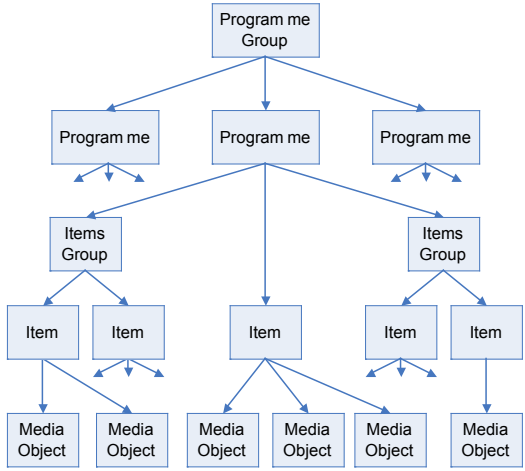
- *Programme Group*: een verzameling van *Programmes* die gecreëerd is na goedkeuring door een commissie of die het resultaat is van een vooropgestelde planning volgens een redactioneel concept. Bijvoorbeeld: het VRT-Nieuws.

- *Programme*: een audiovisueel werk dat gecreëerd en vastgelegd werd na goedkeuring door een commissie. Bijvoorbeeld: Het Journaal.

- *Items Group*: een verzameling van (*Programme*) *Items* als samengesteld onderdeel van een *Programme* of die volgens een redactiebeslissing samen horen. Als een *Item* in dezelfde nieuwsuitzending meer dan één keer aan bod komt, dan worden die *Items* in een *Items Group* gegroepeerd. Men toont bijvoorbeeld eerst een vooraf opgenomen reportage en enkele *Items* verder is er een live interview over hetzelfde onderwerp.

- *Item* of *Programme Item*: een samengesteld onderdeel van een *Programme*, bepaald door een redactionele beslissing. Het kan zowel zelfstandig als door zijn plaats in een *Programme* geïdentificeerd worden. Een *Item* is bijvoorbeeld gemarkeerd volgens het moment waarop het in een nieuwsuitzending verschijnt.

- *Media Object* of *MOB*: één component van één mediatype van een *Programme* of *Item*. Een *MOB* is continu in de tijd en een fysiek object, in tegenstelling tot de vorige logische concepten. Bijvoorbeeld: een bestand met de werkelijke video van een nieuwsuitzending.



Figuur 13: Het model van P/Meta[71]

Andere entiteiten worden in de *definition layer* (cf. §5.2.3.2, 'Wat is P/Meta?') door hun context bepaald. Eén van de belangrijkste entiteiten is een *Brand*. Dit is een collectie activa met een herkenbare collectieve identiteit, bijvoorbeeld *Één* of *Canvas*. Twee andere bepalende entiteiten zijn personen en organisaties die betrokken zijn bij de creatie, het beheer en de controle van de inhoud.

5.2.3.6 Lijst van de componenten van P/Meta

De semantiek van metadata voor de uitwisseling van audiovisueel materiaal wordt voor de P/Meta-standaard bepaald in een lijst van attributen (1), referentiedata (2) en *transaction sets* (3).

Een attribuut (1) is het meest eenvoudige element dat informatie kan bevatten. Het is mogelijk om een attribuut uniek te identificeren aan de hand van een code en een naam, bijvoorbeeld het attribuut met code 'A1' en naam 'ADDRESS_DELIVERY_CODE'. De standaard bepaalt ondubbelzinnig de courante betekenis van elk attribuut, het type van de waarde (bijvoorbeeld *Boolean*, *Integer*, *Uncontrolled Text*, *Controlled Code*,…), een externe referentie, bekende alternatieve benamingen en eventueel enkele voorbeelden. De betekenis van een attribuut kan toenemen door het gebruik ervan te contextualiseren. Zo kan hetzelfde attribuut in verschillende contexten (her)gebruikt

71 Figuur is ontleend aan De Sutter, Notebaert en Van de Walle (2006), p. 224.

worden om bijvoorbeeld uiteenlopende B2B-doelen te dienen. Het attribuut *Language Code* kan bijvoorbeeld zowel worden ingezet om de taal van de originele dialoog van een *Item* aan te geven als om de taal van de eigendomsrechten aan te duiden.

Van alle attributen waarvan het waardetype een *Controlled Code* betreft, moet een lijst voorhanden zijn met alle toegelaten waarden en hun precieze betekenis (2). Dergelijke waardenlijsten worden door P/Meta aangeboden of door een externe bron als EBU, ISO of SMPTE.

Behalve attributen en referentiedata biedt P/Meta ook *transaction sets* (3) of functionele groeperingen van P/Meta-attributen en/of andere P/Meta-sets aan. Een voorbeeld is de set 'S12 PERSON_DETAILS', die bestaat uit enkele attributen, zoals 'A89 PERSON_LAST_NAME' en 'A88 PERSON_FIRST_NAME', en aangevuld wordt met de subset 'S13 ADDRESS'. De P/Meta-standaard bevat een aantal vooraf gedefinieerde *sets* die geconstrueerd zijn als bouwstenen voor de opzet van gemeenschappelijke data-uitwisselingen. Ook kunnen nieuwe logische sets gecreëerd worden om aan de eigen specifieke transactievereisten te voldoen. Deze sets worden geconstrueerd volgens een syntaxis en notatiewijze die door de standaard aangegeven zijn. De voorgedefinieerde *sets* zijn ontwikkeld om in de volgende domeinen alvast een generieke oplossing te bieden: metadata voor identificatie en herkenning, beschrijvende metadata, technische metadata, metadata met betrekking tot transacties, transmissies, rechten en andere.

Het P/Meta-schema laat toe om de logische inhoud en de betekenis van informatie te beschrijven, onafhankelijk van de technische implementatie. Men kan aan de eisen van het P/Meta-schema voldoen zonder rekening te houden met een bepaald platform, een coderingsstandaard, een transactieprotocol of zelfs de gekozen taal in het geval van *Controlled Code*-attributen.

5.2.3.7 IPEA: Innovatief Platform voor Elektronische Archivering

Twee grote Vlaamse omroepen, de commerciële omroep VMMa en de publieke omroep VRT, het technisch ondersteunend bedrijf Videohouse en verschillende universitaire onderzoeksgroepen die aan het IBBT verbonden zijn, bundelden hun krachten in het IPEA-project (2005-2006).[72] IPEA onderzocht de eisen met betrekking tot de overstap van een tapegebaseerde

72 Zie de projectwebsite IBBT (2007a) en de niet gepubliceerde deliverable 'Digitale archivering op nationaal en internationaal vlak: een stand van zaken', Hauttekeete, Dekeyser, et al. (2006).

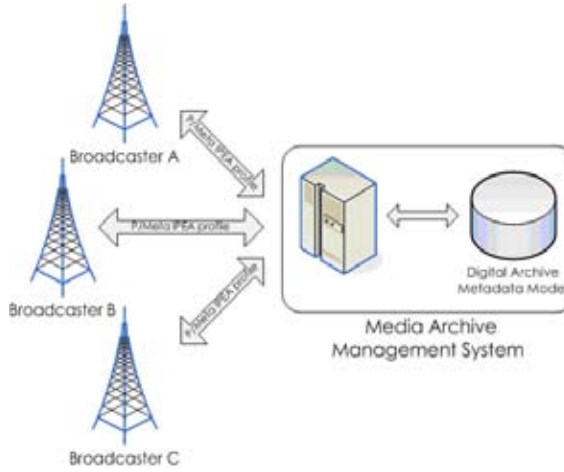
productie– en archiveringsomgeving naar een volledig bestandsgebaseerde workflow. Het project stelde onder meer de ontwikkeling van een algemene gemeenschappelijke standaard voor de uitwisseling en archivering van audiovisuele data voorop.

Eén van de cruciale aspecten van het IPEA-project betrof dan ook de creatie van een gemeenschappelijk metadatamodel,[73] waarin een gestandaardiseerde semantische beschrijving, een gestandaardiseerde syntaxis en een definitie van nuttige ontologieën centraal stonden.

Dit leidde tot de definitie van twee metadatastandaarden: een intern model en een uitwisselingsmodel. De eerste standaard is een ERD[74] waarin de layout en de relaties van het digitale archief gedefinieerd worden. Het tweede model is bedoeld voor externe gebruikers met het oog op de invoer en afhaling van media in of uit een digitaal archief. In dit model wordt gebruik gemaakt van het zogenaamde IPEA-profiel, een uitwisselingsstandaard die een subset van de EBU P/Meta 1.1-standaard is. Deze subset beantwoordt aan de noden van de genoemde Vlaamse omroepen. De keuze voor de internationaal genormeerde specificatie P/Meta vereenvoudigde de definitie van de semantiek, de syntaxis en de taal waarmee tussen verschillende actoren over programma's gecommuniceerd kan worden.

73 Dit werd gerealiseerd in werkpakket 4 van IPEA: WP4, 'Creation of a standardized metadata model', cf. IBBT (2007a).

74 In een ERD of Entity Relationship Diagram wordt een conceptueel model grafisch voorgesteld.



Figuur 14: IPEA uitwisselingsmodel[75]

P/Meta legt dus de semantiek en de syntaxis van alle elementen voor de interface strikt vast en kon voor het IPEA-profiel nauwkeurig gevolgd worden. P/Meta biedt een eenduidige definitie van de semantiek in een exhaustieve lijst van alle attributen die belangrijk kunnen zijn voor de beschrijving van een programma en de specificatie van de mogelijke waarden van de attributen. P/Meta bepaalt ook de syntaxis van de interface door aan te geven op welke manier de attributen zinvol met elkaar kunnen communiceren. De syntaxis is echter opgesteld met het oog op interpreteerbaarheid door computers en is bijgevolg vrij abstract.

In tegenstelling tot de opzet van P/Meta, waarin van mogelijke implementaties en noodzakelijke technologieën geabstraheerd wordt, was voor het IPEA-project een consensus wel noodzakelijk. Vervolgens werd geopteerd voor een implementatie van het IPEA-profiel in XML en een controle van de gecreëerde documenten door XML-schema's.

In november 2005 lanceerde EBU versie 1.2 van P/Meta waarin, mede onder impuls van de resultaten van het IPEA-project, enkele *attributes* en *sets* aan versie 1.1 werden toegevoegd.[76]

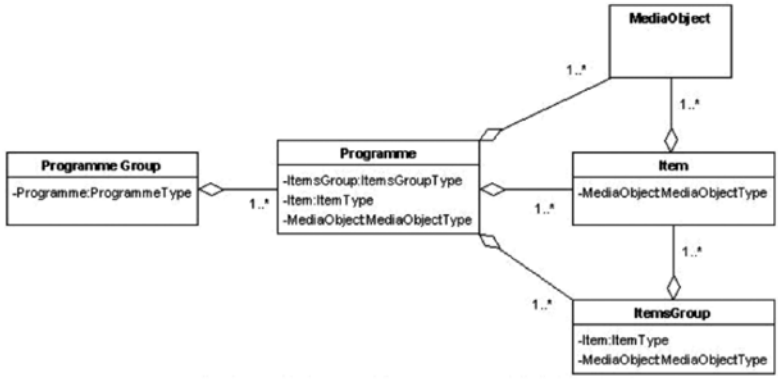
75 Figuur is ontleend aan De Sutter, Notebaert, et al. (2006), p. 185.

76 EBU (2005).

5.2.3.8 P_META 2.0

Sinds juli 2007 is versie 1.2 vervangen door P/Meta 2.0.[77] Deze versie is het resultaat van een aantal opmerkelijke wijzigingen. De meest opvallende wijzigingen zijn de toewijzing van namen aan structuren die voordien naamloos waren, de optimalisering van de datastructuur voor het gebruik van XML en de opsplitsing van de originele P/Meta-metadataset in een algemene toolkit en afzonderlijke toepassingsgerichte specificaties, bijvoorbeeld voor de uitwisseling van programma's. Die aanpassingen zorgen er echter voor dat versie 2.0 moeilijk compatibel is met de eerdere versies.

In P/Meta 2.0 wordt een *Brand* niet meer als een *exchange concept* gezien. Men beschouwt het nu als een element dat door zijn context bepaald wordt en nog steeds gedefinieerd is als een collectie van activa met een herkenbare collectieve identiteit. De entiteit *Items Group* wordt in 2.0 meer op de voorgrond geplaatst waardoor de logische structuur van het P/Meta-model er nu uitziet als in figuur 15.



Figuur 15: P/Meta 2.0[78]

Aangezien XML bij de creatie van P/Meta 2.0 een prominente rol speelde, introduceert de nieuwe standaardversie een aantal conventies die de overstap van P/Meta 1.2 naar XML mogelijk maken. Zo zijn alle P/Meta-attributen XML-elementen of -attributen geworden en is de notatiewijze aangepast: i.p.v. de *set* 'S12 PERSON_DETAILS', wordt in de huidige standaard het element 'PersonDetails' van het type 'pmeta:PersonDetailsType' genoteerd. Het element 'PersonDetails' bevat onder andere de elementen 'PersonLastName'

77 EBU (2007).

78 Figuur is ontleend aan EBU (2007).

en ‘PersonFirstName’ van het type ‘string’ (t.e.m. versie 1.2 gedefinieerd als de attributen ‘A89 PERSON_LAST_NAME’ en ‘A88 PERSON_FIRST_NAME’) en het element ‘Address’ van het type ‘pmeta:AddressType’ (vroeger de set ‘S12 PERSON_DETAILS’).

5.2.4 SMEF-DM

Het Standard Media Exchange Framework (SMEF) Data Model bepaalt een set van datadefinities voor alle informatie met betrekking tot productie, ontwikkeling, gebruik en beheer van mediaobjecten. Het model werd ontwikkeld met de bedoeling de samenwerking en uitwisseling van informatie tussen systemen te garanderen door middel van een gemeenschappelijk kader voor de gedeelde data.[79]

SMEF is ontwikkeld door de Media Data Group (BBC Technology) in opdracht van de BBC. Het model is afgewogen tegen andere relevante standaarden zoals MPEG-7, P/META, ISAN/V-ISAN en TV Anytime, die vervolgens hebben bijgedragen tot de definitie van het SMEF-datamodel.

Het SMEF-datamodel is gebaseerd op metadata die op media betrekking hebben. Het bereik van de metadata gaat verder dan een beschrijvend karakter. Het model bevat namelijk ook de nodige gegevens om *business*-processen te ondersteunen, wat kan gaan van commissie en het vastleggen van video tot transmissie en archivering. SMEF-DM is bruikbaar in de sectoren van televisie, radio en het web (media) en biedt ondersteuning aan zowel analoge als digitale services.

Het SMEF-datamodel geldt als de centrale bron van datadefinities voor de ontwikkeling van applicaties voor de BBC. Het biedt een initiële set van datadefinities aan voor BBC-projecten en kan tijdens projecten voortdurend als toetssteen fungeren.

Hoewel het model aanvankelijk ontwikkeld is voor specifiek gebruik binnen de BBC, zijn de datadefinities voldoende generisch om ook buiten de BBC-context toepasbaar te zijn. Sommige referenties in SMEF-DM zijn echter heel specifiek voor de BBC, zoals het World Service Programme Numbers. De BBC heeft bijvoorbeeld ook een specifieke interpretatie van het begrip ‘programme week’ (dat vervat zit in de entiteit PRGRAMME_WEEK_CALENDAR_YEAR). De gebruikers buiten de BBC-omgeving moeten hiervan op de hoogte zijn, ook al is het model toepasbaar in de algemene omroepwereld.

79 BBC (2007).

Zoals eerder gesteld, bestaan er relaties tussen SMEF, Dublin Core en P/Meta, die hier bondig toegelicht worden. Dublin Core (§5.2.1) is een metadatastandaard die doorgaans naast andere metadatastandaarden, waaronder SMEF, gebruikt wordt. Hoewel dit niet specifiek in de documentatie van SMEF-DM opgenomen is, is het vrij eenvoudig om de SMEF-attributen om te zetten naar een equivalente Dublin Core-*descriptor*.

P/Meta en SMEF zijn uiterst compatibel met elkaar. P/Meta richt zich vooral op de *business-to-business*-uitwisseling van programma-informatie en –data, terwijl SMEF veeleer voor de interne informatiesystemen gebruikt wordt. Niettemin is het mogelijk om directe relaties te leggen tussen P/Meta- en SMEF-attributen. Die mogelijkheden zijn ruim gedocumenteerd.[80]

De enorme hoeveelheid entiteiten en relaties binnen het SMEF-datamodel komt de leesbaarheid van een voorstelling in één diagram echter niet ten goede. Daarom wordt het model in acht diagrammen onderverdeeld. Deze diagrammen zijn complementair en bieden als het ware een vereenvoudigd perspectief op het hele datamodel. Een individuele entiteit kan dan in meerdere diagrammen voorkomen. De acht diagrammen hebben betrekking op de volgende concepten: *editorial object*, *media object*, materiaalinstantie, subject en referentie, commissie, editoriaal genre en beschrijving, contract en rol, publicatie en publiek.

SMEF is een groot en complex model. Om te vermijden dat de terminologie in verschillende contexten steeds anders geïnterpreteerd zou worden, wordt ze vrij concreet geformuleerd. Het volgende overzicht verduidelijkt de gehanteerde terminologie.

Een *editorial object* in SMEF betreft een volledig programma of item. Andere namen voor een editoriaal object zijn bijvoorbeeld ‘werk’ of ‘episode’. De term ‘editoriaal object’ kan in verschillende entiteiten voorkomen:

- EDITORIAL_OBJECT_GROUP: representeert iedere groep programma’s of items met het oog op promotie of verkoop.
- EDITORIAL_OBJECT_CONCEPT: beschrijft de eigenschappen van één programma of werk dat op alle versies van dat programma van toepassing is.
- EDITORIAL_OBJECT_VERSION: beschrijft een versie van een editoriaal object voor een specifiek doel.

80 Zie o.m. BBC (2007) en Mauthe en Thomas (2004).

Het *Image Format Type* definieert de geometrische eigenschappen van een beeld of beeldapparaat. De BBC stelde een *Publication Format Code* voor om deze informatie te presenteren. Deze code bestaat uit zes karakters en volgt het formaat aabccd; waarbij:

- aa= Active Image Aspect Ratio
- b = Display Format
- cc= Raster Aspect Ratio
- d = Protected Aspect Ratio

De codes aa en cc kunnen de volgende waarden hebben:

- 16 = 16:9
- 15 = 15:9
- 14 = 14:9
- 12 = 12:9 = 4:3

De code b kan de volgende waarden hebben:

- P = Pillarbox
- L = Letterbox
- F = Full Frame
- M = Mixed Formats

Hoewel het *Display Format* uit de drie andere parameters afgeleid kan worden, wordt het toch expliciet in de code opgenomen.

Het ACQUISITION_BLOCK is een verzameling audio *items* in gepubliceerde vorm, bijvoorbeeld een CD of record. Meerdere *acquisition blocks* vormen een *catalog* van opgenomen muziek- en spraakitems. Het gaat hier zeker niet altijd over CD's maar het geldt veeleer als een aanduiding van de beschikbaarheid van die set records. Een voorbeeld hiervan is 'The Best of Des O`Connor'. MUSIC_SPEECH_SOUND_ITEM_IN_BLOCK is een individueel *item* in een

acquisition block. Dit kan bijvoorbeeld een track zijn van een CD, bijvoorbeeld de track 'Moon River' van de collectie 'The Best of Des O'Connor'. De beschrijvende informatie over dit item is opgenomen in één van de subelementen van EDITORIAL_OBJECT_VERSION.

Een *media object* is de beschrijving van een component van een *editorial object*, bijvoorbeeld de audio, video of ondertiteling. De entiteit MEDIA_OBJECT bevat dan de metadata met de algemene en editoriale informatie over een *media object*, bijvoorbeeld de editoriale beschrijving van een audioclip. UNIQUE_MATERIAL_INSTANCE bevat de attributen die de opslag van een *media object* beschrijven. Eén van de attributen van een unieke materiaalinstantie is UMID (Unique Material Identifier) die de SMPTE-standaard voor de identificatie van materiaal volgt.[81] MEDIA_OBJECT_GROUP bepaalt de editoriale en conceptuele verbanden tussen *media objects*. Een voorbeeld hiervan is een set van specifieke audioclips die samen gegroepeerd zijn als een fase in het productieproces. De entiteiten STORAGE en STORAGE_TYPE duiden aan op welke locatie en in welke vorm het unieke materiaal wordt bewaard.

De groep *Location*, *Story* en *Classification* betreft een samenvatting van de verschillende manieren waarop individuele programma's en items worden gecatalogeerd en geclassificeerd. De locatieset CONTENT_LOCATION duidt de locaties aan die betrekking hebben op een *media asset*. STORY omschrijft het thema dat van toepassing is op de verschillende versies van programma's of items.

Een andere groep entiteiten beschrijft gegevens met betrekking tot transmissie en publicatie van editoriale objecten. De kernentiteit hiervan is PUBLICATION_EVENT die de geplande en actuele transmissie en/of publicatie van een programma representeert. Een *publication event* kan onderverdeeld worden in andere *publication events*. Dat is handig voor *publication events* die als het ware een container vormen en weer onderverdeeld kunnen worden in andere *publication events*.

Personen en organisaties worden door een andere groep entiteiten beschreven. Zij kunnen verschillende rollen in het *media asset management* uitoefenen, bijvoorbeeld een cameraman, een director of de eigenaar van rechten op een bepaalde locatie. SMEF ondersteunt de verschillende rollen die een persoon of organisatie in deze context kunnen vervullen. De entiteit PERSON vertegenwoordigt dus een belangrijke actor in de sector. De entiteit ORGANISATION behandelt groepen zoals bijvoorbeeld een erkend, onafhan-

81 Zie o.m. SMPTE (2003).

kelijk productiehuis. De entiteit PERSON_LINK_ORGANISATION ondersteunt mogelijke onderlinge relaties van personen en organisaties. De entiteit ROLE beschrijft dan de taak of de verantwoordelijkheid van de persoon en/of organisatie. De associatie kan een contract inhouden en wordt voorgesteld door de entiteit CONTRACT en CONTRACT_LINE.

5.2.5 MARC/MARC21

MARC is een acroniem van Machine-Readable Cataloguing.[82] MARC is een standaard voor de representatie en de communicatie van bibliografische en aanverwante informatie en dit in een vorm die de computers kunnen lezen. De standaard wordt onderhouden door de Amerikaanse Library of Congress en vindt zijn oorsprong in de jaren zestig als een digitale vorm van bibliotheekfiches. De hoofdfunctie van de standaard was dan ook de vereenvoudiging en bespoediging van het terugvinden van boeken in een bibliotheek. De data-elementen van MARC zijn bijgevolg de basis van de meeste bibliotheekcatalogi. Er bestaan immers geen alternatieve standaarden of modellen met een gelijkaardige verfijningsgraad.

MARC ondersteunt acht soorten materiaal. Onder het type *sound recordings* ressorteren alle soorten geluid, behalve muziek. Hieronder vallen dus onder meer mondelinge historische bronnen. Verder bevat MARC ook zeven types records. Zo bevat het type *computer file* bijvoorbeeld een gedigitaliseerde versie van een mondelinge historische bron en *manuscript (textual) language material* betreft bijvoorbeeld de transcriptie van een mondelinge historische bron.

Een (bibliografisch) MARC-record bestaat uit meerdere velden: auteursvelden, velden met titelinformatie, enzovoort. Deze velden worden verder onderverdeeld in subvelden.

82 Zie LOC (2008e) voor de website van MARC. Andere nuttige referenties met MARC-documenten, crosswalks, e.a. zijn: LOC (2008c) (de definities van de verschillende velden en subvelden), (LOC, 2008g) (documentatie en toolkits m.b.t. implementatie in XML), LOC (2004) (over het MARC-veld met preservatie-informatie), McCallum (2002) en Spicher (1996) (algemene referentieartikels). Bestaande crosswalks of mappingen met andere standaarden: Baca, Clarke, et al. (2009) ('pathways' met verschillende standaarden), Clarke (2001) (VRA Core naar MARC), LOC (1999) (o.a. USMARC en EAD), LOC (2008a, 2008f) (MARC en Dublin Core), LOC (2008d) (MARC en MODS).

De tekstuele namen van de velden, zoals auteur en onderwerp, worden vervangen door tags die bestaan uit een driecijferige code. Die code beschrijft dan welke gegevens in het veld staan.

Subvelden worden gescheiden door middel van een karakter (bv. $), aangevuld met een subveldcode die aangeeft welke gegevens volgen.

Sommige velden worden verder gedefinieerd door indicatoren. Dit zijn twee posities die een karakter tussen 0 en 9 kunnen bevatten. Het tweede karakter kan bijvoorbeeld aangeven dat een aantal volgende karakters bij de sortering door de computer genegeerd moet worden. Dit is bijvoorbeeld van toepassing bij familienamen die met 'van' beginnen.

Een eenvoudig voorbeeld van een MARC-entry:

245 10 $aInterview met een oudstrijder$h[sound recording].
260 ## $aKortrijk$bVereniging voor oudstrijders$c1999.
300 ## $a1 minidisc$bdigital, ATRAC, stereo.
500 ## $aInterview met een oudstrijder 1940-1944.
500 ## $atranscriptie beschikbaar.
511 0# $aInterview afgenomen door X

De eerste regel bevat een veld met de code 245, een *title statement*. De indicatoren hebben de waarde 1 en 0 en het veld bevat de subvelden $a (de eigenlijke titel) en $h (het medium).

Om de records overzichtelijker te maken en bewerking van de records te vereenvoudigen, is later de MARC-XML standaard ontworpen die de records in een XML-bestand voorstelt.

Het voorbeeld toont de hoge mate van verfijning van het MARC-formaat, maar ook de daarmee gepaard gaande complexiteit, en toch is het formaat compact. Veldnamen zoals 'plaats van publicatie' worden immers door een korte code vervangen.

Bovendien is de veldcodering logisch opgebouwd, wat de complexiteit eveneens doet afnemen. Zo betekent 6XX een veld met informatie over het onderwerp en duidt X00 op een naam.

MARC kent geen semantische zoekfunctie. Er wordt namelijk enkel gezocht naar de gegeven sleutelwoorden in de verschillende velden zonder dat rekening gehouden wordt met de betekenis of het concept van de sleutelwoorden.

Behalve bibliografische records, met een bespreking van kenmerken van resources, beschrijven andere types records bijvoorbeeld een classificatie of ze geven informatie over namen of onderwerpen.

Voordelen

- Hoge mate van verfijning
- Wijdverbreid en courant
- Mogelijke weergave in XML-syntaxis

Nadelen

- Complex
- Geen hiërarchische opbouw
- Geen semantiek
- Moeilijk leesbaar voor 'leken' (wegens de numerieke codes)

5.2.6 MODS

MODS, het Metadata Object Description Scheme, werd, zoals MARC21, ontwikkeld door de Library of Congress, maar is veel jonger.[83] De standaard wordt onderhouden door de Network Development and MARC Standards Office en houdt rekening met de input van gebruikers. In 2002 werd een eerste versie gepubliceerd en momenteel is men toe aan versie 3.3. MODS kan beschouwd worden als een XML-afgeleide van MARC21 en leent zich meer voor de beschrijving van objecten in een digitale omgeving. MODS kent in die optiek dan ook heel wat voordelen ten opzichte van zijn zogenaamde antecedent, waarvan het ontstaan zich veeleer in een traditionele papieren bibliotheekomgeving situeerde.

Als XML-schema moest de MODS-standaard data kunnen bevatten van bestaande MARC21-bestanden. MODS bevat een subset van MARC-velden en maakt gebruik van taalgebaseerde tags in plaats van numerieke tags zoals in

83 Zie LOC (2008h) voor de homepage van MODS en LOC (2008i) voor een toelichting bij de elementen en attributen van MODS versie 3.3. Zie ook de referenties Guenther (2003) en McCallum (2004).

MARC. In sommige gevallen hergroepeert de standaard elementen van de MARC21-standaard.

MODS kan in de volgende toepassingsgebieden gebruikt worden:

- Als een SRU-formaat[84]
- Als een uitbreidingsschema van METS (cf. §6.2)
- Voor de weergave van metadata bij het harvesten
- Voor de beschrijving van bronnen in XML
- Voor de weergave van een vereenvoudigd MARC-record in XML

MODS vult andere metadataformaten aan. Voor sommige toepassingen, zeker voor MARC-toepassingen, kent MODS behoorlijk wat voordelen ten opzichte van andere schema's. Zo is de elementenset rijker dan die van Dublin Core en eenvoudiger dan het volledige MARC-formaat.[85] In vergelijking met het ONIX-schema zijn MODS-elementen dan weer beter compatibel met bibliotheekdata.[86] Het MODS-schema is bovendien gebruiksvriendelijker dan bijvoorbeeld MARC door het gebruik van linguïstische tags in plaats van numerieke tags.

Naast deze voordelen zijn er nog enkele interessante aspecten van MODS te noemen:

- De elementen erven de semantiek van MARC.
- Sommige data worden in MODS anders gegroepeerd dan in het meer uitgebreide MARC: wat bij MARC aan de hand van meerdere data-elementen wordt beschreven, kan in MODS bijvoorbeeld met één data-element beschreven worden.

84 *Search Retrieval via URL*, cf. LOC (2009c).

85 Voor crosswalks en mappings tussen MODS en andere standaarden, zie o.m. Baca, Clarke, et al. (2009) (crosswalks tussen verschillende descriptieve metadatastandaarden) en LOC (2008d) (mapping tussen MARC en MODS). Voor een afweging tussen Dublin Core en MODS, zie o.a. LOC (2003) (in het kader van het project American Memory van de LOC).

86 ONIX (Online Information eXchange) is een applicatie voor de elektronische uitwisseling van boekmetadata. Het is een *open source* software dat een schema aanbiedt voor de online beschrijving van boeken. Cf. EDItEUR (2009).

- Terwijl MARC het gebruik van een cataloguscode eist, is dat bij MODS niet het geval.
- Verschillende elementen hebben optioneel een ID-attribuut waardoor men gemakkelijk links op elementniveau kan leggen.

Zoals eerder aangegeven heeft MODS een subset elementen van MARC21 overgenomen. Aangezien MODS de representatie van MARC-data toelaat, beoogt de standaard een conversie van de core-elementen waarbij de meer specifieke data dus genegeerd kunnen worden. Het MODS-schema streeft echter geen *round-tripability* met MARC21 na. Dit houdt in dat een MARC21-record naar MODS geconverteerd kan worden maar niet terug naar het originele MARC21-record.

MODS wordt in XML geserialiseerd. Hiertoe definieert MODS hoofdelementen, subelementen en attributen van de hoofd- en subelementen. De inhoud van de elementen wordt enkel op het laagste niveau ingevuld om 'mixed elements' te vermijden. Bijvoorbeeld als <titleInfo> enkele subelementen voor <title>, <partNumber> en <partName> bevat, dan fungeert <titleInfo> enkel als *wrapper*-tag waaronder de meer specifieke elementen <title>, <partNumber> en <partName> vallen.

Attributen kunnen op elk niveau met elementen in verband gebracht worden en worden ook met dat betreffende element gedefinieerd. Enkele gebruikelijke attributen zijn: *type*, *encoding* en *authority*.

Een MODS-document begint met een schemadeclaratie die de *namespace* aangeeft. In een (groep) record(s) is deze schemadeclaratie voor elk element optioneel, aangezien de MODS-*namespace* in het record zelf wordt aangegeven. Toch is het heel gebruikelijk om het prefix 'mods:' voor elk element te gebruiken wanneer een MODS-record gecombineerd wordt met XML-data van een andere *namespace*, bijvoorbeeld een MODS-record in een METS-document.

In MODS is geen enkel element verplicht. De enige voorwaarde luidt dat iedere MODS-beschrijving ten minste één element bevat.

Voorbeeld van een MODS-record, een hoofdstuk uit een boek:

```
<mods xmlns:xlink="http://www.w3.org/1999/xlink" version="3.0" xmlns:
xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://www.loc.
gov/mods/v3" xsi:schemaLocation="http://www.loc.gov/mods/v3 http://www.loc.
gov/standards/mods/v3/mods-3-0.xsd">
<titleInfo>
        <title>Models, Fantasies and Phantoms of Transition</title>
</titleInfo>
 <name type="personal">
        <namePart type="given">Ash</namePart>
        <namePart type="family">Amin</namePart>
        <role>
                <roleTerm type="text">author</roleTerm>
        </role>
</name>
 <typeOfResource>text</typeOfResource>
 <relatedItem type="host">
        <titleInfo>
                <title>Post-Fordism</title>
                <subTitle>A Reader</subTitle>
        </titleInfo>
        <name type="personal">
                <namePart type="given">Ash</namePart>
                <namePart type="family">Amin</namePart>
                <role>
                        <roleTerm type="text">editor</roleTerm>
                </role>
        </name>
        <originInfo>
                <dateIssued>1994</dateIssued>
                <publisher>Blackwell Publishers</publisher>
                <place>
                        <placeTerm type="text">Oxford</placeTerm>
                </place>
        </originInfo>
        <part>
                <extent unit="page">
                        <start>23</start>
                        <end>45</end>
                </extent>
        </part>
```

</relatedItem>
<identifier>Amin1994a</identifier>
</mods>

5.2.7 CDWA

CDWA (Categories for the Description of Works of Art) werd in de jaren negentig ontworpen door de Art Information Task Force, kortweg AITF, een discussiegroep van kunstwetenschappers, museumcurators, kunstbibliothecarissen, informatiemanagers, technische specialisten en andere medespelers in de kunsten- en museumsector die de noodzaak van richtlijnen voor de beschrijving van artistieke producten erkenden.

De standaard wordt gehanteerd voor de beschrijving van data uit kunstdatabanken.[87] Het gaat om een conceptueel kader voor het beschrijven en opvragen van informatie over kunstwerken, architectuur, e.a. CDWA omvat 512 categorieën waarvan een kleine kern toelaat de minimale gegevens te verschaffen om het kunstwerk te identificeren en te achterhalen. Daarenboven omvat de CDWA ook discussiefora, basisregels voor het catalogiseren en voorbeelden.

De categorieën leveren een kader waar bestaande informatiesystemen naar gemapt kunnen worden en op basis waarvan men nieuwe systemen kan ontwikkelen. Discussies in het kader van CDWA stellen woordenschatten voor en suggereren beschrijvende toepassingen die de informatie in de verschillende systemen meer compatibel en toegankelijk maken.

Het gebruik van het CDWA-kader moet bijdragen tot de integriteit en de levensduur van de data en moet de datamigratie naar nieuwe systemen in de toekomst vereenvoudigen. Eén van zijn voornaamste kwaliteiten is dat het de eindgebruiker zal helpen in zijn zoektocht naar betrouwbare informatie, zonder dat rekening wordt gehouden met het systeem waarin de data opgeslagen zijn.

Verder is CDWA hiërarchisch en relationeel opgevat, zodat gegevens over curator, plaats, maker,… eenmaal beschreven en bijgehouden worden in afzonderlijke bestanden en gekoppeld kunnen worden aan gegevens over de werken waarop die metadata van toepassing zijn.

87 Zie Getty (2006) voor een overzicht van de elementen van CDWA.

Het CDWA Lite is een XML-schema dat de kernelementen (*core*) bevat voor de beschrijving van een kunstwerk of cultureel object en dat gebaseerd is op het CDWA en het CCO. CDWA Lite wil bijdragen aan de vereniging van catalogi en bibliotheken die gebruik maken van het OAI-PMH-protocol[88] voor de uitwisseling van data.

Het CCO (Cataloguing Cultural Objects: A Guide to Describing Cultural Works and their images) levert een aantal voorschriften en richtlijnen aan voor de selectie, ordening en formattering van data die gebruikt worden om catalogusrecords aan te vullen.[89] Het gebruikt hiervoor informatie die gerelateerd is aan een subset van de CDWA-categorieën en de VRA Core-categorieën (cf. §5.2.8).[90]

5.2.8 VRA Core

VRA Core (Visual Resources Association Data Standards Committee) laat toe visueel cultureel erfgoed te beschrijven. De standaard schrijft zowel de metadata-elementen als hun mogelijke onderlinge structuur voor. Behalve een beschrijving van de werken categoriseert de standaard ook digitale afbeeldingen van het kunstwerk in kwestie.

In de meest recente versie, 4.0, is de implementatie van elementen in XML-syntaxis verwezenlijkt.[91] Daarbij werden in eerste instantie de zogenaamde *element qualifiers* uit versie 3.0 geherdefinieerd en geconverteerd naar subelementen en attributen volgens de XML-syntaxis. Vervolgens werd ook de onderverdeling in types records gewijzigd. In versie 4.0 onderscheidt men nu drie mogelijke types records, beschreven door de XML-elementen *collection*, *work* en *image*,[92] die als zogenaamde *wrappers* van het betreffende

88 Voluit The Open Archives Initiative Protocol for Metadata Harvesting. Cf. Lagoze en Van de Sompel (2008).

89 Zie Baca (2007) voor een bespreking van CCO en CDWA Lite.

90 Zie Baca, Clarke, et al. (2009) voor crosswalks tussen CDWA Lite en andere metadatastandaarden.

91 Zie VRA (2007a) voor de website van de huidige versie 4.0, VRA (2007c) voor een lijst van de elementen en VRA (2007b) voor hun definitie en beschrijving. Zie ook nog MITLibraries (2007) (over VRA Core op de site van MITLibraries, met nuttige links en referenties) en Kessler (2007) ('the story behind VRA'). Documentatie over mogelijke crosswalks is te vinden in Baca, Clarke, et al. (2009) (verschillende metadatastandaarden) en Clarke (2001) (VRA Core 3.0 met MARC21).

92 In versie 3.0 was enkel sprake van 'work' en 'image'.

record fungeren. Dit illustreert meteen ook de hiërarchische opbouw van de standaard.

Een object van het type *work* is een unieke entiteit, zoals een *object* of een *event*. Voorbeelden hiervan zijn schilderijen, beeldhouwwerken of voorstellingen. Een record van het type *image* is een visuele representatie van een *work* in zijn geheel of een deel ervan. Een *collection* is een aggregatie van *work* of *image*-records. Deze groepering was nodig om het catalogiseren op collectieniveau toe te laten.

Het enige noodzakelijke element in een VRA Core 4.0-record bevat de informatie die een record eenduidig kan identificeren. In niet-XML-formaat betekent dit dat er minstens een *work*-, *collection*- of *image*-element in het record aanwezig moet zijn. In XML-formaat gaat het om de *wrapper* die deze informatie bevat.

Het is uiteraard aangewezen om in het record, behalve de identificatie, ook beschrijvende elementen toe te voegen met het oog op een vlotte en eenvoudige *retrieval* van het record. Hier gaat het voornamelijk om de elementen die informatie geven over de fundamentele vragen over het object: wat, wie, waar en wanneer. De volgende aangewezen elementen geven mogelijk antwoord op deze vragen en staan in voor een accurate beschrijving van het object:

WORK TYPE (wat)

TITLE (wat)

AGENT (wie)

LOCATION (waar)

DATE (wanneer)

VRA Core 4.0 baseert zich op richtlijnen van het CCO (cf. §5.2.7). Die richtlijnen stellen dat men rekening moet houden met *display* en *indexing requirements*. Dat wil zeggen dat datawaarden van een specifiek metadataelement geformatteerd moeten zijn in de vorm waarin ze getoond worden. Tegelijkertijd moeten deze datawaarden ook afzonderlijk geformatteerd worden en moeten ze gekoppeld zijn aan thesauri om de *retrieval* van de data te vereenvoudigen.

Daarom heeft het VRA Core XML-schema elk element van twee subelementen voorzien: *display* en *notes*. Die subelementen worden binnen de *wrapper* van het element 'genest':

```
<materialSet>
    <display>oil on canvas</display>
<notes source="Art Bulletin, v.87, no.1 (March 2005)">Medium originally thought to be
tempera. Oil medium discovered in tests at Uffizi in 2003</notes>
    <material type="medium" vocab="AAT" refid="300015050">oil paint</material>
    <material type="suppert" vocab="AAT" refid="300014078">canvas</material>
</materialSet>
```

In het facultatieve subelement *notes* kan men vrije tekst of annotaties toevoegen die nog niet door andere attributen beschreven werden. Indien de annotatie op het hele record betrekking heeft, dan moet men het element *description* gebruiken.

Het element *record type* uit versie VRA Core 3.0 werd in 4.0 vervangen door een element *work, collection* of *image*. Dit element wordt als het ware gebruikt als opslagplaats voor de administratieve metadata. De elementen worden met twee attributen aangevuld:

- *Id*: een unieke identificatiecode van het XML-record.
- Het globale *refid*-attribuut: een lokaal nummer, code of adres die het record op een unieke manier identificeren binnen de context van het *source*-attribuut.
- Het globale *source*-attribuut: de set of de omgeving waartoe het record behoort, bijvoorbeeld de naam van een museum.

Samengevat functioneren *work*, *image* of *collection* in het XML-schema dus als overkoepelende *wrappers* waarin de andere elementen, die in sets gegroepeerd zijn, zich bevinden. De attributen *id*, *refid* en *source* van een *work*, *image* of *collection-wrapper* zorgen voor een unieke identificatie in verschillende contexten. Het *id*-attribuut identificeert een XML-record binnen een bestand met meerdere XML-records. *Refid* en *source* identificeren een XML-record binnen het systeem vanwaar het afkomstig is.

Voorbeeld van een VRA Core 4.0-record:

```
<work id="w_98765432" refid="14363" source="History of Art Visual Resources Collection, UCB">
    <agentSet>
        <display></display>
        <notes></notes>
        <agent></agent>
    </agentSet>
    <dateSet></dateSet>
    <culturalContextSet></culturalContextSet>
    <descriptionSet></descriptionSet>
    ...
</work>
```

Momenteel bestaan er twee XML-schema's van de metadatastandaard VRA Core 4.0: een *restricted* en een *unrestricted* schema. De *unrestricted* versie legt geen beperkingen op aan de datawaarden in een VRA Core 4.0-record, terwijl de *restricted* versie dit wel doet.

Ten slotte moet aangehaald worden dat het doel van VRA Core 4.0 *file sharing* is. Dit betekent dat de standaard tekortschiet om alle data in XML op te slaan. Vermoedelijk zal de standaard uitgebreid worden met andere subelementen en attributen om aan de noden van een organisatie te voldoen. Als uitwisselingsschema volstaat de VRA Core 4.0 echter.

5.2.9 EAD

EAD staat voor Encoded Archival Description, een metadatastandaard die ontwikkeld is door de bibliotheek van de University of Berkeley in Californië. De standaard ontstond uit de behoefte om nog meer informatie in te voeren dan mogelijk is met MARC.[93]

De eisen luidden onder meer:

- weergave van uitgebreide en intergerelateerde beschrijvende informatie
- behoud van hiërarchische relaties tussen verschillende niveaus van beschrijving

93 Zie LOC (2008b) voor de homepage van EAD, LOC (2006a) voor voorbeelden van EAD-records in XML, DEN (2007) (toelichting door DEN bij EAD).

- weergave van beschrijvende informatie die afkomstig is van verschillende hiërarchische niveaus
- navigatie in een hiërarchische informatiestructuur
- ondersteuning voor elementspecifieke indexering en navigatie.

EAD is SGML-gebaseerd maar ondersteunt ook XML. De toegelaten elementen voor de beschrijving van een handschriftencollectie en de ordening van die elementen (wat zijn de nodige elementen, welke elementen zijn binnen andere elementen toegelaten,...) worden in de EAD Document Type Definition (DTD) gespecificeerd. De gespecificeerde tagset van EAD bevat 146 elementen en wordt zowel gebruikt voor de beschrijving van een collectie in zijn geheel als voor de encodering van de verschillende collectieniveaus (delen van collecties en afzonderlijke archiefstukken). De vele mappingsmogelijkheden tussen EAD en andere metadatastandaarden, waaronder MARC en Dublin Core, doen de flexibiliteit en de interoperabiliteit van de data toenemen.[94]

Een beschrijving in EAD bestaat uit verschillende delen:

- De EAD-*header* bevat de titel en gedetailleerde informatie over de collectie en het document. De elementen in de *header* worden vaak ook gemapt naar Dublin Core-elementen.
- De archiefbeschrijving bestaat uit de Data Item Description (DID) met eventuele aanvullende beschrijvingen. Het grootste deel betreft een volledige inventaris van de collectie.

De DID bevat dus een volledige beschrijving van de collectie, waaronder ook gegevens over de beheerder (persoon of organisatie), de taal, een korte toelichting,... De DID kan aangevuld worden met de volgende elementen:

- een biografische beschrijving van de betrokken persoon of organisatie
- een uitgebreide beschrijving van de collectie
- beschrijving van objecten die met de collectie in verband staan

94 Voor documentatie bij crosswalks en mappings, zie Baca, Clarke, et al. (2009) (verschillende metadatastandaarden), LOC (1999) ('EAD-crosswalks') en het artikel McCrory (2005) voor een commentaar bij EAD-crosswalks.

- objecten die deel uitmaken van de collectie maar die ervan gescheiden zijn (bijvoorbeeld voor een speciale behandeling, omwille van specifieke opslagbehoeften,…)
- een lijst van onderwerpen of trefwoorden voor de collectie
- gegevens over het materiaal in de collectie.

De inventaris van de collectie wordt progressief in kleinere stukken opgedeeld, die steeds 'fijner' beschreven worden. Dit zorgt ervoor dat men bij zoekopdrachten of bij het inventariseren de gewenste informatiediepte kan bepalen.

Door de Research Libraries Group worden via een soort 'coördinatiecentrum' richtlijnen aangeboden.[95] Leden kunnen informatie uitwisselen, die vervolgens geïndexeerd wordt en door een interface doorzoekbaar wordt. Zo kunnen onderzoekers met één enkele query in honderden collecties tegelijk zoeken.

95 RLG (2002).

Voorbeeld van een EAD-bestand:

```
<filedesc>
    <titlestmt>
        <titleproper>Interview met een oudstrijder
            <date>1999</date>
        </titleproper>
        <author>Vereniging voor oudstrijders</author>
    </titlestmt>
    <notestmt>
        <note>
        <p>
            <subject>Wereldoorlog II</subject>
        </p>
    </note>
    </notestmt>
</filedesc>
```

Voordelen

- Ondersteunt hiërarchie
- Kan gemapt worden naar MARC en Dublin Core

Nadelen

- SGML is minder gebruiksvriendelijk.

5.2.10 SPECTRUM

SPECTRUM is een open standaard die wordt onderhouden en gestuurd door de Britse MDA. De standaard beschrijft procedures voor de documentatie, behandeling en identificatie van objecten. Bovendien wordt onder meer aandacht besteed aan rechtenbeheer, uitleenbeheer en risicobeheer. SPECTRUM evolueert voortdurend en kan dan ook aanhoudend uitgebreid worden.

SPECTRUM is een Britse standaard die ontwikkeld werd met de hulp en het inzicht van honderden ervaringsdeskundigen in de museumbranche.

De standaard wordt bijgevolg beschouwd als de 'industriestandaard' voor documenteren.[96]

5.2.11 ISAD(G)

ISAD(G) staat voor General International Standard Archival Description en is een archiefstandaard die regels voorschrijft voor de beschrijving van archiefcollecties en -objecten.[97] Hierin wordt vooral een hiërarchische voorstelling van groot (een collectie) naar klein (één object) beoogd (zoals in EAD) waarbij telkens de relaties tot de andere niveaus worden aangegeven. Op een gelijkaardige manier worden ook regels voorgeschreven voor de opgave van referenties, titels, datering, enzovoort.

Een afgeleide standaard van ISAD(G) is SEPIADES (SEPIA Data Element Set).[98] Deze standaard is gericht op de beschrijving en het beheer van fotografische collecties en bevat eenentwintig kernelementen (de *core*) die aangevuld kunnen worden met meer dan vierhonderd andere elementen. SEPIADES is eveneens hiërarchisch opgevat, net zoals ISAD(G). De standaard voorziet echter niet in een eigen codering en moet dus veeleer als een handleiding voor collectiebeschrijvingen worden opgevat. Voor de opslag van records wordt Dublin Core aangeraden.

5.2.12 ISAAR[99]

5.2.12.1 Achtergrond en doelstelling

ISAAR(CPF), voluit de International Standard Archival Authority Record for Corporate Bodies, Persons and Families, is een norm die richtlijnen biedt voor beschrijvingen van entiteiten (organisaties, personen en families) die betrokken zijn bij de vorming en het beheer van archieven.

Die beschrijvingen hebben o.m. de volgende doelstellingen:

96 Zie Collections Trust (2009) (de referentiesite van SPECTRUM) en Shepherd (2002) voor crosswalks en een vergelijking tussen ISAD(G) (cf. §4.2.11) en SPECTRUM.

97 Zie ICA (2000) (de referentiepagina van ISAD(G)). Voor crosswalks, zie LOC (1999) (met EAD) en Shepherd (2002) (vergelijking met SPECTRUM).

98 Klijn (2004).

99 De omschrijving van ISAAR is gebaseerd op de Nederlandse vertaling van het ICA-document: Archiefschool en VVBAD, ICA (2006).

- de beschrijving van een organisatie, persoon of familie als eenheid in de context van een archivistisch beschrijvingssysteem,

- de regeling van de creatie en het gebruik van ontsluitingstermen in archivistische beschrijvingen,

- de documentatie van de relaties tussen verschillende archiefvormers onderling en tussen die entiteiten en de door hen gevormde archieven of andere bronnen die op hen betrekking hebben.

De beschrijving van archiefvormers is een wezenlijke taak van archivarissen, ongeacht of de beschrijvingen zich in papieren of in digitale systemen bevinden. Hiervoor is een volledige en geactualiseerde documentatie van de context van de archiefvorming en het gebruik noodzakelijk. Vooral de herkomst van de archieven is van belang.

ISAD(G) (§5.2.11) kan men als de parallelnorm van ISAAR beschouwen. ISAD(G) richt zich immers op de beschrijving van contextuele informatie in archivistische beschrijvingen op ieder niveau. Zoals aangegeven suggereert ISAD(G) ook om contextgegevens afzonderlijk vast te leggen en te onderhouden, zodat deze beschrijvingen eenvoudig gekoppeld kunnen worden aan andere gegevens die (een aspect van) het betreffende archief beschrijven.

Voor de afzonderlijke beschrijving en het onderhoud van dergelijke contextgegevens kan men meerdere argumenten opwerpen. Zo kan men bijvoorbeeld beschrijvingen van archiefvormers en contextgegevens koppelen aan beschrijvingen van archiefstukken van diezelfde archiefvormer(s) die zich mogelijk in verschillende archieven bevinden. Bovendien kan men ze bijvoorbeeld in verband brengen met beschrijvingen van andere bronnen, zoals bibliotheekmateriaal of museumobjecten, die met de entiteit verband houden. De beschrijving van dergelijke relaties komt het archiefbeheer ten goede en ondersteunt het onderzoek.

Bewaarplaatsen die archiefstukken van eenzelfde bron beheren, kunnen de contextgegevens over de bron eenvoudiger uitwisselen of ernaar refereren indien deze op een gestandaardiseerde manier beheerd zijn. Standaardisatie is vooral van belang in een internationale context aangezien het delen of koppelen van contextinformatie mogelijk nationale grenzen overschrijdt. Het internationale karakter van archivering, zowel vroeger als nu, stimuleert in ieder geval verdere internationale standaardisatie, die dan weer de uitwisseling van contextgegevens zal bevorderen. Processen zoals kolonisatie, immi-

gratie en handel hebben bijvoorbeeld bijgedragen aan het multinationale karakter van archivering.

ISAAR staat in voor de creatie van consistente, geschikte en duidelijke beschrijvingen van archiefvormende organisaties, personen en families, om het gemeenschappelijk gebruik van archivistische geautoriseerde beschrijvingen te bevorderen. De standaard moet worden gebruikt in combinatie met bestaande nationale normen of als basis voor de ontwikkeling van nationale regels.

Archivistische geautoriseerde beschrijvingen zijn vergelijkbaar met bibliografische geautoriseerde beschrijvingen aangezien ze beiden de creatie van gestandaardiseerde ontsluitingstermen ondersteunen. De naam van de archiefvormer van de beschrijvingseenheid is één van de belangrijkste ontsluitingstermen. Ontsluitingstermen kunnen kwalificaties meekrijgen die essentieel zijn om de identiteit van de benoemde entiteit kenbaar te maken, zodat een nauwkeurig onderscheid kan worden gemaakt tussen verschillende entiteiten met een gelijke of gelijkaardige naam. De archivistische beschrijvingen moeten echter aan meer eisen voldoen dan bibliografische. Dat heeft te maken met het belang dat archivistische beschrijvingssystemen hechten aan de documentatie van archiefvormers en de context van archiefvorming. Archivistische beschrijvingen reiken daarom verder en bevatten doorgaans meer gegevens dan bibliografische.

Het hoofddoel van ISAAR is dus algemene richtlijnen te bezorgen voor de standaardisatie van beschrijvingen van archiefvormers en de context van archiefvorming, die instaan voor:

- het raadplegen van archieven, door middel van beschrijvingen van de context en hun relaties met beschrijvingen van vaak verschillende en fysiek verspreide archiefstukken,
- het begrip van de context waarin archieven zijn ontstaan met het oog op een inzicht in de betekenis en het belang van die archieven,
- een nauwkeurige identificatie van archiefvormers en hun relaties met andere entiteiten. Zo worden ook administratieve veranderingen binnen organisaties of veranderingen in de persoonlijke omstandigheden van individuen en families gedocumenteerd,
- de uitwisseling van die beschrijvingen tussen instellingen, systemen en/of netwerken.

5.2.12.2 **Vorm**

In eerste instantie bepaalt ISAAR welke gegevens in een archivistische geautoriseerde beschrijving kunnen voorkomen. De inhoud van de elementen wordt bepaald door de instantie die verantwoordelijk is voor de beschrijvingen.

Ieder gegevenselement van ISAAR bestaat uit:

- de naam van het beschrijvingselement,
- een formulering van het doel van het beschrijvingselement,
- een formulering van de regel(s) die op het element van toepassing is (zijn),
- de relevantie ervan en voorbeelden die de toepassing van de regel illustreren.

De paragrafen zijn enkel genummerd om ze te kunnen citeren. De nummering wordt dus best niet gebruikt om beschrijvingselementen aan te duiden of om de volgorde of structuur van de beschrijvingen voor te schrijven.

De beschrijvingselementen zijn in vier velden ingedeeld: identiteit, beschrijving, relaties en beheer.

Van alle elementen zijn bij iedere beschrijving vier elementen noodzakelijk: de soort entiteit, de geautoriseerde naam/namen, de bestaansperiode en een identificatiecode van de geautoriseerde beschrijving.

De aard van de beschreven entiteit en de eisen van het systeem of netwerk waarin de verantwoordelijke voor de beschrijving werkt, zullen bepalen welke optionele beschrijvingselementen in een bepaalde beschrijving worden gebruikt en of deze elementen in een verhalende en/of gestructureerde vorm worden gepresenteerd.

Veel beschrijvingselementen in een volgens ISAAR(CPF) opgestelde beschrijving zullen als ontsluitingstermen dienst doen. Regels en afspraken over de standaardisatie van ontsluitingstermen kunnen zowel nationaal als volgens taal ontwikkeld worden. Dat geldt ook voor de woordenlijsten en afspraken voor de opstelling of selectie van de inhoud van deze elementen.

De norm geeft nogal wat voorbeelden. Deze zijn echter louter illustratief en mogen in geen geval als voorschriften of uitbreidingen op de regels

beschouwd worden. Ieder voorbeeld wordt gedocumenteerd met een vermelding van de instantie die het geleverd heeft en eventuele verdere aantekeningen. Deze worden steevast voorafgegaan door de opmerking: 'N.B. verwar de bronverwijzing van het voorbeeld of mogelijke aantekeningen niet met het voorbeeld zelf'.

Zoals gezegd is de norm bedoeld om samen gebruikt te worden met ISAD(G) en met nationale archivistische beschrijvingsnormen. Als die normen binnen een archivistisch beschrijvingssysteem of netwerk samen worden toegepast, dan kunnen geautoriseerde beschrijvingen met beschrijvingen van archieven en vice versa verbonden worden. Dit kan bijvoorbeeld aan de hand van de elementen 'Naam van de archiefvormer(s)' en 'Institutionele geschiedenis / Biografie' in een beschrijving volgens ISAD(G). ISAAR wordt bovendien in combinatie gebruikt met nationale normen en afspraken. Archivarissen kunnen bijvoorbeeld nationale normen volgen bij de bepaling van herhalende elementen. Terwijl in veel landen slechts één geautoriseerde naam voor een bepaalde entiteit is toegestaan, laat men het in andere landen toe om meerdere namen te noteren.

ISAAR behandelt slechts een gedeelte van de noodzakelijke voorwaarden voor de uitwisseling van archivistische geautoriseerde gegevens. Een succesvolle digitale uitwisseling van gegevens via computernetwerken is immers ook afhankelijk van het gebruik van een geschikt communicatieformaat door de bewaarplaatsen. Encoded Archival Context (EAC) wordt als geschikt communicatieformaat aanbevolen. EAC ondersteunt de uitwisseling via het World Wide Web van archivistische geautoriseerde beschrijvingen die opgesteld zijn volgens ISAAR(CPF). Het is een Document Type Definition (DTD) in XML en SGML.

5.3 Metadatastandaard voor preservering

5.3.1 PREMIS[100]

5.3.1.1 Achtergrond

PREMIS is een metadatastandaard voor langetermijnbewaring. Ze werd ontwikkeld door een werkgroep van internationale experts op het gebied van bewaring en metadata, die in juni 2003 werd opgericht. De leden vertegenwoordigen verschillende sectoren zoals bibliotheken, musea, archieven, overheidsinstellingen en de privé-sector. De werkgroep wilde een set implementeerbare preserveringsmetadata[101] ontwikkelen, die ondersteund wordt door richtlijnen en aanbevelingen voor de creatie, het beheer en het gebruik ervan. In mei 2005 verscheen een eerste versie van de bevindingen in het rapport *Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group*.[102] Het rapport bevatte onder meer talrijke bronnen over preserveringsmetadata, maar het is in eerste instantie een 'datawoordenboek' met definities van preserveringsmetadata, die gebaseerd is op het referentieel OAIS-model. Naast de *Data Dictionary* publiceerde de werkgroep ook een set XML-schema's voor de ondersteuning van het woordenboek in digitale archiveringssystemen.

5.3.1.2 Doelstelling

De PREMIS *Data Dictionary* omschrijft 'preserveringsmetadata' als de informatie die een bewaarplaats of digitaal archief nodig heeft om het digitale proces van langetermijnbewaring te ondersteunen. De werkgroep onderzocht hiervoor metadata die de volgende functies documenteren:

- uitvoerbaarheid
- weergave

100 De beschrijving van PREMIS is deels gebaseerd op Higgins (2007).

101 Hiermee worden metadata bedoeld die nodig zijn om de langetermijnbewaring ('preservatie') van digitale data te garanderen.

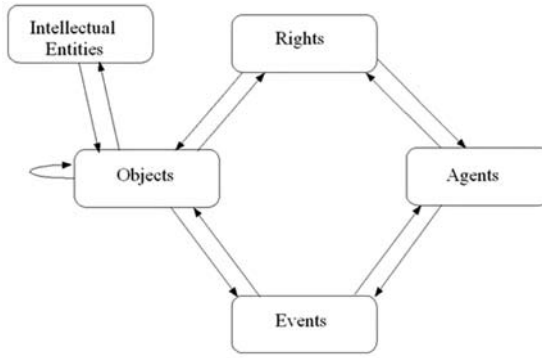
102 Ondertussen ge-updated met een versie van maart 2008, PREMIS Editorial Commitee (2008).

- verstaanbaarheid
- authenticiteit
- identiteit

Het is duidelijk dat het hier moet gaan om verschillende soorten metadata: administratieve metadata (o.m. rechtenbeheer), technische en structurele metadata. Vooral de documentatie van de herkomst of 'geschiedenis' van een digitaal object en de documentatie van de relaties tussen de verschillende digitale data, staan centraal.

5.3.1.3 Vorm

De werkgroep ontwikkelde eerst een eenvoudig datamodel om de semantische eenheden in de *Data Dictionary* te definiëren. Het datamodel bestaat uit vijf entiteiten die belangrijk zijn voor de doelstelling van digitale langetermijnbewaring: *Intellectual Entities*, *Objects*, *Events*, *Agents* en *Rights* (zie figuur 16).



Figuur 16: Datamodel PREMIS[103]

- *Intellectual Entity*: een intellectuele eenheid met het oog op de beschrijving of het beheer van het digitaal object, bijvoorbeeld een boek of pagina's in een boek, een foto, een databank.
- *Object*: een discrete eenheid van informatie in digitale vorm.

103 Figuur is ontleend aan PREMIS Editorial Committee (2008), p. 5.

- *Event*: een actie met betrekking tot een *object* of *agent*.

- *Agent*: een persoon, organisatie of softwaretoepassing die in relatie staat met een *event* van een *object* of met de *rights* van of op het *object*.

- *Rights*: de rechten of toestemmingen van een *object* of *agent*.

De *Data Dictionary* definieert semantische eenheden. Elke semantische eenheid wordt verbonden met één van de vijf entiteiten. In die optiek kan een semantische eenheid beschouwd worden als een eigenschap van een entiteit. Zo is de semantische eenheid 'grootte' bijvoorbeeld een eigenschap van een *object*. Semantische eenheden hebben ook een waarde. De 'grootte' van een bepaald *object* is bijvoorbeeld '843200004'.

In sommige gevallen kunnen semantische eenheden de vorm aannemen van een container die een set gerelateerde eenheden groepeert. Die gegroepeerde semantische eenheden worden dan de semantische componenten van de container genoemd. Sommige containers zijn uitbreidbaar en kunnen metadata van een extern schema bevatten. Dit laat toe om PREMIS uit te breiden met elementen die geen betrekking hebben op langetermijnbewaring maar die bijvoorbeeld voor de beschrijving wel nuttig kunnen zijn.

Tussen twee entiteiten is een relatie mogelijk, die op verschillende manieren geïnterpreteerd kan worden. De bewering 'object A heeft als formaat B', kan bijvoorbeeld als een relatie tussen A en B opgevat worden. Het PREMIS-model behandelt formaat B echter als een eigenschap van Object A. PREMIS reserveert het concept relatie enkel voor de associaties tussen entiteiten, bijvoorbeeld tussen twee of meer *objects* of tussen een *object* en een *agent*.

Een *object* bestaat uit drie subtypes: bestand, bitstroom en representatie.

- Een bestand is een genoemde en geordende opeenvolging van bytes die door het besturingssysteem gekend is. Een bestand kan 0 of meer bytes bevatten en heeft een bestandsformaat, een toegangspermissie en eigenschappen van een bestandssysteem zoals grootte en datum van de laatste wijziging.

- Een bitstroom betreft de data in een bestand die betekenisvolle gemeenschappelijke eigenschappen hebben. Een bitstroom kan enkel naar een bestand worden omgevormd mits de toevoeging van een bestandsstructuur (*headers*,...) en de herformattering van de bitstroom naar een

bepaald bestandsformaat. Bijvoorbeeld audiodata in een WAV-bestand, een beeld in een TIFF 6.0-bestand.

- Een representatie is een set van bestanden, met de structurele metadata die nodig zijn voor een volledige en degelijke vertolking van een *intellectual entity*. Een artikel kan bijvoorbeeld volledig in een PDF-bestand zitten. Het PDF-bestand is dan de representatie van het artikel. Een ander artikel kan gerepresenteerd zijn door twaalf TIFF-bestanden, één voor elke pagina, en een XML-bestand dat de structuur beschrijft. De dertien bestanden vormen dan de representatie van het artikel.

Een *event* betreft metadata in verband met acties. De beschrijving van acties die een digitaal object veranderen, is essentieel voor de documentatie van de herkomst en dus voor de garantie op authenticiteit. Het hangt af van de bibliotheek welke acties als *event* worden opgenomen. Volgens het datamodel kunnen *objects* op twee manieren met *events* geassocieerd worden. Als een *object* gerelateerd is aan een tweede *object* via een *event*, dan wordt de *identifier* van de *event*-eenheid opgenomen in de relatiecontainer als de semantische component *relatedEventIdentification.* Als een *object* geassocieerd wordt met een *event* zonder dat het in relatie staat met een ander *object*, dan wordt de *identifier* van het *event* opgenomen in de container *linkingEventIdentifier*. Als een bibliotheek bijvoorbeeld een XML-bestand (*object* A) oplaadt en er een genormaliseerde versie van creëert (*object* B) via een applicatie (*event* 1), dan kan dit als volgt in een *relationship* worden beschreven:

relationshipType = “derivation”
relationshipSubType = “derived from”
relatedObjectIdentification
relatedObjectIdentifierType = “local”
relatedObjectIdentifierValue = “A”
relatedObjectSequence = “not applicable”
relatedEventIdentification
relatedEventIdentifierType = “local”
relatedEventIdentifierValue = “1”
relatedEventSequence = “not applicable”

Een *agent* is belangrijk, maar heeft geen centrale plaats in de *Data Dictionary*. Die schrijft enkel een manier voor om een *agent* te identificeren en kent er een classificatie aan toe (persoon, organisatie of software). Dit is uiteraard niet voldoende maar wordt in PREMIS verder buiten beschouwing gelaten. In het datamodel is een relatie mogelijk tussen de entiteit *agent* en de

entiteit *event*, maar niet tussen *agent* en *object*. *Agents* beïnvloeden *objects* enkel indirect via een *event*. Omdat een *agent* verschillende rollen kan vervullen in meerdere *events*, is de rol van een *agent* een eigenschap van de *event*-entiteit.

De semantische eenheden die in PREMIS beschreven worden, zijn met elkaar verbonden via een aantal structurele afspraken die de *Data Dictionary* mee organiseren en haar implementatie ondersteunen. Die afspraken hebben betrekking op het gebruik van *identifiers*, als een manier om relaties te leggen of om metadata aan *objects* te relateren.

Instanties van *objects*, *events*, *agent*s en *rights* zijn uniek identificeerbaar via een set van semantische eenheden onder de *identifier*-container. Deze semantische eenheden volgen een gelijkaardige syntaxis en structuur, ongeacht het type entiteit:

[entity type]Identifier
[entity type]IdentifierType: domain in which the identifier is unique
[entity type]IdentifierValue: identifier string

Een voorbeeld van een *object*:
ObjectIdentifier
ObjectIdentifierType: NRS
ObjectIdentifierValue: http://nrs.harvard.edu/urn-3:FHCL.Loeb:sa1

Een voorbeeld van een *event*:
EventIdentifier
EventIdentifierType: NRS
EventIdentifierValue: 716593

Het *identifier*-type 'NRS' duidt aan dat de *identifier* uniek is binnen het domein van de *Name Resolution Service* die de *identifiers* toekent. Als de bibliotheek digitale objecten en hun metadata uitwisselt, is het nodig dat het type van de *identifier* wordt meegegeven.

Identifiers zijn herhaalbaar voor *objects* en *agents*, maar niet voor *rights* en *events*. *Objects* en *agents* kunnen verschillende identiteiten hebben in een globale omgeving en in verschillende systemen. Daarom is het noodzakelijk dat aan deze entiteiten verschillende *identifiers* kunnen worden toegewezen. De context van *rights* en *events* is beperkt tot de bewaarplaats, bibliotheek of archief, waardoor één *identifier* telkens volstaat.

5.4 Conceptuele modellen

5.4.1 FRBR[104]

5.4.1.1 Achtergrond en doelstelling

FRBR staat voor Functional Requirements for Bibliographic Records. Het gaat om een conceptueel model dat gebruikt wordt in de bibliografische wereld en dat in 1998 door de IFLA (International Federation of Library Associations and Institutions) in een rapport werd gepresenteerd. De tekst biedt een duidelijk gedefinieerd en gestructureerd kader waarin data van bibliografische records gerelateerd worden aan de noden van gebruikers van deze records. De gebruiker staat dus centraal in dit kader. De twee belangrijke concepten in FRBR zijn bijgevolg 'gebruikerstaken' en 'bibliografische entiteiten'.

5.4.1.2 Vorm

Het FRBR-model omschrijft de vier taken *Find*, *Identify*, *Select* en *Obtain*, die het vervolg van het model bepalen:

- *Find*: vindt entiteiten die aan bepaalde criteria van de gebruiker voldoen. Deze voorwaarden zijn gebaseerd op de attributen of relaties van die entiteit.
- *Identify*: identificeert entiteiten. De taak bevestigt bijvoorbeeld dat de beschreven entiteit overeenkomt met de gezochte entiteit of maakt een onderscheid tussen entiteiten met zeer gelijkaardige eigenschappen.
- *Select*: selecteert entiteiten die aan de behoefte van de gebruiker beantwoorden. De taak kiest bijvoorbeeld een entiteit omdat die aan bepaalde gebruikersverwachtingen zoals de inhoud of het formaat voldoet.
- *Obtain*: verleent toegang tot beschreven entiteiten.

De bibliografische entiteiten kunnen in groepen worden ingedeeld. De meest bekende elementen van het model zijn de entiteiten in groep 1. Die bevatten de producten of resultaten van een intellectuele of artistieke inspanning en zijn in bibliografische records beschreven. Het gaat echter om 'conceptuele'

104 Zie de rapporten en versies op IFLA (2008a).

entiteiten, wat betekent dat ze zeker geen concrete records in een databank kunnen representeren. Het gaat in het bijzonder om de vier volgende entiteiten:

- *Work*: een intellectuele of artistieke creatie
- *Expression*: de intellectuele of artistieke realisatie van een *work*.
- *Manifestation*: de fysieke belichaming van een *expression* of *work*
- *Item*: een versie van een *manifestation*

Een *work* is een abstract concept, het achterliggend idee, vóór het in een vorm wordt vastgelegd. Een *expression* is de representatie van het abstract concept, zoals woorden of muzieknoten. Deze representatie is nog steeds conceptueel en dus niet tastbaar. Een *work* kan verschillende *expressions* hebben, bijvoorbeeld in verschillende talen. Een *manifestation* is een set van fysieke zaken die een *expression* of *work* bevatten. Een *manifestation* kan verschillende *expressions* bevatten, bijvoorbeeld in het geval van een CD waarop verschillende liedjes staan, die samen een *expression* zijn van een individueel *work*. Een *item* ten slotte is een concreet exemplaar van een *manifestation*. Die kan fysiek zijn maar kan even goed een kopie van een bestand zijn.

De entiteiten in groep 2 zijn de verantwoordelijken voor de intellectuele of artistieke inhoud, de fysieke productie en distributie of het beheer van de entiteiten van groep 1. Het FRBR-rapport omschrijft slechts twee entiteiten in groep 2: *person* en *corporate body*, hoewel *family* hier vaak als derde entiteit wordt opgegeven. Deze entiteiten zijn verantwoordelijk voor een entiteit in groep 1. Bijvoorbeeld: de auteur creëert een *work*, de vertaler realiseert een *expression*, de uitgever staat in voor een *manifestation* en een bibliotheek bezit een *item*.

Entiteiten in groep 3 zijn de onderwerpen van *works*. Iedere entiteit uit groep 1 of groep 2 ressorteert onder deze categorie, alsook de bijkomende entiteiten *concept*, *object*, *event* en *place*.

Het FRBR-rapport biedt dus een conceptueel model en geen concreet datamodel. Het model is bijgevolg onderworpen aan beslissingen op implementatiegebied. Verschillende implementaties vertegenwoordigen verschillende functionaliteiten, die dan weer enkel mogelijk zijn in systemen die het FRBR-principe implementeren. Bibliotheekcatalogi die gebaseerd zijn op

de FRBR-principes kunnen de zoekresultaten eenvoudiger groeperen. Zo kan bijvoorbeeld een lijst met alle werken van een bepaalde auteur worden weergegeven. Daarnaast kan men verschillende expressies van een werk presenteren, bijvoorbeeld gegroepeerd volgens formaat, taal, uitvoerder of volgens om het even welk attribuut.

Het FRBR-model is in alle bibliotheken bruikbaar, hoewel OCLC-studies aangetoond hebben dat niet alle werken baat hebben bij een toepassing van de FRBR-principes. Er is een zekere kost verbonden aan de implementatie van het FRBR-model. FRBR zal vooral bruikbaar zijn voor literaire of artistieke werken omdat hier doorgaans meerdere versies of uitvoerders van bestaan.

5.4.2 CIDOC-CRM

Het CIDOC conceptueel referentiemodel werd ontwikkeld door de ICOM/CIDOC Documentation Standards Group. Het model schrijft een ontologie voor culturele erfgoedinformatie voor. Het CIDOC *Conceptual Reference Model* (CRM) levert de definities en structuur voor de beschrijving van expliciete en impliciete concepten en relaties in de documentatie van cultureel erfgoed. CIDOC-CRM wil de kennis over informatie met betrekking tot cultureel erfgoed promoten door een gemeenschappelijk en uitbreidbaar semantisch kader op te zetten waarnaar alle informatie over cultureel erfgoed gemapt kan worden. CIDOC-CRM wil een gemeenschappelijke taal voorstellen voor domeinexperts en ontwikkelaars om de eisen van informatiesystemen te formuleren en het model moet als gids dienen voor het conceptueel modeleren. Op die manier wil men een gemeenschappelijke semantiek aanleveren om te kunnen bemiddelen tussen de verschillende informatiebronnen over cultureel erfgoed zoals die gepubliceerd zijn door musea, bibliotheken en archieven.

Een duidelijke visie op de ontologie is noodzakelijk. Er moet immers goed overwogen worden wat al dan niet door de ontologie ondersteund wordt. Hier moet een onderscheid worden gemaakt tussen een theoretische en een praktische invalshoek.

Het theoretische bereik van CIDOC-CRM betreft het domein dat CIDOC-CRM wil toepassen indien er voldoende tijd en middelen zouden zijn. Het praktische perspectief is een subset van het theoretische en dekt de realisaties van CIDOC-CRM tot nu toe. Dat wil zeggen dat er mappings worden voorzien die de vertaling van en naar de brondocumenten verzorgen. Deze invalshoek kan dus veranderen naargelang andere standaarden relevant of belangrijk worden.

De theoretische invalshoek van CIDOC-CRM betreft de nodige informatie voor de wetenschappelijke beschrijving van collecties cultureel erfgoed samen met de nodige vertalingen om informatie te kunnen uitwisselen tussen heterogene informatiebronnen. Met cultureel erfgoed worden alle types van materiaal bedoeld die verzameld en tentoongesteld worden door de musea en aanverwante instituten. Dit betreffen dus collecties, sites en monumenten die te maken hebben met de geschiedenis, etnografie, archeologie, historische monumenten en collecties kunstwerken. Voor een definitie van cultureel erfgoed wordt verwezen naar de omschrijving door ICOM.[105] CIDOC-CRM stelt als doelstelling voorop dat de kwaliteit van de beschreven informatie voldoende degelijk moet zijn voor academisch onderzoek. CIDOC-CRM richt zich dan ook hoofdzakelijk op museumprofessionals en onderzoekers.

CIDOC-CRM is ook voornamelijk gericht op de beschrijving van contextuele informatie. Dit houdt de historische, geografische en theoretische achtergrond in van de tentoongestelde items, waardoor hun waarde en betekenis toenemen. De nadruk in het model ligt dus op de beschrijvende metadata. Informatie voor administratie en beheer vallen buiten de opzet van CIDOC-CRM.

De praktische invalshoek van CIDOC-CRM is dus een subset van de theore tische. De elementen van de volgende datastructuren zijn tot nu toe opgenomen in CIDOC-CRM. Ze worden geverifieerd door de mappings die worden bijgevoegd bij de ondersteunende documentatie. De volgende lijst geeft hun status weer.

Volledig:

- Dublin Core
- Art Museum Image Consortium (AMICO)
- Encoded Archival Description (EAD)
- MDA SPECTRUM
- Natural History Museum (London) John Clayton Herbarium Data Dictionary
- National Museum of Denmark

105 ICOM (2009).

- International Federation of Library Associations and Institutions (IFLA) Functional Requirements for Bibliographic Records (FRBR)
- OPENGIS
- Association of American Museums Nazi-era Provenance Standard
- MPEG-7
- Research Libraries Group (RLG) Cultural Materials Initiative DTD

In voorbereiding:

- Consortium for the Computer Interchange of Museum Information (CIMI) Z39.50 profile
- Council for Prevention of Art Theft Object ID
- The International Committee for Documentation of the International Council of Museums (CIDOC)
- The International Core Data Structures for Archaeological and Architectural Heritage
- Core Data Index to Historic Buildings and Monuments of the Architectural Heritage
- English Heritage MIDAS – A Manual and Data Standard for Monument Inventories
- English Heritage SMR 97
- Hellenic Ministry of Culture POLEMON Data Dictionary

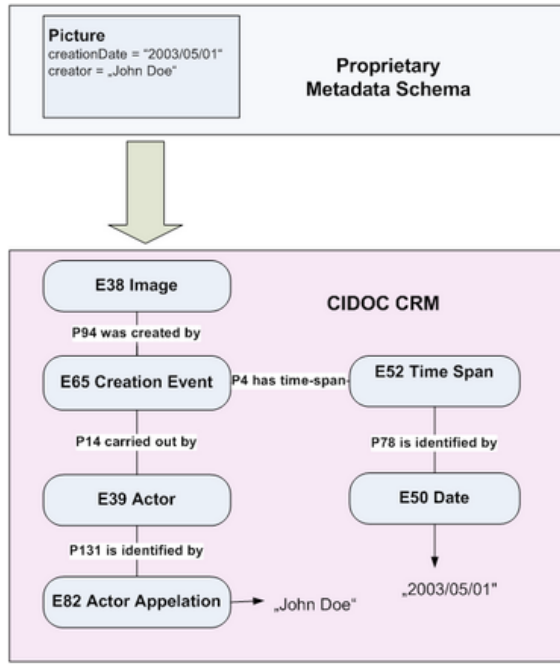
Gewenst:

- FENSCORE
- Sydney University TimeMapper

- Data Service Standards in Archaeology
- Digital Library Metadata
- International Council on Archives (ICA) ISAD(G) – International Standard Archival Description (General)
- Visual Resources Association Core Categories – VRA Core
- Machine Readable Cataloguing – MARC
- CIMI SGML DTD
- Getty Categories for the Description of Works of Art – CDWA
- RSLP Collection Description
- MODES OBJECT FORMAT

Het kader voorziet 84 klassen en 141 properties. Klassen worden doorgaans met de notatie 'Enn' weergegeven, properties met 'Pnn', gevolgd door de betreffende naam van de klasse of *property* tussen haakjes. Het grootste voordeel van CIDOC-CRM is zijn algemeenheid. Hierdoor kan vrijwel elke standaard of ieder metadataschema in de culturele erfgoedsector gemapt worden naar CIDOC-CRM. Een ander voordeel van deze standaard is de gelaagdheid, die zeer specifieke zoekvragen toelaat.

Een groot nadeel van de CRM is zijn complexiteit. Algemeen wordt aangenomen dat de standaard niet bedoeld is om rechtstreeks aan de eindgebruiker getoond te worden. Applicaties die CIDOC-CRM gebruiken, zoals I-MASS en SCULPTEUR, tonen ofwel een vereenvoudigde versie van het schema ofwel loodsen ze de eindgebruiker met een aantal vragen naar de gewenste informatie. Het is kenmerkend dat bepaalde properties in een metadataschema niet kunnen gemapt worden naar één enkele property maar naar een ketting van gerelateerde klassen en properties. Een metadatarecord waarin een object bijvoorbeeld beschreven wordt met de velden 'CreationDate' en 'Creator', wordt naar CIDOC-CRM gemapt als 'P94(was created by) – E65(CreationEvent) – P14(carried out by) – E39(Actor) – P131(is identified by) – E82(Actor Appellation)' en 'P94(was created by) – E65(CreationEvent) – P4(has time-span) – E52(Time-span) – P78(is identified by) – E50(Date)'. Een voorbeeld van een mapping wordt in figuur 17 schematisch weergegeven.



Figuur 17: Voorbeeld mapping naar CIDOC-CRM[106]

5.4.3 ABC[107]

Naast SPECTRUM en ISAD(G) moet ook het ABC-model toegelicht worden. Dat model is het resultaat van 'The Harmony Project'[108] en is ontworpen als gemeenschappelijk conceptueel model dat de interoperabiliteit tussen meerdere metadata-ontologieën van verschillende domeinen moet vergemakkelijken. ABC is niet als een metadatawoordenschat opgevat maar als een model dat als basis kan dienen voor het ontwerp van specifieke ontologieën.

De belangrijkste doelstelling van het model is de weergave van het volledige traject van een object. Zo worden de creatie, de evolutie en de overgangen die het object ondergaat, beschreven, bijvoorbeeld: waar en wanneer het interview is afgenomen, wie het afnam, transfers naar andere media, enzovoort. Op deze manier kan de hele levenscyclus van een object worden opge-

106 Figuur is ontleend aan Haslhofer en Hecht (2005).

107 Lagoze en Hunter (2001). Zie ook de bespreking van het ABC-model in Mannens, Paridaens, et al. (2007), p. 93-94.

108 Harmony (2009).

vraagd en dankzij de bidirectionele relaties kan ook informatie opgevraagd worden over elk object dat gedurende zijn levenscyclus verbonden is met dat object.

Het ABC-model voorziet ook in een hiërarchie voor objecten en eigenschappen. Zo kunnen objecten worden georganiseerd in een hiërarchie waarbij elke subklasse extra informatie over het object levert. Die hiërarchie vereenvoudigt de interoperabiliteit tussen verschillende standaarden dankzij 'partial understanding'. Indien een bepaald object namelijk niet over een tegenhanger in de andere standaard beschikt, dan kan men op een hoger niveau op zoek gaan naar een equivalent.

De hiërarchie voor eigenschappen laat eveneens toe die te verfijnen met zogenaamde subeigenschappen. Hierdoor kan men niet alleen een gewenst informatieniveau bepalen (internen krijgen toegang tot alle informatie, externen enkel tot een bepaald niveau) maar ook de nauwkeurigheid van de zoekactie.

De hiërarchieën, levenscycli en bidirectionele relaties laten bovendien toe om van eenzelfde object/item verschillende representaties te verkrijgen.

5.4.4 GAMA[109]

GAMA, Gateway to Archives of Media Art, is een interdisciplinair project dat van start ging op 1 november 2007. Aan het project nemen 19 organisaties deel uit de Europese cultuur-, kunst- en technologiesector. Doelstelling van het project is de ontwikkeling van een centraal online portaal dat de toegang verleent tot verschillende Europese mediakunstcollecties voor geïnteresseerden, curatoren, artiesten, academici en onderzoekers. Het project resulteerde onder meer in een ontologie die zich richt op de beschrijving van mediakunsten. De Europese Commissie steunt het project met 1.2 miljoen euro via het *e*contentplus programma.[110]

Mediakunst, de creatie van kunstwerken met behulp van nieuwe mediatechnologieën, is een van de populairste eigentijdse kunstgenres. In mediakunst wordt de spanning tussen cultuur en technologie en de ambigue rol van nieuwe mediatechnologieën onderzocht. Tegelijkertijd brengt mediakunst die nieuwe technologieën onder de publieke aandacht. In dit opzicht hebben archieven van mediakunst een belangrijke taak te vervullen. Toch zijn die

109 GAMA (2009).

110 Europe's Information Society (2008).

archieven vaak moeilijk toegankelijk en worden ze amper geconsulteerd en gebruikt, wat in schril contrast staat met hun grote betekenis en relevantie in het moderne kunstenlandschap. Het is die discrepantie die in het GAMA-project centraal staat en weggewerkt moet worden.

Het consortium van het GAMA-project bestaat uit de belangrijkste Europese leveranciers van mediakunst. De media die door de partners worden aangeboden, omvat ongeveer 55% van alle mediakunstwerken die online voor Europese culturele archieven en verdelers toegankelijk zijn. Het doel is dan ook om een centraal platform op te stellen dat toegang verleent tot al deze mediakunstarchieven. Het platform moet zich richten op de gebruiker en verschillende talen ondersteunen. Het moet het gebruik, het hergebruik en de zichtbaarheid van de mediakunst en de aangereikte media vergroten. De *gateway* zou uiteindelijk moeten evolueren tot het centraal zoekportaal voor Europese mediakunst.

De toegang tot verschillende archieven moet bovendien geoptimaliseerd worden ongeacht de specifieke structuur van de archieven, de gebruikte metadata, de taal, de digitale formaten en doelstelling. Om aan deze eisen te voldoen, ontwikkelde GAMA een nieuwe ontologie voor de beschrijving van mediakunstwerken.

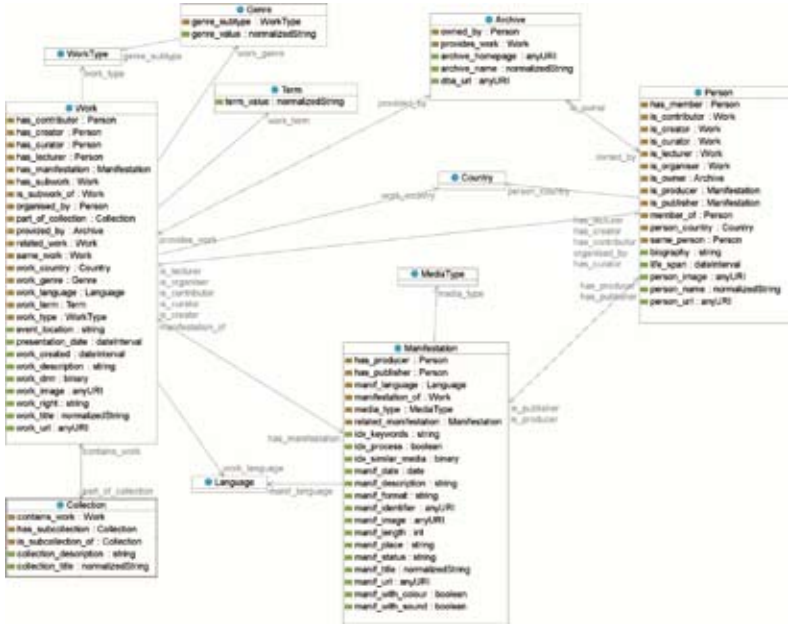
Het metadataschema van GAMA is beschreven in RDF. Klassen, eigenschappen en datatypes zijn de bouwblokken van het schema. Ze behoren allemaal tot dezelfde namespace gama. Het schema bestaat uit elf klassen, die kunnen opgedeeld worden in twee groepen: 'entiteiten' en 'enumeraties'.

De entiteiten bestaan uit de volgende klassen:

gama:Work	kunstwerken, events, en andere bronnen
gama:Person	persoon, instituut, collectief
gama:Manifestation	fysieke representaties van werken
gama:Archive	een archief
gama:Collection	collecties van werken

De enumeraties zijn klassen met gefixeerde instanties en bestaan uit de volgende klassen:

gama:WorkType	lijst van types van werken
gama:MediaType	lijst van mediatypes
gama:Genre	gelaagde hiërarchie van genres
gama:Term	lijst van veelgebruikte termen m.b.t. werken
gama:Country	lijst van landen
gama:Language	lijst van talen



Figuur 18: GAMA metadataschema[111]

5.4.5 FRAR[112]

Catalogi van bibliotheken, musea en archieven bestaan doorgaans uit een set georganiseerde data die de beheerde informatie beschrijven. Voor de groe-

111 Figuur is ontleend aan Simko (2008), p. 26.

112 IFLA (2008c). Cf. *Working group on FRANAR*, http://www.ifla.org/VII/d4/wg-franar.htm {24/12/2008}.

pering van verschillende werken van één auteur of edities van één bepaald werk, bestaat de nood aan gecontroleerde toegangspunten voor auteurs en titels. Sommige auteursnamen of titels kennen verschillende schrijfwijzen, die door de toegangspunten allemaal gekend zijn en met elkaar in verband worden gebracht. Autoriteitscontrole houdt dan zowel het beheer van geautoriseerde schrijfwijzen in als de identificatie van entiteiten die door deze toegangspunten voorgesteld worden. De eindgebruiker kan zo elke mogelijke vorm van de auteursnaam of titel gebruiken om de gewenste informatie te achterhalen.

Voor deze doelstelling werd in 1999 FRANAR opgericht, een werkgroep rond Functional Requirements and Numbering of Authority Records. De groep stelde verschillende doelen voorop, waaronder het onderzoek naar de nodige criteria voor een *authority record*[113] en het onderzoek naar de haalbaarheid van een internationaal gestandaardiseerde nummering voor *authority data*: ISADN (International Standard Authority Data Number).

In het eerste rapport van de werkgroep, *FRAR Functional Requirements for Authority Data*, wordt een conceptueel model beschreven voor de analyse van de functionele benodigdheden voor *authority data*, met het oog op de ondersteuning van autoriteitscontrole en de internationale uitwisseling van die data. Meer specifiek biedt het document een conceptueel model dat ontwikkeld is als:

- een referentiekader dat data uit *authority records* verbindt met noden van gebruikers van de records
- en een leidraad voor de internationale uitwisseling van *authority data*, zowel binnen als buiten de bibliotheeksector.

Het model richt zich op data, zonder rekening te houden met de wijze waarop die verpakt zijn, bijvoorbeeld in records. In feite gaat het om een uitbreiding van de FRBR-standaard, die dezelfde methodologie hanteert.

- Eerst gebeurt de identificatie van de objecten waarin de eindgebruiker geïnteresseerd is. Ieder object of entiteit dient vervolgens als een soort toegangspunt tot een cluster van data. Het model kan ook relaties tussen verschillende types van objecten of entiteiten leggen.

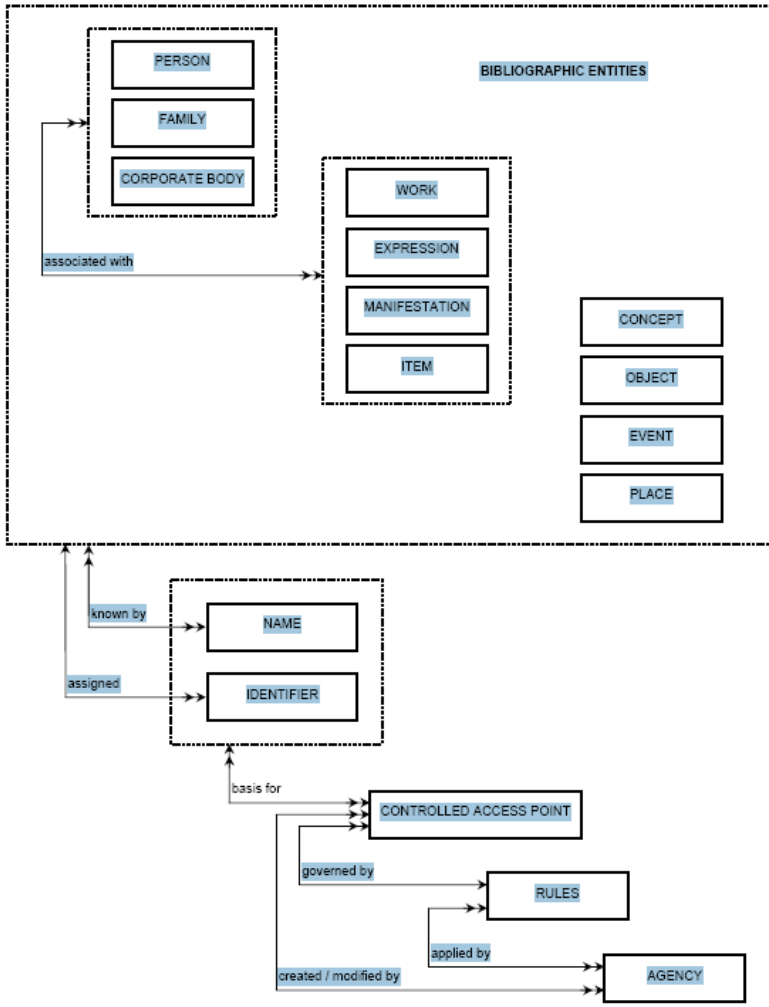
113 *Authority records* bestaan uit geaggregeerde informatie over een bepaalde instantie of entiteit waarvan de naam gebruikt wordt als toegangspunt tot de bibliografische records ervan.

- Na de identificatie van de entiteiten en de bepaling van relaties tussen de verschillende entiteiten, volgt de identificatie van de primaire eigenschappen of attributen van elke entiteit.

Entiteiten uit de bibliografische wereld, zoals die van FRBR, zijn gekend door hun namen en eventuele *identifiers*. Bij autoriteitscontrole worden die namen en *identifiers* als toegangspunt tot de informatie gebruikt. De entiteiten waarop *authority data* zich richten, zijn entiteiten die door FRBR geïdentificeerd werden: *person*, *corporate body*, *work*, *expression*, *manifestation*, *item*, *concept*, *object*, *event* en *place* (en eventueel *family*). Die zijn in figuur 19 bovenaan terug te vinden. Onderaan bevinden zich de namen van de entiteiten, de *identifiers* ervan en de gecontroleerde toegangspunten, die gebaseerd zijn op deze namen en *identifiers* en die werden geregistreerd in *authority files*. Ook de regels die gehanteerd werden bij de opstelling van de catalogus worden in het diagram opgenomen. Zo kan een auteur in een bepaalde catalogus de fysieke persoon zijn met al zijn pseudoniemen, terwijl in een andere catalogus ieder pseudoniem als een afzonderlijke auteur wordt beschouwd. De tekening verduidelijkt ook de relaties tussen de naam (en *identifier*) en de bibliografische entiteit (*person*,...). Een specifieke instantie van iedere entiteit is via één of meerdere namen gekend en ook omgekeerd kan elke naam aan één of meerdere instanties van die entiteiten worden gerelateerd. Een specifieke instantie van een entiteit kan ook worden gerelateerd aan één of meerdere *identifiers*, maar een *identifier* kan slechts met één specifieke instantie van een entiteit in verband staan. In het diagram (figuur 19) is bovendien rekening gehouden met relaties die kunnen bestaan tussen *persons*, *corporate bodies*, *families* enerzijds en *works*, *manifestations*, *expressions* en *items* anderzijds.

Het onderste deel van het diagram laat de associaties zien tussen de namen (en *identifiers*) van de entiteiten en de gecontroleerde toegangspunten van de entiteiten en de associaties tussen entiteiten en de regels voor die entiteiten. Een toegangspunt kan bijvoorbeeld gebaseerd zijn op de combinatie van twee namen en/of *identifiers*, zoals mogelijk het geval is bij werken die aan de hand van naam en titel geïdentificeerd worden. Die toegangspunten worden bepaald door regels die door verschillende instanties toegepast kunnen worden. Een toegangspunt kan dus worden gecreëerd of veranderd door een instantie.

Er is nog een ander type relatie mogelijk, namelijk tussen de instanties van twee verschillende types entiteiten. Het kan bijvoorbeeld gaan om een relatie tussen een persoon en een organisatie. Maar die relaties zijn in het diagram niet weergegeven.



Figuur 19: FRAR diagram[114]

Het model levert entiteitsdefinities die voornamelijk van twee standaarden afkomstig zijn, namelijk FRBR (Functional Requirements for Bibliographic Records) en GARR (Guidelines for Authority Records and References). De entiteiten in het model zijn: *person*, *corporate body*, *family*, *work*, *expression*, *manifestation*, *item*, *concept*, *object*, *event*, *place*, *name*, *identifier*, gecontroleerd toegangspunt, regels en agentschap. Iedere entiteit bevat een aantal

114 Figuur is ontleend aan Patton (2005), p. 8.

attributen die vooral ontleend zijn aan FRBR, GARR en ISAAR(CPF). Verder specificeert het conceptueel model ook alle mogelijke relaties tussen de entiteiten. Voor een verdere uitdieping van de definities van de entiteiten, hun attributen en relaties wordt verwezen naar de specificaties van FRAR.

In de praktijk verloopt het proces voor catalogisering volgens de volgende drie stappen.

- De bibliografische beschrijving: de catalograaf maakt bibliografische beschrijvingen van de bronnen die in de bibliotheek aanwezig zijn. De door hem toegepaste regels zijn gebaseerd op de bronnen waarvan de data zijn afgeleid, de volgorde en de vorm van de individuele data-elementen.

- De formulering van toegangspunten: er worden toegangspunten gecreëerd, zowel van de geautoriseerde vormen van de naam die een auteur, titel, onderwerp,... representeert als van de variabele vormen van de geautoriseerde vorm.

- De registratie van toegangspunten: de catalograaf registreert de gecontroleerde toegangspunten in *authority files*. De registratie van een nieuw toegangspunt kan eventueel leiden tot de herziening van reeds bestaande toegangspunten.

5.4.6 LCSH[115]

5.4.6.1 Achtergrond en doelstelling

De Library of Congress Subject Headings is een thesaurus van indexeringen (*subject headings*) die door de Library of Congress onderhouden wordt. Bibliografische records vormen het toepassingsgebied van de indexeringen. De indexeringen zijn namelijk van toepassing op alle items van een bibliografisch record. Op die manier wordt de toegang tot items met betrekking tot een bepaald onderwerp in een catalogus vereenvoudigd.

LCSH heeft al ruime ingang gevonden in bibliotheken. Hierdoor wordt zoeken naar items in bibliotheekcatalogi die dezelfde zoekstrategie en de LCSH-thesaurus gebruiken, geüniformeerd. Ondanks het brede bereik van LCSH en zijn wijdverbreid gebruik hebben sommige bibliotheken nood aan nog andere indexeringen. Zo heeft de National Library of Medicine van de Verenigde Staten

115 LOC (2009a).

een eigen thesaurus ontwikkeld, de Medical Subject Headings (MeSH).[116] Veel universiteiten maken van de beide indexeringen gebruik. Een samenwerkingsverband tussen de National Library of Canada en een aantal afgevaardigden van de LCSH resulteerde in een aanvullende set van Canadese Subject Headings (CSH), die gebruikt wordt voor typisch Canadese onderwerpen.[117]

De LC Subject Headings zijn gepubliceerd in vijf volumes maar kunnen uiteraard ook online geraadpleegd worden. De lijst met topics wordt wekelijks aangepast.[118]

De *subject headings* staan in voor uniforme toegangspunten van termen, namen en titels met betrekking tot het onderwerp of genre van een werk. Wie in een bepaald onderwerp geïnteresseerd is maar geen weet heeft van titels of auteurs, is zo in staat om informatie over gerelateerde onderwerpen te vinden. Wanneer een werk dus handelt over een persoon, organisatie, plaats of ander werk, dan kunnen zowel de namen van de betreffende persoon, organisatie of plaats als de uniforme titel als *subject heading* worden gekozen.

Er bestaat echter een duidelijk onderscheid tussen catalogiseren en indexeren. LCSH is opgesteld om het hoofdonderwerp van een bepaald werk aan te duiden. In het algemeen moet 20% van het werk te maken hebben met het onderwerp, vooraleer de uniforme naam van het onderwerp als *subject heading* voor dat werk kan gebruikt worden. In dit opzicht mag de toewijzing van *subject headings* niet verward worden met een gedetailleerde indexering. Die indexeringen vermelden soms een onderwerp dat slechts eenmaal in een bepaald werk voorkomt of vernoemd wordt. De *subject headings* gelden dus veeleer als een beknopte, gestandaardiseerde samenvatting van dat werk, zonder te veel in detail te treden. Niettemin moet gestreefd worden naar een zo groot mogelijke specificiteit in de toewijzing van *subject headings*. Elk onderwerp kan worden onderverdeeld in verdere subcategorieën. Over het gebruik van categorieën en subcategorieën bestaan enkele regels, die in het volgende voorbeeld toegelicht worden.

116 MeSH (2008).

117 CSH (2008).

118 Cf. LOC (2009a).

- Als drie of minder subcategorieën van een bepaald onderwerp besproken worden:

  - Als de subcategorieën samen het hele onderwerp beschrijven, dan wordt de naam van het onderwerp als subject heading toegewezen.

    Bijvoorbeeld: een boek over gewervelde en ongewervelde dieren. Dieren zijn ofwel gewerveld of ongewerveld dus wordt 'dieren' als *subject heading* toegewezen en niet de termen 'gewervelde dieren' en 'ongewervelde dieren'.

  - Als de subcategorieën samen slechts een deel van een bepaald onderwerp beschrijven, dan worden de subcategorieën als subject headings gebruikt.

    Bijvoorbeeld: een boek over muizen en ratten. Beide zijn knaagdieren maar er bestaan nog veel andere soorten knaagdieren. Als *subject headings* worden daarom twee termen meegegeven, namelijk 'muizen' en 'ratten' in plaats van 'knaagdieren'.

- Als drie of meer subcategorieën van een bepaald onderwerp worden besproken, dan wordt de naam van het onderwerp als *subject heading* meegegeven.

  Bijvoorbeeld: een boek over muizen, olifanten, beren en herten. Als subject heading wordt de term 'zoogdieren' gebruikt.

Vervolgens bestaan ook enkele richtlijnen voor het toegelaten aantal *subject headings*.

- Veel werken behandelen meer dan één onderwerp. Ieder onderwerp moet dan als *subject heading* aan dat werk worden toegewezen.

- Soms worden verschillende aspecten van hetzelfde onderwerp besproken, waardoor er meerdere *subject headings* nodig zijn om het werk goed te catalogiseren.

  Bijvoorbeeld: een werk over een bepaald economisch probleem op een bepaalde plaats. Een eerste subject heading duidt dan het probleem aan. Als het probleem een specifiek onderdeel kent dat gerelateerd is aan de plaats, dan wordt het lokale probleem als subject heading gekozen.

Een tweede subject heading kan de plaats zijn, waarvan een onderdeel 'Economic conditions' is.

- Een werk kan ook een onderwerp behandelen dat zich over meerdere niveaus laat beschrijven. Bijvoorbeeld: een algemene discussie over een concept wordt geïllustreerd met een case study, die van toepassing is in een bepaalde context. De *subject headings* moeten de verschillende niveaus reflecteren als minstens 20% van het werk aan ieder niveau gewijd is.

  Bijvoorbeeld: Women $z Nicarague; Women

Algemeen wordt aangenomen dat een boek adequaat kan beschreven worden door middel van zes *subject headings* en het gebruik van meer dan tien *subject headings* voor een bepaald werk moet vermeden worden. De volgorde van de *subject headings* komt bovendien overeen met de volgorde van belangrijkheid.

5.4.6.2 **Vorm**

Er wordt een onderscheid gemaakt tussen de hoofdheading en zijn subdivisie(s). Overeenkomstige MARC-velden worden tussen haken weergegeven.

- Hoofdheading
  - Onderwerp - Topic (MARC tag 650) (*topical heading*): een concreet object, een categorie van objecten, mensen of dieren, een abstract concept, geloof, proces of fenomeen, een instituut,...

    Een *topical heading* kan een enkele term of een zin zijn:

    - een term: waarschijnlijk de meest gangbare vorm.

      Bijvoorbeeld: Women, Savannas, Housing

    - een zin: hiervoor bestaan verschillende mogelijke constructies.
      - Directe volgorde:

        Bijvoorbeeld: Housing policy, Foreign exchange administration

- Omgekeerde volgorde: omgekeerde *headings* worden gescheiden door een komma. De meer significante term wordt eerst gegeven, gevolgd door een bepaling.

  Bijvoorbeeld: Authors, Mexican
  Farms, small

- Term en een bepalende term: de bepalende term staat tussen ronde haken en duidt de context van het onderwerp aan.

  Bijvoorbeeld: Stress (Physiology), Stress (Psychology)

- Een zin over een topic mag voorzetsels bevatten.

  Bijvoorbeeld: Violence in motion pictures

- Een zin mag gerelateerde termen bevatten die met elkaar verbonden zijn door het woord 'and'.

  Bijvoorbeeld: Banks and Banking, Cities and Towns

- Soms bestaat een zin uit termen die met elkaar verwant zijn. De zin bestaat dan uit de twee gerelateerde termen, en 'etc.'

  Bijvoorbeeld: Comic books, strips, etc.

- Een combinatie van de bovenvermelde structuren.

- Naam – van persoon, organisatie of conferentie (MARC tag 600, 610 of 611): de vorm van alle namen moet identiek zijn ongeacht de functie van het object dat een naam krijgt toegewezen. (bijvoorbeeld author, responsible body of subject).

- Uniforme titel (MARC tag 630): hiervoor geldt dezelfde regel als voor namen.

- Een geografische plaats (MARC tag 651): men onderscheidt twee categorieën:
  - Plaatsen die een jurisdictie hebben of hadden, zoals landen, steden of provincies. Zo'n plaats heeft een overheid die als corporate author beschouwd kan worden.

    Bijvoorbeeld: Argentinië, Buenos Aires (Argentinië)
  - Plaatsen zonder jurisdictie.

    Bijvoorbeeld: Olympus, Mount (Greece)
    Atlantic coast (Nicaragua)
- Subdivisies: *subject heading strings*

  Een *subject heading* kan bestaan uit een *string*, met een *heading* en één of meer subdivisies. Deze subdivisies worden met een bepaalde *string* gespecificeerd.

  Bijvoorbeeld: Farms, Small $z Colombia
  Camus, Albert, $d 1913-1960 $v Congresses
  Women $z Italy $x Social conditions

5.4.6.3 Enkele beperkingen

De lijst van LCSH is beperkt. Ze is immers gebaseerd op *headings* die werkelijk al gebruikt zijn om onderwerpen te beschrijven van de werken die in de Library of Congress gecatalogiseerd zijn. De keuze en vorm van een *heading* zijn niet noodzakelijk up-to-date. Wekelijks wordt de lijst bijgewerkt. Zo is 'Man' bijvoorbeeld verouderd en vervangen door 'Human being'. De vorm van de *subject headings* kan behoorlijk variëren. Terwijl vroeger vaak de omgekeerde volgorde werd gehanteerd voor *subject headings* die uit zinnen bestaan, wordt tegenwoordig vaak de directe volgorde gebruikt. Zo wordt de verouderde vorm 'Societies, Primitive' veranderd naar 'Primitive societies'.

5.4.7 GETTY Thesauri[119]

De thesauri van het Getty-instituut worden geproduceerd en onderhouden door de Getty Vocabulary Program. Ze volgen de ISO en NISO standaard voor

119 De volgende tekst is grotendeels gebaseerd op de GETTY-website Getty (2009a).

de constructie van thesauri. Ze bevatten termen, namen en andere informatie met betrekking tot mensen, plaatsen en concepten die gerelateerd zijn aan kunst, architectuur en materiële cultuur.

De Getty-thesauri kunnen op drie verschillende manieren worden gebruikt:

- voor de invoer van data en dus beschrijving van zaken,
- als kennisbank die informatie aan onderzoekers levert,
- als *search assistant* om de toegang voor de eindgebruiker tot online bronnen te vereenvoudigen.

De drie belangrijkste thesauri zijn:

Art & Architecture Thesaurus (AAT): een gestructureerde woordenschat die opgebouwd is rond 34.000 concepten, waaronder 131.000 termen, beschrijvingen, bibliografische citaten en andere informatie over beeldende kunsten, architectuur, decoratieve kunsten, archivalische zaken en materiële cultuur.

Union List of Artist Names (ULAN): bevat ongeveer 120.000 records, waaronder 293.000 namen en biografische en bibliografische informatie over artiesten en architecten, waaronder ook tal van naamsvariaties en pseudoniemen.

Getty Thesaurus of Geografic Names (TGN): bevat ongeveer 912.000 records, waaronder 1,1 miljoen namen, plaatstypes, coördinaten en beschrijvende nota`s die gericht zijn op belangrijke plaatsen voor de studie van kunst en architectuur.

Deze drie thesauri worden bondig toegelicht.

5.4.7.1 AAT

De thesaurus bestaat uit 34.000 concepten, gerelateerd aan kunst en architectuur. De concepten hebben betrekking op een tijdspanne van de oudheid tot nu. Ieder concept wordt geïdentificeerd aan de hand van een unieke numerieke ID. Aan elk concept zijn termen gekoppeld, gerelateerde concepten, een *parent* (voor de plaats binnen de hiërarchie), bronnen voor de data en nota's. Op die manier bevat de thesaurus 131.000 termen. Die termen worden gebruikt om kunst en architectuur te beschrijven. De termen van een concept bevatten de enkelvoudige vorm, de meervoudsvorm, de natuurlijke volgorde, de omgekeerde volgorde, spellingsvarianten, mogelijke uitspraken

en synoniemen. Van al die termen wordt één aangeduid als de geprefereerde term of de *descriptor*.

De AAT is een hiërarchische databank. De *root* van de hiërarchie wordt de *Top* van de AAT-hiërarchie genoemd. Naast de hiërarchische relaties kent de AAT ook equivalente en associatieve relaties. Het conceptuele kader van hiërarchieën en facetten is ontworpen om een algemene classificatie van kunst en architectuur te bekomen. Het kader is niet subject-specifiek. Dat wil zeggen dat er bijvoorbeeld geen specifieke termen voorhanden zijn voor de beschrijving van een schilderij uit de renaissance.

De 'facetten' vormen de belangrijkste subdivisies van de hiërarchische structuur van AAT. Een facet bevat een homogene klasse van concepten, waarvan de termen karakteristieken bevatten die hen onderscheiden van termen uit een andere klasse. Zo is marmer een substantie die wordt gebruikt voor de creatie van kunst en architectuur. Daarom wordt marmer in het facet over materialen teruggevonden. De facetten zijn conceptueel georganiseerd volgens een schema dat evolueert van abstracte concepten tot meer concrete, fysieke artefacten.

- Geassocieerde concepten: dit facet bevat abstracte concepten en fenomenen die betrekking hebben op de studie en uitvoering van een brede waaier aan menselijke ideeën en activiteiten, waaronder kunst en architectuur in alle mediavormen, maar ook gerelateerde disciplines. Dit facet betreft ook de theoretische en kritische overwegingen, ideologieën, houdingen en sociale en culturele stromingen. Bijvoorbeeld: schoonheid, balans, vrijheid, socialisme,...

- Fysieke attributen: dit facet bevat de perceptuele of meetbare eigenschappen van materialen en artefacten, maar ook eigenschappen van materialen en eigenschappen die niet als afzonderlijke componenten te onderscheiden zijn. Onder deze categorie bevinden zich eigenschappen zoals de grootte en vorm, chemische eigenschappen van materialen, kwaliteiten van textuur en hardheid en eigenschappen zoals oppervlakte, afwerking en kleur. Bijvoorbeeld: rond, broosheid, grenzen,...

- Stijlen en periodes: dit facet bevat algemeen aanvaarde termen om stilistische stromingen en periodes aan te duiden die relevant zijn voor de kunst en architectuur. Bijvoorbeeld: Frans, Louis XIV, Xia, Abstracte Expressionist,...

- Agenten: dit facet bevat termen om mensen, groepen van mensen en organisaties te benoemen die worden geïdentificeerd door een beroep of activiteit, door fysieke of mentale eigenschappen of door een sociale rol. Bijvoorbeeld: religieuze orden, corporaties, landschapsarchitecten,...

- Activiteiten: dit facet verzamelt alle domeinen van inspanningen, fysieke of mentale acties, systematische sequenties van acties, gebruikte methoden en processen bij bepaalde materialen of objecten. Activiteiten kunnen variëren van leertrajecten tot enkele gebeurtenissen, van mentale taken tot fysieke acties. Bijvoorbeeld: archeologie, ontwerpen, analyseren, tentoonstellingen, corrosie, tekenen,...

- Materialen: termen die een fysieke substantie aanduiden, van natuurlijke tot synthetische substanties. Dit facet kan variëren van specifieke materialen tot materiaaltypes die voor een bepaalde functie ontworpen zijn, zoals kleurstoffen, en van grondstoffen tot verwerkte producten, zoals ijzer, klei, plakband, emulgators,...

- Objecten: het grootste facet bevat termen voor alle fysieke of zichtbare objecten die levenloos zijn en het product van een menselijke activiteit. Ook eigenschappen van een landschap die de context leveren voor een bepaald bouwwerk maken hier deel van uit. Bijvoorbeeld: schilderijen, voorgevels, kathedralen, tuinen,...

Een voorbeeld van een record met de betreffende termen:

Figuur 20: AAT termen[120]

120 Voorbeeld is ontleend aan Getty (2008a).

Een verheldering van de gebruikte afkortingen of *flags* in het voorbeeld:

D = Descriptor

AD = Alternative Descriptor

UF = Use For term (een synoniem dat geen *descriptor* of *alternative descriptor* is)

C = Current term

H = Historical term

B = Both: current en historical

U = Unknown

NA = Not Applicable

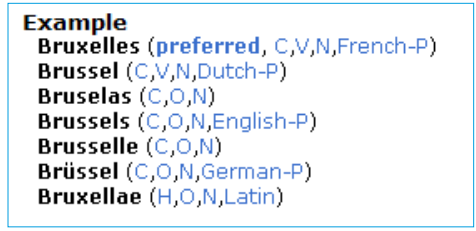
5.4.7.2 TGN

Deze thesaurus bestaat uit 912.000 records van plaatsen, met betrekking tot de prehistorie tot nu. De structuur van de thesaurus is heel gelijkaardig aan die van de AAT. Een record in deze databank duidt een plaats aan. Aan ieder record zijn de volgende zaken gerelateerd: een unieke numerieke ID, namen, de *parent* (de plaats in de hiërarchie van de TGN-thesaurus), geografische coördinaten, nota's, bronnen voor de data, andere relaties en een plaatstype. Dit plaatstype beschrijft de rol van de plaats. Bijvoorbeeld een bewoonde plaats, staat of hoofdstad van een staat. In totaal bevat de thesaurus 1.106.000 namen. Die namen worden zowel in de Engelse als in de lokale taal uitgedrukt, eventueel ook in andere talen en doorgaans wordt bovendien de historische naam meegegeven. Zoals eerder vermeld, bezit een record ook de coördinaten van de plaats. Die coördinaten zijn veeleer als referentie bedoeld en zijn slechts bij benadering juist.

Net als de AAT-thesaurus, is TGN hiërarchisch opgebouwd. De *root* van de hiërarchie is de *Top* van de *TGN*-hiërarchieën. Ze bestaat uit twee facetten: 'World' en 'Extraterrestrial Places'. Het spreekt voor zich dat vooral het facet 'World' veel termen bevat. De plaatsen zijn er gerangschikt volgens subdivisies die de huidige politieke en fysieke wereld representeren, hoewel er ook historische naties in besloten liggen. Behalve hiërarchische relaties kent de TGN ook equivalente en associatieve relaties. TGN is een thesaurus, die de ISO- en NISO-standaarden volgt.

Figuur 21: TGN-hiërarchie[121]

Een voorbeeld van de verschillende benamingen van Brussel, opgeslagen in een record van Brussel:



Figuur 22: TGN namen voor Brussel[122]

Een verheldering van de gebruikte afkortingen of *flags* in het voorbeeld:

Type *flag*:

N = Zelfstandig naamwoord
A = Bijvoeglijk naamwoord
B = Both

121 Voorbeeld is ontleend aan Getty (2004a).
122 Voorbeeld is ontleend aan Getty (2004a).

Historische *flag*:

C = Current
H = Historical
B = Both
U = Unknown
NA = Not Applicable

Lokale *flag*:

V = Vernacular
O = Other
U = Undetermined

5.4.7.3 ULAN

De ULAN is een thesaurus met records voor artiesten. Momenteel bevat de thesaurus ongeveer 120.000 namen van artiesten. De structuur is heel gelijkaardig aan die van de AAT- of TGN-thesaurus. Ieder record heeft een unieke numerieke Id, namen, gerelateerde artiesten, bronnen van de data en nota's. De records hebben betrekking op een periode van oudheid tot heden. In totaal zijn er ongeveer 293.000 namen in de thesaurus opgenomen. Dit zijn mogelijk gewone namen maar het kan ook gaan om pseudoniemen, spellingsvarianten van een naam, de naam in verschillende talen en namen die door de tijd heen gewijzigd zijn (bijvoorbeeld door huwelijk). Eén van de namen krijgt de *flag* 'preferred' mee.

Hoewel de structuur vrij vlak is, is de ULAN als een hiërarchische databank opgebouwd. De *root* van de databank is de *Top* van de *ULAN hierarchies*. De *root* kent twee vertakkingen of 'facetten': *Person* en *Corporate Body*. Ook die thesaurus voldoet aan de normen van ISO en NISO. Dat betekent dat er naast hiërarchische relaties ook equivalente en associatieve relaties tussen de verschillende records mogelijk zijn.

Een voorbeeld van de mogelijke benamingen van Le Corbusier:

Figuur 23: ULAN namen voor Le Corbusier[123]

5.4.8 RAMEAU[124]

5.4.8.1 Achtergrond en doelstelling

RAMEAU staat voor Répertoire d'autorité-matière, encyclopédique et alphabétique unifié. Het is een thesaurus die alle kennisgebieden bestrijkt in de vorm van een lijst van alfabetische instanties. De ontwikkeling van de thesaurus is gestart in 1980. Gelijktijdig, maar onafhankelijk, werd de *Directory of Subject Headings* van de Laval-universiteit van Quebec ontworpen, een vertaling van de *Library of Congress Subject Headings*. Vanaf 1983 werd RAMEAU ontwikkeld in samenwerking met het Franse ministerie voor Nationale Educatie (Le Ministère de l'Education Nationale) en met de publieke informatiebibliotheek (la Bibliothèque Publique d'Information, BPI) onder de naam LAMECH (Liste d'Autorité Matière, Encyclopédique, Collective et Hiérarchisée). In 1987 begonnen de twee instituten samen te werken voor het beheer en de verspreiding van de thesaurus, die vervolgens de naam RAMEAU kreeg. Aanvankelijk bevatte de thesaurus enkel data van de Nationale Bibliotheek maar hij werd naderhand verrijkt met data van de universiteitsbibliotheken. Ondertussen is de thesaurus een nationale indexeringstaal geworden en wordt hij gebruikt door publieke bibliotheken, een aantal onderzoeksbibliotheken en private organisaties.

123 Voorbeeld is ontleend aan Getty (2004b).

124 RAMEAU (2008).

5.4.8.2 **Vorm**

RAMEAU is een indexeringstaal die volgens drie niveaus gestructureerd is:

- Terminologisch niveau: gecontroleerde taal
- Semantisch niveau: hiërarchische taal
- Syntaxis-niveau: geprecoördineerde taal

RAMEAU wordt een gecontroleerde taal genoemd omdat er een controle wordt uitgevoerd op de vorm van de woordenschat, de polyseme woorden en de synoniemen. Elk concept heeft een *vedette*, een geprefereerde term voor dat concept. De *vedette* moet aan een aantal voorwaarden voldoen.

- Er moet een onderscheid bestaan tussen woorden en uitdrukkingen. Dat is een gevolg van het feit dat de thesaurus precoördinatief is.

  Bijvoorbeeld: Travail; Conditions de travail

- Het Franse woord moet gebruikt worden, behalve bij afleidingen uit een andere taal.

  Bijvoorbeeld: Droit d'auteur (en niet Copyright); Westerns

- De meervoudsvorm moet gebruikt worden, mits enkele uitzonderingen (zoals voor abstracte termen).

  Bijvoorbeeld: Vêtements; Conscience

- De *vedette* is het meest gebruikte woord:

  Bijvoorbeeld: Bicyclettes (en niet Vélocipèdes)

Zoals vermeld, controleert RAMEAU ook polyseme woorden, woorden met dezelfde schrijfwijze maar met een verschillende betekenis. RAMEAU onderscheidt polyseme woorden van homoniemen, aangezien er per concept maar één *vedette* kan zijn. Dit wordt opgelost door:

- de toevoeging van een bepaald woord tussen haakjes.

  Bijvoorbeeld: Elasticité; Elasticité (économie politique)

- De toevoeging van een adjectief.

  Bijvoorbeeld: Analyse documentaire; Analyse mathématique

- Een onderscheid tussen de enkelvoudige en meervoudige vorm (verschillende betekenis).

  Bijvoorbeeld: Religion; Religions

RAMEAU controleert bovendien op synoniemen om parallelle indexaties te vermijden. Daarom werd trouwens de *vedette* ook ingevoerd. Andere termen die naar hetzelfde concept verwijzen, worden als *terme exclu (TE)* aangeduid, die naar de *vedette* verwijzen.

Bijvoorbeeld: Bicyclettes EP Vélocipèdes

Bijvoorbeeld: Vélocipèdes VOIR Bicyclettes

(Bicyclettes = vedette; Vélocipèdes = TE)

De uitgesloten termen (TE) kunnen synoniemen of quasi-synoniemen zijn, maar ook afkortingen of acroniemen, of ze hebben een anders geconstrueerde syntaxis.

Bijvoorbeeld: Nantes, Edit de (1598) VOIR Edit de Nantes (1598)

Bijvoorbeeld: Religion – Histoire VOIR Histoire religieuse

RAMEAU is ook een hiërarchische taal op semantisch niveau. Dat wil zeggen dat de betekenis van de *vedette* nog verder gepreciseerd wordt door zijn semantische relaties.

- Hiërarchische relaties TG/TS:

  De generische term wordt aangeduid met TG, de meer specifieke term met TS. Dat wordt ook aangeduid als de verticale categorisatie.

  Bijvoorbeeld: Art TS Architecture

  Bijvoorbeeld: Architecture TG Art

- Associatieve relaties TA/TA: die zorgen voor de horizontale categorisatie.

  Bijvoorbeeld: Architecture TA Construction

  Bijvoorbeeld: Construction TA Architecture

Op het niveau van de syntaxis is RAMEAU precoördinatief. Complexe onderwerpen, die bijvoorbeeld uit verschillende *vedettes* samengesteld zijn, worden zo ook afzonderlijke *vedettes*. RAMEAU leidt uit een complex onderwerp niet automatisch de *vedettes* af, maar construeert een nieuwe *vedette* op het moment van de indexering. Een complex onderwerp bestaat uit een hoofdonderwerp (TV – *tête de vedette*) en subdivisies. Die laatste zijn elementen die achter de TV kunnen staan om het onderwerp meer te preciseren of te vervolledigen.

Bijvoorbeeld: Tourisme rural – France = TV + Subdivisie

De volgorde is hierbij heel belangrijk. Foutief zou zijn: France – Tourisme rural.

5.4.9 Thesaurus architecture et patrimoine[125]

Deze thesaurus betreft een systematische lijst van Franse architecturale werken en Frans erfgoed. Ze bevat 1135 termen over architecturale werken en 2529 termen over erfgoed. De termen over architecturale werken zijn ontleend aan de indexaties van de databank van Mérimée, termen over erfgoed komen voort uit de indexaties van de databank over Frans erfgoed in Palissy. Het erfgoed omvat architecturale elementen, glasramen, meubels, muziekinstrumenten, wetenschappelijke instrumenten, industriële machines en boten.

De lijst is hiërarchisch opgevat. De termen worden eerst volgens functionele categorieën opgedeeld en worden steeds verder verfijnd. Functionele categorieën zijn bijvoorbeeld religieus gebruik, begrafenisgebruik of industrieel gebruik. Zo bestaan er achttien basiscategorieën, die disjunct zijn. Geen enkele term komt bijgevolg in meerdere categorieën voor. Indien de semantiek van de term met andere categorieën overlapt, dan wordt de term in de categorie geplaatst die de voorkeur geniet en in de andere categorieën wordt

125 Architecture & Patrimoine (2009).

de verwijzing ‘voir aussi’ weergegeven. De categorieën bevatten bovendien alle nodige verwijzingen, definities en aantekeningen bij het gebruik.

De thesaurus kent een Amerikaanse en een Engelse versie en er bestaat tevens een Italiaanse vertaling. Op die manier wordt het internationaal, elektronisch raadplegen van de thesaurus bevorderd. Er bestaat echter geen xml-versie van de thesaurus, wat het gebruik ervan nochtans zou vergemakkelijken.

6 Declaratieve containers

6.1 Inleiding

Dit hoofdstuk behandelt samengestelde informatieobjecten en relaties tussen data en metadata. Het betreft objecten die beschrijvende, administratieve en/of structurele metadata combineren tot een informatieobject. Het voordeel van dergelijke objecten is de mogelijkheid ze uit te wisselen en te hergebruiken.

METS (§6.2) is een duidelijk voorbeeld van een standaard die zowel beschrijvende metadata als administratieve en structurele metadata combineert tot een object. Een object binnen METS bevat een beschrijvende metadatasectie (dmdSec), een administratieve metadatasectie (admSec), een sectie die aangeeft welke bronnen tot het object behoren (fileSec), een sectie die de hiërarchische structuur van het digitale object weergeeft (structMap), een sectie die zorgt voor de weergave van hyperlinks tussen de verschillende componenten van een METS-structuur die beschreven zijn in de structMap (structLink). Ten slotte is er een minder gebruikte sectie die de middelen aanlevert om digitale objecten te verbinden met toepassingen of programmacodes, die in combinatie met andere informatie in het METS-document worden gebruikt voor het renderen of weergeven van het digitaal object (behaviorSec). Die secties samen beschrijven een METS-object, dat op die manier kan worden uitgewisseld.

Een ander model voor digitale objecten is LOM (§6.3), die zich meer specifiek toelegt op de beschrijving van leerobjecten. Leerobjecten kunnen zo gemakkelijker hergebruikt worden en het achterhalen van die leerobjecten wordt enorm bevorderd. Het LOM-datamodel specificeert welke aspecten van het leerobject beschreven moeten worden en welke woordenschatten hiervoor in aanmerking komen. Verder schrijft de specificatie ook voor hoe dit model verder uitgebreid kan worden. Het datamodel heeft al ruime ingang gevonden en wordt reeds ondersteund door enkele API's.

ORE (§6.4) is een model voor de beschrijving van aggregaties. Dit zijn informatie-eenheden die bij een samenstelling een logisch geheel vormen. Die

informatie-eenheden kunnen op hun beurt een aggregatie vormen. Een voorbeeld hiervan is een boek dat een aggregatie van hoofdstukken is. De hoofdstukken vormen op hun beurt een aggregatie van pagina's. Het doel van dit model is het promoten van het hergebruik van de samengestelde objecten. De vorming van aggregaties gebeurt via een *resource map*, een gecodeerde beschrijving van de aggregatie. De *resource map* (ReM) beschrijft de aggregatie als een set van bronnen, en mogelijk ook de types en de relaties tussen de bronnen. Door zowel aan de aggregatie als aan de *resource map* die de aggregatie beschrijft, een URI toe te kennen, worden dit gewone webbronnen die men kan uitwisselen.

Ten slotte wordt MPEG-21 DIDL (§6.5) besproken. Zoals in §4.3.4 al werd aangehaald, worden in het MPEG-21 kader complexe, digitale objecten beschreven in de Digital Item Declaration (DID), met behulp van de Digital Item Declaration Language (DIDL). DIDL introduceert een set abstracte concepten die samen een datamodel voor complexe digitale objecten vormen. Het DIDL-datamodel herkent de volgende entiteiten: een container als een groep van containers of items, een item als een groep van items of componenten, een component als een groep van bronnen, een bron die een individuele datastream voorstelt en tot slot secundaire informatie over een container, item, component of bron. De DIDL-specificatie voorziet abstracte definities voor alle entiteiten en hun onderlinge relaties. Hoe de data tot een digitaal object gestructureerd worden, is afhankelijk van de implementatie. Zo kan een muziekalbum op verschillende manieren met DIDL beschreven worden. Iedere song kan een item zijn, maar het album kan ook worden voorgesteld als een enkel item met songs als de componenten. De representatie van een object zal uiteindelijk afhangen van de applicatie die men voor ogen heeft.

6.2 METS[126]

6.2.1 Achtergrond en doelstelling

De Metadata Encoding en Transmission Standard, kortweg METS, is een specificatie voor de beschrijving en uitwisseling van digitale objecten en hun eigenschappen. METS is een open standaard die door de bibliotheekgemeenschap ontworpen werd.[127]

126 Zie ook de tekst in Mannens, Paridaens, et al. (2007), p. 86-88.

127 LOC (2009b). Documentatie bij METS: METS Editorial Board (2007).

6.2.2 Vorm

METS is XML-gebaseerd en biedt de middelen om metadata op te slaan voor zowel het beheer als de uitwisseling van digitale objecten. Door de XML-basis kent METS een hiërarchische structuur, waardoor ze de hiërarchie van digitale objecten kan uitdrukken. Een METS-document wordt uit verschillende METS-elementen opgebouwd. Die elementen bestaan dan weer uit meerdere secties:

```
<mets>
    <dmdSec/>
    <amdSec/>
    <fileSec/>
    <structMap/>
    <structLink/>
    <behaviorSec/>
</mets>
```

De secties voorzien mogelijkheden voor de uitdrukking van verschillende types metadata (administratief, beschrijvend,...) en informatie.

De secties dmdSec (*Descriptive Metadata Section*) en amdSec (*Administrative Metadata Section*) dienen als een soort *wrappers* waarin elementen van andere schema's toegevoegd kunnen worden. Deze *wrappers* zorgen er bijgevolg voor dat METS uitbreidbaar en modulair is. Voor de inhoud van de *wrappers* biedt METS geen eigen woordenschat en syntaxis. Die worden verzorgd door de standaard die binnen de *wrappers* gebruikt worden. In de praktijk bestaan er al extensieschema's die van deze techniek gebruik maken, bijvoorbeeld voor Dublin Core en MARC/XML. De data in de *wrappers* hoeven niet strikt tekstueel te zijn. Ook binaire formaten zoals MARC21 kunnen hierin opgeslagen worden.

Een voorbeeld van een dmdSec en een amdSec:

```
<mets:dmdSec ID=”DMD1”>
    <mets:mdWrap MIMETYPE=”text/xml” MDTYPE=”MODS”>
        <mets:xmlData>
            <mods:mods version=”3.1”>
            <mods:titleInfo>
                <mods:title>Interview met een oudstrijder</mods:title>
            </mods:titleInfo>
            <mods:name type=”personal”>
                <mods:namePart>Jan De Smedt</mods:namePart>
            </mods:name>
            <mods:typeOfResource>audio</mods:typeOfResource>
            </mods:mods>
        </mets:xmlData>
    </mets:mdWrap>
</mets:dmdSec>

<mets:amdSec>
    <mets:rightsMD ID=”ADMRTS1”>
        <mets:mdWrap MDTYPE=”OTHER” OTHERMDTYPE=”METSRights”>
            <mets:xmlData>
                <rts:RightsDeclarationMD RIGHTSCATEGORY=”PUBLIC DOMAIN”>
                    <rts:Context CONTEXTCLASS=”GENERAL PUBLIC”>
                        <rts:Constraints CONSTRAINTTYPE=”RE-USE”>
                            <rts:ConstraintDescription>
                                Het verdelen en/of kopiëren van dit
                                object is enkel toegelaten mits
                                toestemming van de rechthebbenden.
                            </rts:ConstraintDescription>
                        </rts:Constraints>
                    </rts:Context>
                </rts:RightsDeclarationMD>
            </mets:xmlData>
        </mets:mdWrap>
    </mets:rightsMD>
</mets:amdSec>
```

Na dmdSec en amdSec volgt fileSec (*File Section*), waarin bijgehouden wordt welke bestanden tot het beschreven object behoren. Dit kan door de opslag van het digitale object zelf of door een link naar dit bestand.

```
<mets:fileSec>
    <mets:fileGrp USE="archive image">
        <mets:file ID="epio1m" MIMETYPE="audio/wav" ADMID="TECHWAV01">
            <mets:FLocat xlink:href=http://www.xxxx.com/01.wav
            DOCTYPE="URL"/>
        </mets:file>
    </mets:fileGrp>
</mets:fileSec>
```

Na de sectie fileSec volgt de sectie structMap, waarin de hiërarchische structuur van het digitale object wordt weergegeven. Dit laat toe om de opbouw van het digitale object weer te geven. De sectie structMap laat toe om meerdere hiërarchische structuren per object weer te geven. Zo kan men zowel een logische als een fysieke hiërarchie beschrijven. Een interview kan bijvoorbeeld in één bestand zijn opgeslagen (fysieke hiërarchie) maar meerdere 'onderwerpen' bevatten (logische hiërarchie). De weergave van de hiërarchie gebeurt aan de hand van divisies:

```
<mets:structMap TYPE="physical">
    <mets:div TYPE="book" LABEL="Het leven tijdens WOII" DMDID="DMD1">
        <mets:div TYPE="page" LABEL="Blank page"/>
        <mets:div TYPE="page" LABEL="Page i: Main title page"/>
        <mets:div TYPE="page" LABEL="Page ii: Blank page"/>
        <mets:div TYPE="page" LABEL="Page iii: Title page"/>
    </mets:div>
</mets:structMap>
```

Tot slot is er de sectie structLink. Die zorgt voor de weergave van hyperlinks tussen de verschillende componenten van een METS-structuur, beschreven in de structMap.

Een minder gebruikte sectie is de zogenaamde behaviorSec. Die voorziet METS van de middelen om digitale objecten te verbinden met toepassingen of programmacodes die in combinatie met andere informatie binnen het METS-document worden gebruikt voor het renderen of weergeven van het digitale object.

METS biedt ook verschillende profielen aan, die helpen bij de creatie van METS-documenten. Een profiel geeft een vrij detaillistische beschrijving van een klasse van METS-documenten. Voor elk profiel is een schema beschikbaar. Profielen leiden ook programmeurs bij hun creatie van software voor het gebruik en de verwerking van METS-documenten. Bovendien bevorderen ze ook de interoperabiliteit van digitale bibliotheken.

Een profiel bestaat uit een dertiental componenten, waaronder de titel, een abstract van de uitbreidingsschema's en een voorbeelddocument.

Voordelen

- Uitbreidbaar en modulair dankzij de *wrapper*-secties

Nadelen

- Mogelijke veiligheidsproblemen bij het invoegen van programmacodes in de sectie behaviorSec
- Kleine gebruikersgemeenschap

6.3 LOM[128]

6.3.1 Achtergrond en doelstelling

LOM, of Learning Objects Metadata Standard, is een IEEE-standaard die ontworpen is voor de beschrijving van zogenaamde leerobjecten. Dit kan multimedia content zijn, educatieve content, leerobjectieven, enzovoort. De standaard is ontworpen met het oog op een minimale set attributen die nodig zijn om de leerobjecten te beheren, te lokaliseren en te evalueren. De standaard ondersteunt onder andere ook *security*, *privacy* en evaluatie.[129]

6.3.2 Vorm

LOM definieert een basisschema voor de beschrijving van de hiërarchie van data-elementen voor leerobjecten. Op het hoogste niveau zijn er negen categorieën te onderscheiden.

128 Zie ook de tekst in Mannens, Paridaens, et al. (2007), p. 89-90.

129 IEEE (2002).

- *General*: met algemene informatie over het leerobject in zijn geheel.
- *Lifecycle*: met informatie over de geschiedenis en de huidige staat van een leerobject, samen met de factoren die het leerobject tijdens zijn evolutie hebben beïnvloed.
- *Meta-metadata*: met informatie over de metadata zelf.
- *Technical*: met informatie over de technische eisen en karakteristieken van het leerobject.
- *Educational*: met informatie over het educatieve en pedagogische karakter van het leerobject.
- *Rights*: met informatie over de intellectuele eigendomsrechten.
- *Relation*: maakt het mogelijk om de relatie tussen verschillende leerobjecten weer te geven.
- *Annotation*: commentaren over het educatieve gebruik van het leerobject en het tijdstip van en verantwoordelijke voor de toevoeging van deze commentaren.
- *Classification*: de beschrijving van het leerobject in relatie tot een specifiek classificatiesysteem.

LOM specificeert voor elk element een naam, een toelichting, een grootte, een voorbeeldwaarde, een datatype en nog enkele andere basisdetails. Een voorbeeld van een dergelijk element is 'Technical.Location'. Het gaat hier om een element 'Location' binnen het element 'Technical'. Dat element geeft informatie over de plaats van het leerobject, bijvoorbeeld een URL. Sommige elementen kennen een beperkte woordenschat, opgenomen in een lijst van toegelaten waarden. Andere waarden worden ook toegelaten maar ten koste van een lagere semantische interoperabiliteit. LOM laat verder ook de uitbreiding van data-elementen toe, maar die mogen geen LOM-elementen vervangen, omdat dit semantische interoperabiliteit in de weg zou staan. Een voorbeeld hiervan is de toevoeging van een element 'Naam', dat verward kan worden met het data-element 'General.Title'.

Voor LOM bestaan reeds bindingen naar RDF en XML.[130] Een LOM-element zou er in XML-vorm als volgt kunnen uitzien:

```
<lom xmlns="http://ltsc.ieee.org/xsd/LOMv1p0">
    <general>
        <title>
            <string xml:lang="nl">Interview met een oudstrijder</string>
        </title>
        <language>nl</language>
    </general>
    <technical>
        <location type="URI">http://www.interviews.org/oudstrijderx3242.mp3
        </location>
    </technical>
</lom>
```

Tot slot voorziet de LOM-standaard ook in een mapping naar Unqualified Dublin Core.

Voordelen

- Heel flexibel en uitbreidbaar
- Mapping mogelijk naar Dublin Core en binding met RDF
- Uitgebreid softwareaanbod[131]

Nadelen

- Geen voorzieningen voor ontologieën

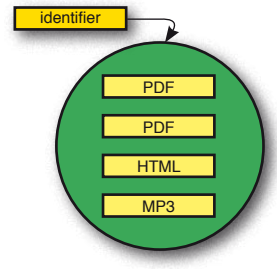
130 Nilsson, Palmér en Brase (2003) en IEEE LTSC (2007).

131 Friesen (2005).

6.4 ORE

6.4.1 Achtergrond

Samengestelde informatieobjecten, zogenaamde *compound objects*, zijn aggregaties van gescheiden informatie-eenheden die een logisch geheel vormen wanneer ze worden samengesteld. Een voorbeeld hiervan is een gedigitaliseerd boek, dat een aggregatie is van hoofdstukken die op hun beurt uit pagina's zijn opgebouwd. Een ander voorbeeld is een publicatie, die een aggregatie is van tekst en ondersteunend materiaal, zoals datasets, software tools, video-opnames van de experimenten,...



Figuur 24: Een samengesteld informatieobject[132]

Verschillende informatiesystemen, zoals *content management systems*, leveren ondersteuning voor de opslag en identificatie van en toegang tot de samengestelde objecten en hun geaggregeerde informatie. In de meeste systemen variëren de componenten volgens semantisch type (artikel, boek, video, dataset,...), mediatype (tekst, beeld, audio, video,...) en mediaformaat (PDF, XML, MP3,...) of kunnen de componenten op hun beurt weer een samengesteld object zijn. De componenten kunnen ook variëren volgens hun netwerklocatie. Sommige componenten van een samengesteld object kunnen lokaal opgeslagen zijn, andere bevinden zich op een ander netwerk.

De informatiesystemen verzorgen de opslag en identificatie en leveren de toegang tot deze samengestelde objecten op een architectuurspecifieke manier. Omdat het web tegenwoordig het uitgelezen platform is voor interoperabiliteit en webgebaseerde toepassingen – waarbij onder meer gerefereerd kan

132 Figuur is ontleend aan Lagoze en Van de Sompel (2007a), 'Figure 1 – A compound information object: identified aggregation of multiple components'.

worden aan de online *search engines* die nu de belangrijkste informatiebronnen zijn – zullen deze informatiesystemen hun objecten echter ook op het web presenteren. Dat kan door de associatie van een URI met elke component van het samengesteld object, zodat de bronnen door het web via de URI geïdentificeerd kunnen worden. Web services en toepassingen, zoals browsers en crawlers, kunnen deze URI's gebruiken om de gepaste representaties van de bronnen te verkrijgen.

Jammer genoeg is de manier waarop dergelijke informatiesystemen hun samengestelde objecten publiceren op het web niet perfect en ontbreekt er een algemeen aanvaarde standaard. In veel gevallen gaan bepaalde geavanceerde functionaliteiten verloren wanneer deze objecten op het web gepubliceerd worden. Meestal is de publicatie op het web gericht op de eindgebruikers en niet op *agents*, wat wel het geval is voor crawlers. De structuur van het object zit doorgaans vervat in onder meer *splash*-pagina's of *user interface widgets*. Die benadering maakt de essentiële structuur van het object onduidelijk voor machinegebaseerde applicaties zoals browsers en crawlers.

In het voorbeeld van het gescand boek waarvan aan alle pagina's een HTTP URI is toegewezen, kan een webcrawler bijvoorbeeld op één van die pagina's landen. De crawler kan vanuit die pagina links vinden naar andere pagina's van het boek, naar het hoofdstuk dat deze pagina bevat of naar het boek. Behalve die links kunnen er op de pagina ook links zijn naar bijvoorbeeld informatie over de auteur, de uitgever, annotaties,... Door het gebrek aan semantiek in de links slaagt een webcrawler er niet in een onderscheid tussen de links te maken. De links zijn met andere woorden niet getypeerd of als de links wel type-informatie bevatten, zijn die niet leesbaar voor webcrawlers. Door de afwezigheid van standaarden gaat de notie van een samengesteld object met een duidelijke grens en getypeerde relaties tussen zijn componenten vaak verloren.

Het gebrek aan standaarden tast de performantie van bestaande webservices en toepassingen aan. *Search engines* die gebaseerd zijn op crawlers kunnen bruikbaarder worden als de precisie en gelaagdheid van de resultaten correspondeert met samengestelde objecten in plaats van met de individuele bronnen. De rangschikking van de resultaten van de *search engines* kan worden verbeterd als de links naar de componenten van een object anders behandeld zouden worden dan de links die verwijzen naar de samengestelde objecten.

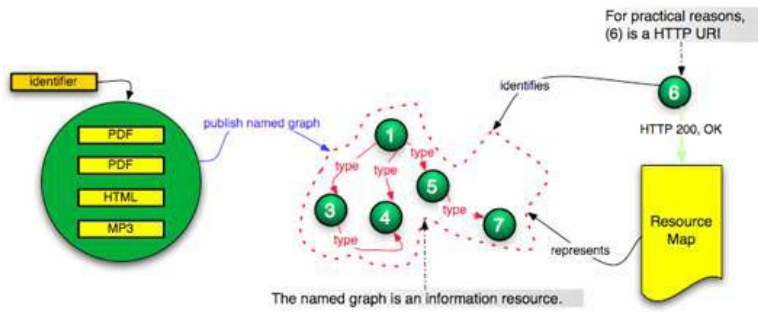
6.4.2 Doelstelling

Het doel van OAI-ORE (Object Reuse and Exchange)[133] is de ontwikkeling van een gestandaardiseerd, interoperabel en machineleesbaar mechanisme dat de informatie van de samengestelde objecten kan uitdrukken. De OAI-ORE-standaarden zorgen ervoor dat *web clients* en toepassingen de logische grenzen van de samengestelde objecten en hun relaties onderling kunnen reconstrueren. Dit betekent een toegevoegde waarde voor de ontwikkeling van diensten die instaan voor de analyse en de reconstructie van samengestelde objecten, zeker in het geval van *e-science* en *e-scholarship*, die de doelapplicaties van ORE vormen.

6.4.3 Vorm

ORE tracht een *interoperability layer* te realiseren, die een gestandaardiseerd middel moet worden om repository- en applicatie-specifieke implementaties van *compound objects* op het web te publiceren.

ORE moet dus in staat zijn om de grenzen van *compound objects* te identificeren. Dit kan door de publicatie van diagrammen (*graphs*) op het web die de relaties binnen de *compound objects* beschrijven. Elk gepubliceerd diagram wordt geïdentificeerd door een URI, zodat het een gewone webbron wordt. ORE doet dit via een *Resource Map*, een geëncodeerde beschrijving van het genoemde diagram (*named graph*).

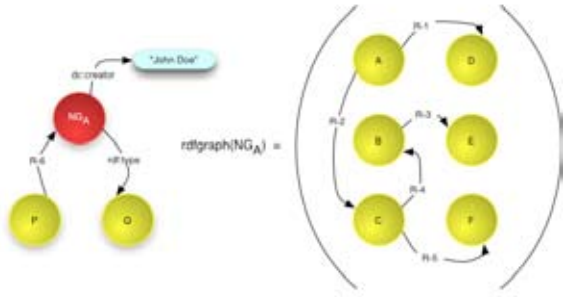


Figuur 25: ORE named graph[134]

133 OAI/ORE (2009).

134 Figuur is ontleend aan Lagoze en Van de Sompel (2007a), 'Figure 6: A named graph is published at a HTTP URI. A Resource Map is available through content negotiation with that HTTP URI'.

Een *named graph* is een uitbreiding van RDF, gebruikt om een naam te associëren met een set van triples – een zogenaamde *graph* of diagram. Een diagram bestaat uit verschillende aspecten. Een *named graph* is een bron, geïdentificeerd aan de hand van een URI. Deze URI kan zowel het subject als het object van triples zijn. Deze triples kunnen bijvoorbeeld het type van de *graph* aangeven of ze kunnen metadata associëren met de *graph*, zoals op het onderstaande diagram te zien is. De *named graph* is niet de RDF-*graph*. Het is een bron met een representatie die een set van triples encodeert. De relatie tussen een *named graph* en een RDF-*graph* die de representatie encodeert, is gedefinieerd via de functie *rdfgraph.*



Figuur 26: ORE rdfgraph[135]

De *Resource Map* (ReM) beschrijft een aggregatie die een set van bronnen vormt en mogelijk ook de types van en de relaties tussen de bronnen. De bronnen in een aggregatie worden daarom geaggregeerde bronnen genoemd.

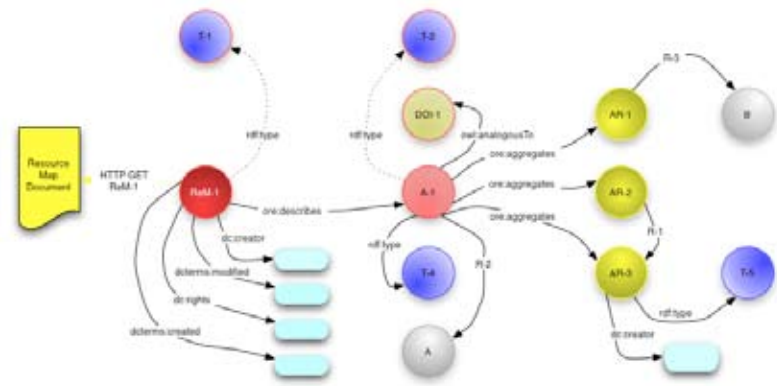
Een aggregatie op het web moet een URI (bv. A-1) hebben, indien ze via het web bereikbaar moet zijn. Het ORE-model maakt het noodzakelijk dat een *Resource Map* precies één aggregatie beschrijft. Een aggregatie kan wel meerdere *Resource Maps* bevatten. Opdat applicaties en *clients* zouden kunnen refereren naar de aggregatie is het noodzakelijk dat de URI A-1 naar de *Resource Map* leidt. Dat wordt op de volgende manieren gerealiseerd:

- de URI van de aggregatie kan geconstrueerd worden via een *fragment identifier aggregation*, die toegevoegd wordt aan de URI van de *Resource Map*.

135 Figuur is ontleend aan Lagoze en Van de Sompel (2007b).

- Indien de aggregaties in de infrastructuur maar één beschrijving kunnen hebben, bestaat er een andere mogelijkheid. Stel dat de aggregatie een URI http://sample.org/ReM-1 heeft, dan kan die verwijzen naar de *Resource Map* via de URI http://sample.org/ReM-1.xml of http://sample.org/ReM-1.rdf, afhankelijk van de serialisatie van de ReM.

Figuur 27 geeft een volledige representatie van een aggregatie.



Figuur 27: ORE grafische voorstelling van een aggregatie[136]

De ReM kan op verschillende manieren worden beschreven. Zo kan men de ReM in RDF/XML of Atom serialiseren.

Voordelen

- Aggregaties
- Bereikbaar voor webcrawlers
- Binding met RDF
- Hiërarchie

Nadelen

- Nog in evolutie

136 Figuur is ontleend aan Lagoze en Van de Sompel (2007b).

6.5 MPEG 21[137]

6.5.1 Achtergrond en doelstelling

Het MPEG-21 Multimedia Framework is een open raamwerk dat instaat voor de levering van multimediale data en de definitie van de consumptie ervan en dit voor alle spelers binnen de leverings- en verbruiksketen. MPEG-21 is gebaseerd op twee essentiële concepten: het digitale item, een fundamentele unit voor distributie en transactie, en het concept van de gebruikers, die interageren met deze digitale items. Samenvattend kan men zeggen dat het hoofddoel van MPEG-21 erin bestaat een technologie te definiëren die gebruikers helpt bij de toegang tot digitale items of de uitwisseling, het verbruik, de verhandeling of de manipulatie ervan.

De gebruiker is een entiteit die binnen de MPEG-21-omgeving interageert met een andere gebruiker of die gebruik maakt van een digitaal item. Gebruikers kunnen individuen, verbruikers, gemeenschappen, organisaties, bedrijven, consortia, regeringen, enzovoort zijn. Ze worden geïdentificeerd aan de hand van hun relatie tot een andere gebruiker voor een bepaalde interactie. Puur technisch maakt MPEG-21 geen onderscheid tussen verbruiker en *provider*, die daarom beiden als gebruikers worden beschouwd. Een gebruiker kan op verschillende manieren (publiceren, verbruiken,…) van *content* gebruik maken. Toch kan hij specifieke of zelfs unieke rechten en verantwoordelijkheden hebben, afhankelijk van de interactie met andere gebruikers binnen MPEG-21.

6.5.2 Vorm

Als basis biedt MPEG-21 een raamwerk waarin een gebruiker over een digitaal item interageert met een andere gebruiker. Een interactie kan onder andere de creatie, de archivering of het afleveren van content zijn. MPEG-21 bestaat ondertussen uit 18 delen die men in vijf categorieën kan indelen.

De eerste twee categorieën zijn *declaration* en *identification*. De DID (*Digital Item Declaration*) is een XML-schema waarin de *Digital Item Declaration Language* (DIDL) wordt gedefinieerd. Hierin wordt de structuur van complexe digitale objecten beschreven, waaronder de relatie tussen verschillende items. Vandaar dat die veeleer thuishoren bij de beschrijving van zogenaamde *compound objects*.

137 Zie de tekst in Mannens, Paridaens, et al. (2007), p. 95-96. Zie o.a. Bormans en Hill (2002); MPEG/ITEC (2005).

```
<dia:container>
    <dia:item>
        <dia:component>
            <dia:descriptor/>
            <dia:resource/>
        <dia:component>
    </dia:item>
</dia:container>
```

DII of *Digital Item Identification* neemt de identificatie van digitale items voor haar rekening. DII ondersteunt onder andere de unieke identificatie van digitale items en beschrijvende schema's en van verschillende types digitale items.

```
<Statement>
    <dii:Identifier>
        myID:1234
    </dii:Identifier>
</Statement>
```

Een derde categorie is DRM of het *Digital Rights Management*, dat tot doel heeft rechten en toelatingen weer te geven in een machineleesbare vorm. De uitdrukking van een recht bestaat uit vier entiteiten en hun wederzijdse relaties: de gebruiker aan wie de rechten zijn toegekend, de rechten zelf, het object waarop de rechten van toepassing zijn en de voorwaarden voor het uitoefenen van de rechten.

Om UMA (*Universal Multimedia Access*) mogelijk te maken, werd verder ook een set normatieve tools ontwikkeld: DIA (*Digital Item Adaptation*), die een vloeiende adaptatie van digitale items moet toelaten. MPEG-21 DIA specificeert aldus de syntaxis en de semantiek van mogelijke adaptaties. Die tools kunnen gebruikt worden om *resources* aan te passen naar gelang de (opgelegde) beperkingen in verband met transmissie, opslag, QoS en/of het afspelen van *resources*.

De laatste categorie, DIP (*Digital Item Processing*), heeft betrekking op de mogelijkheid om als eindgebruiker te interageren met een digitaal item. DIP specificeert daarvoor DIM (*Digital Item Method*), een methode gebaseerd op een variant van ECMAScript, die binnen een MPEG-21 *client*-applicatie toegepast kan worden.

7 Digitale archivering: Best Practices

Zoals eerder aangegeven (§3), kent het OAIS-model als conceptueel raamwerk wijdverbreide toepassingen in verschillende internationale preserveringsprojecten en digitale archiveringssystemen. In het bestek van deze state-of-the-art volstaat een selectief en bondig overzicht van enkele projecten. Vervolgens komen twee praktijkvoorbeelden uit Nederland meer in detail aan bod. Die projecten representeren bovendien de twee aspecten waarmee men volgens de OAIS-voorschriften rekening dient te houden. OAIS benadrukt namelijk het onderscheid tussen eisen voor langetermijnbewaring enerzijds en voor consultatie en hergebruik anderzijds. In de ontwikkeling van het e-Depot in de KB Den Haag staat langetermijnbewaring centraal en het MultiMatch-project, waar het instituut voor Beeld en Geluid aan deelneemt, concentreert zich in eerste instantie op consultatie en hergebruik.

CASPAR (Cultural Artistic and Scientific knowledge for Preservation Access and Retrieval) is een project dat mede gefinancierd wordt door de Europese Unie binnen het Sixth Framework Programme (Priority IST-2005-2.5.10, "Access to and preservation of cultural and scientific resources"). Het ging van start op 1 april 2006 en onderzoekt, implementeert en verspreidt innovatieve oplossingen voor digitale preservering, gebaseerd op het OAIS-referentiemodel.[138]

Het project Planets (Preservation and Long-term Access through Networked Services) situeert zich eveneens binnen het Sixth Framework Programme en loopt van 2006 tot 2010. Het belangrijkste doel van Planets is de ontwikkeling van diensten en tools die bijdragen tot de langetermijnbewaring van digitale culturele en wetenschappelijke objecten. Het Planets consortium wordt gecoördineerd door de British Library en bestaat verder uit een aantal Europese nationale bibliotheken, archieven, universiteiten en technologiebedrijven. Ook in dit project wordt het OAIS-model expliciet als basismodel aangehaald.[139]

Het National Digital Information Infrastructure and Preservation Program (NDIIP) wordt gecoördineerd door de Library of Congress in samenwerking

138 CASPAR (2006).

139 Farquhar en Hockx-Yu (2007).

met instellingen in diverse sectoren, zowel binnen als buiten de Verenigde Staten. Voor de beschrijving van de technische infrastructuur van NDIIP wordt grotendeels van OAIS uitgegaan.[140]

Andere voorbeelden van projecten die zich op het OAIS-model baseren:

- Pandora (Preserving and Accessing Networked Documentary Resources of Australia) van de National Library of Australia is een project voor webarchivering.[141]

- OCLC Digital Archive biedt tools voor bibliotheken en archieven met het oog op de archivering van webdocumenten. Daarbij worden enkele onderdelen van het OAIS-model geïmplementeerd, in het bijzonder *ingest, store, disseminate en administration*.[142]

- CEDARS (1998-2002) was een gezamenlijk project van de universiteiten van Oxford, Cambridge en Leeds, dat gericht was op langetermijnbewaring van digitale data. De 'metadata for digital preservation' die in dit project voorgeschreven worden, zijn gebaseerd op het OAIS-model.[143]

- AIHT (Archive Ingest and Handling Test) was een onderzoeksproject van de Library of Congress waarbij verschillende universiteitsbibliotheken betrokken waren. Bij het onderzoek naar de efficiëntie van archiefsystemen werd uitgegaan van OAIS-concepten.[144]

- PROV (Public Record Office Victoria) is een stadsarchief in Australië dat zich baseert op concepten van OAIS.[145]

140 Gladney (2007).
141 Cathro en Boston (2003).
142 Houser (2004) en OCLC (2008) (de website van OCLC Digital Archive).
143 CEDARS (2000).
144 Nelson, Bollen, et al. (2005).
145 PROV (2005).

7.1 Ontwikkeling van het e-Depot in de Koninklijke Bibliotheek van Den Haag

7.1.1 Voorgeschiedenis

De Koninklijke Bibliotheek van Nederland (KB) werd in 1798 opgericht als de Nederlandse nationale bibliotheek. Sinds 1974 is de KB ook een depotbibliotheek. In tegenstelling tot andere nationale bibliotheken is de depotfunctie van de KB niet verplicht: uitgevers mogen zelf bepalen of zij publicaties aan de KB schenken. In 1993 besliste de KB om haar depot uit te breiden met digitale informatie. Zo ontstond de nood aan een systeem om digitale publicaties op te slaan en ze op lange termijn te bewaren.

De KB heeft op het vlak van langetermijnbewaring een pioniersrol vervuld, waarbij ze onder meer talrijke onderzoeksprojecten begeleid heeft. Van 1998 tot 2000 liep onder leiding van de KB het NEDLIB-project (Networked European Deposit Library), waarin een werkgroep van acht Europese nationale bibliotheken en enkele uitgevers de vereisten van een Europees depotsysteem met betrekking tot langetermijnarchivering onderzocht. Het OAIS-model, op dat moment een ISO-standaard in wording, werd in het kader van dit project grondig geanalyseerd en verder uitgewerkt voor bibliotheken en archieven.[146]

Het NEDLIB-project is van grote betekenis geweest voor het internationale onderzoek naar digitale archivering. De belangrijkste conclusies van het project luidden dat het OAIS-model niet alleen een goed model is voor de archivering van ruimtevaartdata, maar ook een goede basis vormt voor de opzet van digitale archieven in bibliotheken en archieven. In navolging van de OAIS-voorschriften concludeerden de NEDLIB-partners dat men de functionaliteit voor archivering gescheiden moet houden van de functionaliteit voor zoeken, authenticatie en autorisatie.

Om een e-Depot op te zetten volgens de richtlijnen van het NEDLIB-project en het OAIS-referentiemodel bestonden er geen onmiddellijke 'out of the box'-oplossingen. Door middel van een Europese aanbesteding zocht de KB naar

146 Cf. o.a. Van der Werf-Davelaar (1999) (verantwoording van de implementatie van OAIS in het Nedlib-project) en Lupovici en Masanès (2000).

een externe technische partner om een elektronisch depot te installeren. In september 2000 tekenden KB en IBM het contract waarmee het DNEP-project (Depot voor Nederlandse Elektronische Publicaties) werd ingezet. Het DNEP-project bestond uit twee delen. Ten eerste werd nagedacht over de ontwikkeling en implementatie van een groots opgezet digitaal archief, het e-Depot. Ten tweede omvatte het DNEP-project ook een grondige studie van de noodzakelijke aspecten voor langetermijnbewaring, aangezien de kennis hierover zich op dat moment nog in een experimentele fase bevond en er geen definities voorhanden waren van de functionele vereisten voor een duurzame bewaring van digitale objecten.[147]

De ontwikkeling van het e-Depot nam twee jaar in beslag en werd op 12 december 2002, samen met het IBM-systeem DIAS (Digital Information and Archival System), als de technische kern van het systeem voorgesteld. Het e-Depot is ingericht voor de bewaring van Nederlandse elektronische publicaties. Daarnaast zal het plaats bieden aan het Nederlands webarchief en masterfiles van gedigitaliseerd materiaal. Omdat informatievoorziening tegenwoordig een mondiale aangelegenheid is, heeft de KB het e-Depot ook opengesteld voor internationale uitgevers als een 'safe space' of 'last resort' voor digitale publicaties.

Ondertussen heeft de KB als een 'safe place' archiveringsovereenkomsten met een groot aantal wetenschappelijke uitgevers, waaronder Elsevier, Kluwer, Biomed Central, Blackwell, Oxford University Press, Springer, Sage Publications en Taylor & Francis.[148]

Na de implementatie van het e-Depot in de KB heeft IBM onlangs nog samengewerkt met de Deutsche Nationalbibliothek voor de implementatie van het digitaal archief 'Kooperativer Aufbau eines Langzeitsarchivs digitaler Informationen' (Kopal), dat gebaseerd is op dezelfde technologie als in de KB, met name DIAS.[149]

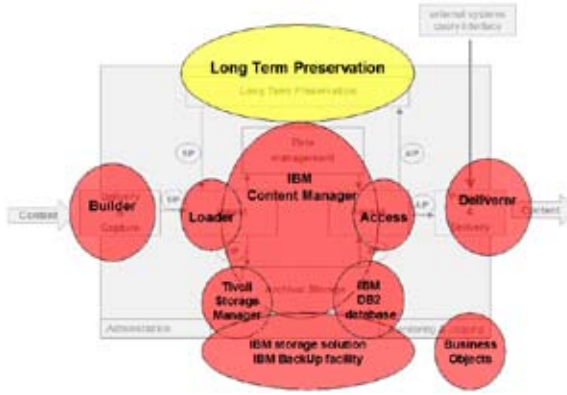
147 De resultaten van deze KB/IBM Long-Term Preservation (LTP) Study kan men nalezen in KB (2002).

148 Steenbakkers (2005).

149 Wollschlaeger (2006).

7.1.2 DIAS-architectuur

De hoofdcomponent van het e-Depot is DIAS (Digital Information Archiving System). DIAS was de eerste concrete realisatie van het OAIS-referentiemodel.[150]



Figuur 28: DIAS configuratie met OAIS[151]

In figuur 28 wordt duidelijk hoe de componenten van het DIAS-systeem, dat erg modulair opgebouwd is, met de functionele entiteiten van het OAIS-model samenvallen (cf. §3):

- *Ingest*: de ontvangst en bewerking van de SIP's van de uitgevers
- *Archival storage*: de opslag, bewaring en *retrieval* van AIP's
- *Data management*: descriptieve informatie en administratieve data
- *Administration*: administratie van het dagelijks beheer
- *Preservation planning*: plannen beheren en preserveringsstrategieën en –acties uitvoeren
- *Access*: toegang tot AIP's en productie van DIP's voor de aflevering aan de *designated community*
- *Monitoring en logging*: registratie en melding van acties.

150 Steenbakkers (2005).

151 Figuur is ontleend aan Steenbakkers (2004), 'Figure 3. DIAS Configuration and OAIS'.

In 2003 startte de KB samen met IBM een project voor de ontwikkeling van het 'preservation subsystem' voor DIAS (het gele ovaal in de figuur). Hiervoor baseerde men zich op de bevindingen van de onderzoeksrapporten van de LTP-studies in het kader van het DNEP-project over een aantal belangrijke aspecten van digitale bewaring, authenticiteit, mediamigratiemanagement, archivering van webpublicaties, e.a.

Volgens de onderzoekers houden langetermijnbewaring en ontsluiting van digitale objecten drie stappen in:[152]

- Archiveren: het toekennen van een *identifier*, het digitale object onderbrengen in een gecontroleerde OAIS-compliant-archiefomgeving en de toevoeging van technische en administratieve metadata. De archiefomgeving moet een aparte eenheid in de ICT-infrastructuur zijn. Het is belangrijk dat alle andere functionaliteiten die niet met archivering te maken hebben, zoals zoeken, authenticatie, e.a., apart gehouden worden. Dit verzekert dat de archiefomgeving een duurzaam systeem is dat onafhankelijk van de rest van de ICT-infrastructuur verder ontwikkeld kan worden.

- De *bitstream* bewaren: om de *bitstream* in de originele structuur te bewaren, moeten proactief een aantal stappen ondernomen worden. De *bitstream* moet regelmatig gekopieerd worden en het medium waarop de *bitstream* bewaard wordt, moet tijdig 'refreshed' worden.

- Toegang tot de digitale objecten, ook op lange termijn, garanderen: om toegang op lange termijn te kunnen garanderen, moeten volgens de DNEP-onderzoeksrapporten de software en de hardware die nodig zijn om het object te kunnen 'afspelen' mee bewaard worden.

Bij de bewaring van digitale objecten op lange termijn moet men ten minste met drie aspecten rekening houden: bewaring van het medium, technologiepreservering en intellectuele preservering.

- Mediumpreservering heeft te maken met het onderhoud van de dragers waarop informatie opgeslagen is (tapes, diskettes, CD-ROMS). Elke drager heeft een beperkte levensduur. Om te verzekeren dat de data op deze dragers niet verloren gaan, worden de volgende technologische oplossingen voorgesteld: backups, 'refreshing' en checksums om fouten in de *bitstream* te ontdekken en te verbeteren.

152 Van Diessen en Steenbakkers (2002).

- Snelle veranderingen in de technologie, vooral in de bestandsformaten en de software om elektronische informatie te 'renderen', vormen een grotere uitdaging dan mediumpreservering. Om de opgeslagen informatie ook binnen 50 of 100 jaar te kunnen bekijken of afspelen bestaan er twee oplossingen: migratie en emulatie. Wanneer de omvang van het archief na verloop van tijd toeneemt tot verschillende terrabytes aan informatie, zijn dergelijke migratieoperaties geen triviale taak meer. Een migratie van alle objecten in een verouderd bestandsformaat kan dan zodanig lang duren dat de objecten tijdelijk niet beschikbaar zijn.

- Intellectuele preservering gaat over de integriteit en de authenticiteit van informatie zoals die origineel werd opgeslagen. Authenticiteit betekent dat men het digitale object kan lezen zoals het oorspronkelijk werd opgeslagen. In een digitale omgeving kunnen objecten heel eenvoudig gewijzigd worden. Het is belangrijk dat er technieken of maatregelen ontwikkeld worden die wijzigingen in de *bitstream* ontdekken en kunnen verhinderen. De weg van ontstaan tot de huidige toestand van het digitaal object moet ook bijgehouden worden.

Om de authenticiteit van de opgeslagen informatie te bewaren, worden vijf interpretatieniveaus onderscheiden. Elk niveau moet worden beschreven om de authenticiteit van een digitaal object te waarborgen.

- *Binary interpretation schemata*: bepalen hoe fysieke karakteristieken van de hardware (elektrisch, magnetisch, optisch) naar bits vertaald worden en verder gesegmenteerd worden in eenheden van specifieke bitlengte.

- *Content interpretation schemata*: beschrijven specifieke bestandsformaten, die de bits vertalen naar menselijk bruikbare concepten: ASCII, bitmap,…

- *Content metatadata interpretation schemata*: beschrijven hoe bijkomende informatie en karakteristieken geassocieerd worden met bepaalde content data-elementen. In het geval van ASCII zijn dit bijvoorbeeld codes voor vette tekst, onderstrepen,...

- *Structure interpretation schemata*: beschrijven de relaties tussen de verschillende componenten. Zij bepalen hoe verschillende elementen samengevoegd kunnen worden tot een geaggregeerde eenheid.

- *Functional interpretation schemata*: omvatten de applicatielogica die gebruikt wordt voor het creëren, wijzigen, verkrijgen, verwijderen en renderen van het digitaal object op een specifieke IT-infrastructuur.

Ieder digitaal archief moet bepalen hoe deze authenticiteit bewaard wordt:

- Beschrijven van de *binary schemata* die binnen de IT-infrastructuur gebruikt worden. Bijvoorbeeld 32-bit, 64-bit, bigEndian, lowEndian.

- Beschrijven van de *content schemata* voor elk objecttype. Bijvoorbeeld JPEG2000v1.2, PDF-1992/A.

- Beschrijven van de *content metadata schemata*. Bijvoorbeeld Times-New-Roman.

- Beschrijven van de *structure schemata* voor elk object. Bijvoorbeeld hoofdstukken, afleveringen, fragmenten.

- Beschrijven van de *functional schemata* en hun impact op de digitale objecten. Bijvoorbeeld om een TIFF-beeld te tonen in een webbrowser moet deze eerst naar JPG geconverteerd worden.

De objectieven en de mogelijkheden van het digitaal archief bepalen welke interpretatieschema's bewaard en geëvalueerd moeten worden.

Op basis van de LTP-studies hebben IBM en de KB een *preservation subsystem* voor het e-Depot ontwikkeld waarin technische metadata geregistreerd worden en waar functies werden aan toegevoegd die nodig zijn voor langdurige bewaring. Dit subsysteem bestaat uit drie componenten: de *preservation manager* voor de registratie van technische metadata, de *permanent access toolbox* (PATbox) en de *preservation processor* voor de uitvoer van de preservatieacties zoals migratie en emulatie.[153]

In de *preservation manager* wordt informatie geregistreerd over de in het e-Depot opgeslagen bestandsformaten. Dit wordt als een essentieel onderdeel van DIAS beschouwd, omdat door middel van technische metadata ook toekomstige hardwaresystemen de *bitstream* van de software en van het digitale object kunnen lezen. Zo wordt aan de gebruikers toegang gegarandeerd.

153 Steenbakkers (2005).

Dat gebeurt volgens een structuur die van *preservation layer models* (PLM) en *view paths* gebruik maakt. Een *preservation layer model* beschrijft de verschillende 'lagen' waarop de software draait. Zo kan een PLM bestaan uit de volgende lagen: dataformaat, viewerapplicatie, besturingssysteem en hardwareplatform, waarbij vervolgens deze lagen nader gespecificeerd worden.

Een *view path* voor het formaat Pyramid TIFF-formaat bijvoorbeeld kan de volgende elementen bevatten: het platform is Intel Pentium, het besturingssysteem is Windows Vista en de applicatie is Photoshop 7. Telkens wanneer één van deze elementen in onbruik raakt, moet migratie of emulatie overwogen worden.

7.1.3 De gegevensarchitectuur van het e-Depot

In het kader van WP3 van BOM-Vlaanderen is vooral de gegevensarchitectuur van het e-Depot interessant.

Uitgevers bezorgen de elektronische publicatie als een informatiepakket, een SIP, die de volgende onderdelen bevat: *essence*, *metadata*, *table of contents* en, eventueel, *rendering software*. De *content bitstream* wordt in de archiefomgeving opgeslagen. Om duplicatie van metadata te vermijden, werd besloten de descriptieve metadata in de KB-catalogus in een eigen formaat op te slaan. Via de KB-catalogus kunnen eindgebruikers het e-Depot raadplegen.

De originele descriptieve metadata en een beperkte set van technische metadata (bestandsformaat, versie van bestandsformaat, grootte van het object,…) zoals die geleverd worden door de uitgevers, worden opgeslagen in de AIP. Specifieke preserveringsmetadata worden in de *preservation manager* opgeslagen.

Nieuwe ontwikkelingen, zowel in de KB als in de digitale bibliotheekwereld, hebben de KB ertoe aangezet na te denken over een vernieuwde gegevensarchitectuur en dit niet enkel voor het e-Depot maar voor alle databanken en catalogi in het beheer van de KB.

Eén van de uitgangspunten voor de nieuwe KB-infrastructuur was het vereiste dat de metadata van alle KB-bronnen, inclusief van het e-Depot, integraal doorzoekbaar zouden zijn zonder voorkennis van die metadata. Deze integrale doorzoekbaarheid vereist een gemeenschappelijk datamodel voor alle metadata, terwijl in de KB-structuur gebruik gemaakt wordt van verschillende datamodellen voor verschillende databases. De databases en catalogi die de KB beheert, worden via verschillende websites aangeboden en hebben

vaak een eigen metadataformaat, meestal toegespitst op een specifiek materiaaltype of een specifieke functionaliteit. Door het bijzonder karakter van de verschillende metadatastandaarden (bijvoorbeeld MARC, ISAD,…) en de beschreven objecten, is het moeilijk om één van deze standaarden als gemeenschappelijk datamodel te nemen.[154]

In plaats van 'proprietary' formaten wilde men gebruik maken van internationale standaarden voor bibliografische en structurele metadata. Specifiek voor langetermijnbewaring in het e-Depot onderzocht men de relevantie van de PREMIS-*data dictionary* voor preserveringsmetadata. Bovendien zouden naast wetenschappelijke publicaties ook andere materialen opgenomen worden in het e-Depot, namelijk websites en gedigitaliseerd cultureel en wetenschappelijk erfgoed. Deze ontwikkelingen hebben een belangrijke impact op het e-Depot. Zo zal de diversiteit van formaten toenemen en de structuur van de objecten zal complexer worden. Daarom volstond het oude metadatamodel niet meer.[155]

Dublin Core (DC) was volgens de KB het enige formaat dat 'generiek genoeg is om gebruikt te worden voor materiaalbeschrijvingen in de diverse sectoren van cultuur en wetenschap én dat binnen die sectoren voldoende geaccepteerd is'.[156] Om bij de mapping van de specifieke metadataformaten naar DC niet te veel informatie te verliezen, werd gekozen voor Qualified Dublin Core (DC(X)). Om samengestelde objecten op te slaan en te beschrijven en de locatie van de subobjecten vast te leggen, werd gekozen voor MPEG 21/DIDL als containerformaat.

De metadata die met de digitale objecten worden aangeleverd, maken integraal deel uit van het digitale object en worden dan ook samen in de langetermijnopslag bewaard. Om het publiek toegang te bieden tot de opgeslagen objecten worden de uitgeversmetadata omgezet naar DC(X) als primair formaat voor descriptieve metadata en ondergebracht in de vernieuwde KB-gegevensinfrastructuur. Via de KB-portal zijn deze metadata doorzoekbaar en kunnen de objecten opgevraagd worden. Aan de beschrijvingen van de objecten worden technische metadata toegevoegd om het gebruikte bestandsformaat eenduidiger te kunnen karakteriseren en om voldoende gegevens voorhanden te hebben om het object op ieder moment door middel van emulatie of migratie te kunnen hergebruiken.

154 Doorenbosch en Van Veen (2007).

155 Sierman (2009).

156 Doorenbosch en Van Veen (2009).

De metadatarecords bestaan uit één of meerdere blokken met verschillende datamodellen:

- altijd één DC(X)-blok voor integrale doorzoekbaarheid,
- optioneel een ander standaardformaat (MARCXML, EAD,…) om de rijkere originele metadata niet te verliezen
- en optioneel nog andere metadatablokken (bijvoorbeeld PREMIS voor preserveringsdoeleinden).

7.2 Instituut voor beeld en geluid: Multimatch

Het Nederlands Instituut voor Beeld en Geluid (B&G), opgericht in 1997, verzamelt, conserveert en geeft toegang tot het audiovisueel erfgoed dat uit historisch of cultuurhistorisch oogpunt van nationaal belang is. Daarnaast ontwikkelt en verspreidt het instituut kennis op het gebied van audiovisuele archivering, digitalisering en mediageschiedenis. B&G brengt verschillende voormalige archieven, zoals het RTV-Archief Publieke Omroepen, het Filmarchief van de Rijksvoorlichtingsdienst, het Omroepmuseum, Film en Wetenschap, Smalfilmmuseum en verschillende particuliere collecties, samen.[157]

De collecties omvatten meer dan 700.000 uur aan radio, televisie, film en muziek (waarvan een beperkt deel gedigitaliseerd), 2 miljoen foto's en 20.000 voorwerpen. Daarmee is Beeld en Geluid een van de grootste audiovisuele archieven van Europa.

De doelstelling van B&G is vierledig:

- het bedrijfsarchief van de Nederlandse omroepen zijn,
- het audiovisueel cultureel erfgoed van Nederland bewaren, beheren en ontsluiten,

157 Beeld en Geluid Instituut (2009).

- dit erfgoed ontsluiten voor het grote publiek via een nieuwe 'Media Experience' (interactief media-museum),

- een kennisinstituut inzake archivering van audiovisueel materiaal zijn.

B&G is opgericht om het duurzaam behoud van het Nederlandse nationale audiovisuele erfgoed te garanderen en het toegankelijk te maken voor zoveel mogelijk gebruikers: professionals, het onderwijs en het grote publiek. Op termijn wil B&G de audiovisuele collectie migreren naar een digitaal archief. Het instituut beschikt nu reeds over een digitale collectie van 10.000 uur.

B&G neemt deel aan verschillende nationale en internationale onderzoeksprojecten met betrekking tot technologieën voor digitale conservering en ontsluiting van audiovisueel materiaal. Eén van die projecten die hier als voorbeeld wordt aangehaald, is MultiMatch, waarin men een Europese versie van het 'Geheugen van Nederland' wil realiseren.[158]

7.2.1 MultiMatch

Het MultiMatch-project ambieert de ontwikkeling van een Europese meertalige zoekmachine voor cultureel erfgoedonderzoek. In het project participeren, naast het Instituut voor Beeld & Geluid, tien andere partners waaronder Fratelli Alinari, Biblioteca Virtual Miguel de Cervantes, OCLC PICA, Dublin City University en Universiteit Amsterdam.

De MultiMatch-zoekmachine-*engine* 'crawled' en indexeert culturele erfgoedsites met gedigitaliseerde objecten, zoals bibliotheken, fotoarchieven, musea en audiovisuele archieven. Behalve tekstuele bronnen behandelt het systeem ook beeldmateriaal en videofragmenten. Bovendien moet het systeem minstens vier talen ondersteunen.

De aandacht gaat vooral uit naar de zoekfunctionaliteit van het systeem. De nadruk ligt dan ook bijna uitsluitend op beschrijvende metadatamodellen. Het probleem is dat in de sectoren van cultureel erfgoed veel verschillende datamodellen in gebruik zijn, wat voor problemen zorgt wanneer men de verschillende collecties wil samenvoegen tot één doorzoekbaar geheel. Niet alleen het gebruik van verschillende metadatamodellen (MARC, ISAD, VRA,…) maar ook het gebruik van verschillende ontologieën, thesauri of gecontroleerde woordenlijsten (LCSH, AAT,…) in de diverse sectoren staan de semantische interoperabiliteit van die collecties in de weg.

158 MultiMatch (2009).

In het kader van het MultiMatch-project achtten de onderzoekers het niet haalbaar om een nieuw schema te ontwikkelen om dit vervolgens in de verschillende sectoren te introduceren. Er werd veeleer gezocht naar een gemeenschappelijke standaard waar de specifieke modellen aan gemapt kunnen worden. Bij de keuze of ontwikkeling van deze gemeenschappelijke standaard onderscheidden de MultiMatch-onderzoekers een drietal bepalende factoren:

- verwachtingen van de eindgebruikers m.b.t. de zoekfunctionaliteiten,
- de specifieke kenmerken van de gebruikte metadataschema's van de instellingen die data zullen aanleveren. De afzonderlijke schema's moeten (semi-)automatisch aan het gemeenschappelijke model gemapt kunnen worden.
- de specifieke kenmerken van de objecten die beschreven moeten worden.

Het onderzoek naar semantische interoperabiliteit van de verschillende collecties van de partners in MultiMatch start met een inventarisatie van een 40-tal metadataschema's, ontologieën en algemene referentiemodellen die gebruikt worden in de culturele erfgoedsector.[159] Een eerste vaststelling hierbij was dat in de verschillende erfgoedsectoren gebruik wordt gemaakt van schema's en ontologieën die specifiek gericht zijn op de beschrijving van objecten in die sectoren en dus met het oog op hun specifieke gebruikers. Elk schema was te specifiek om als gemeenschappelijke standaard te gebruiken. Verder bleek uit de inventaris dat, hoewel er onderlinge crosswalks mogelijk zijn tussen veel van de gebruikte medatatastandaarden, interoperabiliteit tussen de verschillende schema's meestal wordt bekomen door te mappen van en naar Dublin Core (DC).

Het probleem is dat DC minder expressief is dan de meer specifieke schema's, waardoor er steeds informatieverlies optreedt bij het mappen van de specifieke schema's naar DC. Dat kan gedeeltelijk opgelost worden door gebruik te maken van de standaarduitbreidingen op DC, zoals Qualified Dublin Core. Een ander alternatief is mappen naar meer expressieve metadatastandaarden zoals MPEG7/21 of naar referentiemodellen zoals FRBR en/of CIDOC CRM.

159 Oomen en Smulders (2006).

In het vervolgonderzoek (D.2.2.1) wordt gezocht naar een datamodel dat interoperabiliteit tussen de heterogene collecties moet toelaten. Dit zogenaamde Multimatch-datamodel moest aan een aantal voorwaarden voldoen:

- Het metadataschema moet in XML uitgedrukt worden om de interactie met techologieën in het semantic web te vergemakkelijken.
- De metadataschema's van de aanleverende instellingen moeten (semi-)automatisch gemapt kunnen worden naar het Multimatch-model.
- Het schema moet het object in zijn geheel en volgens relevante subonderdelen kunnen beschrijven. Het moet dus een hiërarchische opstelling kennen.
- Het schema moet gebruik maken van een geïntegreerde en gedeelde ontologie.

DC is de internationaal meest gebruikte standaard voor de beschrijving van objecten in de culturele erfgoedsector. Bovendien kunnen de relevante elementen van om het even welk metadataschema gemapt worden naar de DC-velden, evenwel met verlies van informatie. Voor de doeleinden van het Multimatch-project wordt Dublin Core niet expressief genoeg bevonden. Een meer expressieve standaard zoals MPEG 7 voldeed echter ook niet omdat die in de eerste plaats ontwikkeld is om audiovisuele objecten te beschrijven en dus minder geschikt is om bijvoorbeeld fysieke objecten en hun kenmerken te beschrijven. Doorslaggevend is bovendien dat MPEG 7 momenteel nauwelijks gebruikt wordt in de culturele erfgoedsector.

Uiteindelijk wordt geopteerd om een nieuw Multimatch-metadataschema te ontwikkelen, gebaseerd op DCMI-*metadataterms*.[160]

Naast een gemeenschappelijk metadataschema moet ook een gemeenschappelijke semantiek gehanteerd worden bij de invulling van de waarden in de velden van het schema. Dat is mogelijk door het gebruik van relevante standaardthesauri en gecontroleerde woordenlijsten. In het kader van het Multimatch-project werd beslist de Getty-thesauri te gebruiken omdat ze wijd verspreid zijn en door toonaangevende organisaties gebruikt worden.[161] Bovendien bestaan ze in verschillende vormen, waaronder XML.

160 Zie voor gedetailleerde documentatie over het Multimatch-metadataschema: Multimatch (z.j.).

161 Getty (2009a, 2009b). Zie ook §5.4.7.

De drie bruikbare Getty-thesauri zijn:

- Getty Arts and Architecture Thesaurus (AAT): artist descriptions.
- Getty Unified List of Artist Names (ULAN): creators names and information.
- Getty Thesaurus of Geographic Names (TGN): geospatial information.

Er diende wel nog gezocht te worden naar een consequente manier om deze woordenlijsten uit te breiden met ontbrekende termen.

Samengevat zal het metadatamodel van het MultiMatch-systeem er als volgt uitzien:

- Intern: het MultiMatch-metadataschema als een uitbreiding van DCMI-*metadataterms*.
- Uitwisseling: mapping van het MultiMatch-schema naar DC met de vijftien elementen.
- Voor verdere interoperabiliteit binnen de erfgoedsector wordt het MultiMatch-schema gemapt naar CIDOC CRM.

8 Conclusies

In dit rapport werd een overzicht gegeven van opslagformaten, metadatastandaarden en containerstandaarden, die de verschillende niveaus representeren waarop men digitale media dient te beschrijven om hun bewaring op lange termijn te garanderen. Op elk niveau situeren zich immers mogelijke gevaren voor dataverlies indien die beschrijvingen niet adequaat en doordacht gebeuren.

Op het laagste niveau is een digitaal bestand opgebouwd uit bits en bytes die op hardwaresystemen opgeslagen zijn. Deze systemen zijn vaak onderhevig aan zogenaamde *wear-and-tear*. Vaste schijven en tapes hebben een beperkte levensduur. In de loop van de tijd kunnen digitale bitstreams zich door externe invloeden, zoals corruptie van de dragers, wijzigen. Op dit laagste niveau zijn er hardware- en softwareoplossingen beschikbaar om deze fouten te herstellen. Door verschillende versies van de digitale bestanden op meerdere plekken op aarde op te slaan, kunnen rampscenario's zoals dataverlies door overstromingen, branden en diefstal, worden voorkomen.

Op een hoger niveau vormen vele bytes in de vorm van digitale bestanden een representatie van de opgeslagen data. Bestands- en compressieformaten zoals JPEG en AVI beschrijven de wijze waarop de bits omgevormd kunnen worden tot een interpreteerbare multimediapresentatie zoals beeld, video en geluid. Bestandsformaten zijn echter tijdsgebonden. In de jaren tachtig en vroege jaren negentig was WordPerfect bijvoorbeeld een gangbaar bestandsformaat voor de bewaring van tekstuele data. Tegenwoordig kunnen weinig teksteditors deze bestanden nog openen. Wanneer een bestandsformaat in onbruik geraakt, zijn er voor archieven maar twee opties mogelijk om de opgeslagen informatie te behouden: 1) migratie van het oude bestandstype naar een nieuw formaat (bijvoorbeeld van WordPerfect naar PDF), 2) door software-emulatie een werkbare WordPerfect-lezer 'in leven houden'. Beide mogelijkheden hebben voor- en nadelen. Door migratie kan informatie verloren gaan en softwarearchivering door middel van emulatie is zeer complex. Om bestandsformaten op lange termijn toegankelijk te houden en een migratie zonder dataverlies mogelijk te maken, is het gebruik van open standaarden noodzakelijk. Bij gesloten bestandsformaten zijn er altijd softwaretools nodig om de data te renderen. Zoals hierboven beschreven, is de archivering van software (en in extremis een werkende IT-infrastructuur) complexer dan het gebruik van open standaarden. Ook bij de keuze van compressieformaten

dient men rekening te houden met open en gesloten compressieformaten en -technieken die compressie zonder verlies garanderen.

De bestandsformaten vormen een representatie van de opgeslagen informatie. Bij multimediale data is echter niet alleen de opgeslagen informatie belangrijk maar ook de *look-and-feel* moet behouden blijven. Indien bijvoorbeeld door migraties van bestandsformaten de resolutie of kleurinformatie in beelden verloren gaat, dan betekent dit een verlies aan informatie. Net zoals bij digitalisering van informatie in analoge vorm veel aandacht besteed moet worden aan het behoud van alle aspecten van het originele object, zo zal bij migratie van digitale bestanden een rijke beschrijving van de *look-and-feel* noodzakelijk zijn. Digitale informatie is ook een conceptueel object dat altijd in een bestaande IT-infrastructuur geïnterpreteerd moet worden.

De authenticiteit van de digitale informatie is aan grotere gevaren onderhevig dan data in analoge vorm. In het laatste geval is het voldoende om alle karakteristieken van het fysieke object te beschrijven, in het eerste geval dienen de gehele ontstaans- en verwerkingsgeschiedenis gearchiveerd te worden.

Op een nog hoger niveau is contextuele informatie onontbeerlijk voor de interpretatie van het digitale bestand. Op lange termijn zullen de producenten van de informatie immers niet meer beschikbaar zijn om de gearchiveerde dataset toe te lichten. Een datacollectie moet van contextuele data vergezeld worden om een volledige beschrijving van de informatie te bieden die, zonder de hulp van externe experts, voor een welomschreven doelpubliek of *designated community* interpreteerbaar blijft.

Op het hoogste niveau wordt een dataset niet enkel door experts maar ook in organisaties en in een tijdsgebonden discours of jargon geproduceerd. Organisatiestructuren kunnen echter wijzigen of verdwijnen en het discours dat eigen is aan een specifieke (productie)context en eindgebruikersgroep met een gemeenschappelijke achtergrondkennis is eveneens tijdsgebonden. Het is dan ook noodzakelijk voldoende informatie mee over te leveren om de data begrijpelijk te houden.

Preservering van digitale objecten is daarom vanuit minstens drie perspectieven belangrijk: preservering van het medium, preservering van technologie en preservering van de intellectuele inhoud.

Rekening houdend met die perspectieven is een gelaagd metadatamodel nodig om de data op de drie respectieve niveaus volledig en nauwkeurig te beschrijven:

- Binaire schema's beschrijven de data tot op bitniveau.
- Technische schema's beschrijven op een hoger niveau hoe bytes vertaald worden naar concepten die door mensen geïnterpreteerd kunnen worden, zoals beeld, video en geluid.
- Descriptieve schema's geven een inhoudelijke beschrijving van de data, titels, auteurs, programma's en dateringen.
- Preserveringsschema's beschrijven relaties tussen de databestanden en geven contextuele informatie. De schema's geven technische en administratieve informatie over de ontstaansgeschiedenis van de data en eventuele wijzigingen die ze ondergaan.
- Structurele schema's geven een beschrijving van alle delen van een digitaal object en de relaties tussen de digitale objecten onderling.

In dit rapport werden ook gangbare descriptieve metadataschema's aangehaald die mogelijk door de verschillende projectpartners en instellingen gebruikt worden. Bestandsformaten voor multimedia hebben een zeer breed toepassingsdomein en worden in principe door alle betrokken instellingen geproduceerd. Voor descriptieve standaarden zijn er echter veel onderlinge verschillen. Zo zijn er voor bibliografische beschrijvingen van boeken in de bibliotheeksector andere velden belangrijk dan voor de beschrijving van archiefstukken in de erfgoedsector. Beschrijvingen van videobestanden in de omroepsector verschillen van beschrijvingen van videokunst in de museumsector.

Voor sommige sectoren kan gerefereerd worden naar projecten waarin het ontwerp van een gemeenschappelijke (sectorspecifieke) standaard centraal staat of waarin de gebruikte metadataschema's bevraagd en onderzocht werden. IPEA (Innovatief Platform voor Elektronische Archivering)[162] is een IBBT-project, gericht op de omroepsector, waarin P/Meta als generieke descriptieve metadatastandaard voor de betreffende sector gesuggereerd wordt. Deze internationale standaard blijkt namelijk zeer verdienstelijk voor de B2B-uitwisseling van omroepdata (cf. §5.2.3). Voor het digitaal archief van BOM, waarbij de meeste omroepen betrokken zijn, zal P/Meta als omroepstandaard dan ook belangrijk zijn. Uit bevragingen van verschillende erfgoed-

162 IBBT (2007a).

instellingen voor het project Erfgoed 2.0, een IBBT-project waarin de digitale interactie tussen erfgoedinstellingen in de lijn van Web 2.0 en Library 2.0 beoogd wordt,[163] blijkt onder meer dat in de erfgoedsector een geleidelijk proces van standaardisatie aan de gang is maar dat dit nog lang niet voltooid is. Een suggestie voor een standaard is dan ook noodzakelijk. Hetzelfde geldt voor de museumsector. Ook hier kan gerefereerd worden aan digitale samenwerkingsinitiatieven zoals het project MOVE (Musea Oost-Vlaanderen in Evolutie)[164], waarvoor op termijn een gemeenschappelijke museumstandaard dient afgesproken te worden. Ten slotte wordt hier ook het recent project 'Van Horen zeggen' aangehaald, waarin met verschillende erfgoedinstellingen onderzocht werd hoe men mondelinge bronnen kan bewaren en ontsluiten.[165] Ook hierin werden verschillende formaten, metadatastandaarden en containerformaten met elkaar afgewogen. Deze haalbaarheidsstudie kon zich baseren op bevindingen van het IBBT-project POKUMON (Podiumkunsten Multimediaal Ontsloten), dat zich richtte op de Vlaamse (digitale) archiefwerking van hoofdzakelijk audiovisuele archiefinstellingen.[166] Het spreekt voor zich dat het BOM-vl-project rekening zal houden met bevindingen en conclusies van deze projecten.

Onder meer op basis van de genoemde projectresultaten mag men besluiten dat het vinden van een grootste gemene deler die de beschrijvingswijze van àlle mogelijke materiaalsoorten dekt, een onhaalbare opgave is. Iedere sector met zijn specifieke materiaalsoorten en data stelt immers afzonderlijke eisen met betrekking tot metadata. Een dergelijke algemene generieke standaard zou tot onnodig veel (meta)dataverlies leiden terwijl het raadzaam is om met het oog op langetermijnbewaring de detaillistische en volledige metadata van de verschillende sectoren mee te archiveren en te bewaren.

Het gelaagd metadatamodel dat voorgesteld zal worden, moet dit probleem ondervangen. Er zal namelijk gestreefd worden naar een model dat in zijn uniforme basislaag zo algemeen mogelijk is en in de verfijningslagen meer specifieke metadata bevat die relevant zijn voor de betreffende toepassingsgebieden. Als tussenlaag wordt aan de verschillende sectoren echter voorgesteld een sectorspecifieke metadatastandaard te gebruiken. Samengevat komt het er op neer dat iedere instelling voor haar materiaal uitmaakt welke specifieke metadata van belang zijn en dat ze dus het eigen archiveringssysteem behoudt. Vervolgens past de instelling, afhankelijk van de sector

163 IBBT (2004-2006).

164 MOVE (2009).

165 Walterus (2009).

166 IBBT (2007b).

waartoe ze behoort of van het materiaal dat ze bezit, de afgesproken (en door BOM-vl voorgestelde) sectorspecifieke metadatastandaard toe (bijvoorbeeld MARC voor de bibliotheeksector, EAD voor de archiefsector, P/Meta voor de omroepsector, enz.). Ten slotte zullen de sectorspecifieke metadata, die zoals gezegd ook in het gelaagd metadatamodel opgenomen zullen worden, gemapt worden naar een generieke sectoroverschrijdende metadatastandaard die het beheer en de doorzoekbaarheid van het volledige digitale archief zal mogelijk maken. Voor deze generieke laag wordt vaak (Qualified) Dublin Core geopperd.

Bij de ontwikkeling van een metadatamodel voor de archivering van digitale multimedia moet men dus rekening houden met metadatabeschrijvingen op alle niveaus, van bitlevelbeschrijvingen tot beschrijvingen van de intellectuele inhoud. Om dit te verwezenlijken zijn descriptieve, technische, administratieve, structurele en contextuele metadata nodig. In zijn generieke basislaag zien de beschrijvingen van de uiteenlopende gearchiveerde digitale materiaalsoorten er identiek uit. Op een fijner niveau worden ook alle sector- en materiaalspecifieke metadata bewaard.

9 Bibliografie

3GPP (2009). *3GPP Specification Detail*, 3GPP, http://www.3gpp.org/ftp/Specs/html-info/26244.htm {29/01/2009}.

Adler, M., T. Boutell, J. Bowler and et al. (2003). *W3C Recommendation: Portable Network Graphics (PNG) Specification (Second Edition). Information Technology - Computer Graphics and Image Processing - Portable Network Graphics (PNG): Functional Specification. ISO/IEC 15948: 2003 (E)*, edited by David Duce: Oxford Brookes University, W3C, http://web4.w3.org/TR/PNG/ {21/01/2009}.

Adobe (2008a). *TIFF*, Adobe Systems Inc., http://partners.adobe.com/public/developer/tiff/index.html {21/01/2009}.

Adobe (2008b). *Video File Format Specification. Version 10*, 44 p., Adobe Systems Inc., http://www.adobe.com/devnet/flv/pdf/video_file_format_spec_v10.pdf {29/01/2009}.

Adobe Developers Association (1992). *TIFF (Revision 6.0)*, 121 p., Adobe Systems Inc., Mountain View, CA, http://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf {21/01/2009}.

Allinson, J. (2006). *OAIS as a Reference Model for Repositories. An Evaluation. Revision 0.5*, 17 p., UKOLN, University of Bath, http://www.ukoln.ac.uk/repositories/publications/oais-evaluation-200607/Drs-OAIS-evaluation-0.5.pdf {19/01/2009}.

Apple (2005). *Introduction to Quicktime Overview*, Apple Inc., http://developer.apple.com/documentation/QuickTime/RM/Fundamentals/QTOverview/QTOverview_AIntro/chapter_1_section_1.html {29/01/2009}.

Apple Computer Inc. (1989). *Audio Interchange File Format: "AIFF". A Standard for Sampled Sound Files (Version 1.3)*, 31 p., Apple Computer Inc., Cupertino, CA, http://www-mmsp.ece.mcgill.ca/Documents/AudioFormats/AIFF/Docs/AIFF-1.3.pdf {21/01/2009}.

Apple Computer Inc. (1991). *Draft. Audio Interchange File Formatt AIFF-C. A Revision to Include Compressed Audio Data*, 41 p., Apple Computer Inc., http://www-mmsp.ece.mcgill.ca/Documents/AudioFormats/AIFF/Docs/AIFF-C.9.26.91.pdf {21/01/2009}.

Architecture & Patrimoine (2009). *Architecture & Patrimoine Homepage*, Ministère de la Culture et de la Communication, http://www.culture.gouv.fr/culture/inventai/patrimoine/ {28/01/2009}.

ATSC (2005). *Digital Audio Compression Standard (AC-3, E-AC-3). Revision B. Doc A/52B*, 236 p., Washington DC: ATSC (Advanced Television Systems Committee), http://www.atsc.org/standards/a_52b.pdf {20/01/2009}.

Attig, J., A. Copeland and M. Pelikan (2004). 'Context and Meaning: The Challenges of Metadata for a Digital Image Library within the University', In: *College & Research Libraries* 65, no. 3, pp. 251-61, http://www.ftrf.org/ala/mgrps/divs/acrl/publications/crljournal/2004/crlmay04/copeland.pdf {28/01/2009}.

Baca, M. (2007). 'CCO and CDWA Lite: Complementary Data Content and Data Format Standards for Art and Material Culture Information', In: *VRA Bulletin* 34, no. 1, pp. 69-75, Getty Research Institute, http://www.vraweb.org/seiweb/readings-prep/CCOandCDWA_Lite-Baca.pdf {26/01/2009}.

Baca, M., S. Clarke, J. Eklund, A.J. Gilliland, P. Harpring, M.S. Woodley and E. O'Keefe (2009). *Metadata Standards Crosswalk*, Murtha Baca, J. Paul Getty Trust, http://www.getty.edu/research/conducting_research/standards/intro-metadata/crosswalks.html {28/01/2009}.

BBC (2007). *SMEF Data Model*, http://www.bbc.co.uk/guidelines/smef/ {23/01/2009}.

Beagrie, N. (2004). 'The Continuing Access and Digital Preservation Strategy for the Uk Joint Information Systems Committee (JISC)', In: *D-Lib Magazine* 10, no. 7/8, pp. 1082-9873, http://www.dlib.org/dlib/july04/beagrie/07beagrie.html {19/01/2009}.

Beeld en Geluid Instituut (2009). *Multimatch. Meertalige zoekmachine*, Nederlands Instituut voor Beeld en Geluid, http://instituut.beeldengeluid.nl/index.aspx?ChapterID=8599 {28/01/2009}.

Betacam PALsite (2000). *Betacam PALsite. The Betacam Web Resource*, A. Barnett and M. Evans, http://betacam.palsite.com/ {29/01/2009}.

Blackstock, S. (z.j). *LZW and GIF Explained*, http://www.cis.udel.edu/~amer/CISC651/lzw.and.gif.explained.html {21/01/2009}.

Bormans, J. and K. Hill (ed.) (2002). *MPEG - 21 Part 9 - File Format', §5.9 in MPEG-21. Overview V.5. ISO/IEC JTC 1/SC 29/WG11/NS231*, Shangai: ISO/IEC, http://www.chiariglione.org/mpeg/standards/mpeg-21/mpeg-21.htm {28/01/2009}.

Brindley, L. (2000). 'Taking the British Library Forward in the Twenty-First Century', In: *D-Lib Magazine* 6, no. 11, http://www.dlib.org/dlib/november00/brindley/11brindley.html {19/01/2009}.

Carlyle, A. (2006). 'Understanding FRBR as a Conceptual Model: FRBR and the Bibliographic Universe', In: *Library Resources & Technical Services Year* 50, no. 4, pp. 264-274, http://projects.ischool.washington.edu/acarlyle/Papers/Carlyle_FRBR_2006.htm {29/01/2009}.

CASPAR (2006). *Caspar Project. Cultural, Artistic and Scientific Knowledge for Preservation, Access and Retrieval*, European Commission. Sixth Framework Programme, http://www.casparpreserves.eu/ {28/01/2009}.

Cathro, W. and T. Boston (2003). *Development of a Digital Services Architecture at the National Library of Australia*, National Library of Australia, http://www.nla.gov.au/nla/staffpaper/2003/cathro1.html {28/01/2009}.

CCSDS (2002). *Reference Model for an Open Archival Information System (OAIS). Blue book. Issue 1*, 148 p., CCSDS, http://public.ccsds.org/publications/archive/650x0b1.pdf {19/01/2009}.

CEDARS (2000). *Metadata for Digital Preservation: The Cedars Project Outline Specification. Draft for Public Consultation*, CEDARS, http://www.leeds.ac.uk/cedars/colman/metadata/metadataspec.html#note6 {28/01/2009}.

Chiariglione, L. (1996). *Short MPEG-1 Description*, MPEG, http://www.chiariglione.org/mpeg/standards/mpeg-1/mpeg-1.htm {19/01/2009}.

Chiariglione, L. (2000). *Short MPEG-2 Description*, MPEG, http://www.chiariglione.org/mpeg/standards/mpeg-2/mpeg-2.htm {19/01/2009}.

Clarke, S. (2001). *VRA Core 3.0 Mapping to MARC 21 (Bibliographic Format)*, Indiana University, http://php.indiana.edu/~fryp/marcmap.html {28/01/2009}.

Coalson, J. (2008). *FLAC: Free Lossless Audio Codec*, http://flac.sourceforge.net/ {20/01/2009}.

Collections Trust (2009). *The SPECTRUM Standard*, Collections Trust, http://www.collectionslink.org.uk/manage_information/spectrum {28/01/2009}.

COM (2006). *The CIDOC Conceptual Reference Model Homepage*, COM (International Council of Museums), http://cidoc.ics.forth.gr/ {29/01/2009}.

CompuServe Inc. (1990). *GIF: Graphics Interchange Format (sm) Version 89a*, 34 p., Colombus, Ohio: CompuServe Inc., http://www.w3.org/Graphics/GIF/spec-gif89a.txt {21/01/2009}.

CORDRA (2006). *Content Object Repository Discovery and Registration/Resolution Architecture*, CORDRA Management Group, http://cordra.net {26/01/2009}.

CoreCodec (2005-2009a). *Matroska. Audio Tags Example*, CoreCodec Inc., http://www.matroska.org/technical/specs/tagging/example-audio.html {29/01/2009}.

CoreCodec (2005-2009b). *Matroska*, CoreCodec Inc., http://www.matroska.org/contact/index.html {29/01/2009}.

Crofts, N., M. Doerr, T. Gill, S. Stead and M. Stiff (ed.) (2006). *Definition of the CIDOC Conceptual Reference Model. Version 4.2.1*, 93 p., ICOM/CIDOC, http://cidoc.ics.forth.gr/docs/cidoc_crm_version_4.2.1.pdf {29/01/2009}.

CSH (2008). *CSH: Canadian Subject Headings*, Library and Archives Canada, http://www.collectionscanada.gc.ca/6/23/ {28/01/2009}.

DCI (2008). *Digital Cinema Initiatives System Requirements and Specifications for Digital Cinema. Version 1.2*, Digital Cinema Initiatives, LLC, http://www.dcimovies.com/specification/index.tt2 {19/01/2009}.

DCMI (2006). *DCMI Preservation Community*, DCMI, http://dublincore.org/groups/preservation/ {23/01/2009}.

DCMI (2009a). *Dublin Core Metadata Initiative Documents*, DCMI, http://dublincore.org/documents/ {23/01/2009}.

DCMI (2009b). *The Dublin Core Metadata Initiative Homepage*, DCMI, http://dublincore.org/ {23/01/2009}.

De Sutter, R., S. Notebaert, L. Hauttekeete and R. Van de Walle (2006). 'IPEA: The Digital Archives Use Case.' In: *Archiving 2006*, pp. 182-186, http://en.scientificcommons.org/12799607 {29/01/2009}.

De Sutter, R., S. Notebaert and R. Van de Walle (2006). 'Evaluation of Metadata Standards in the Context of Digital Audio-Visual Libraries', In: *Lecture notes in Computer Science* 4172, pp. 220-231, http://www.springerlink.com/content/x1120290t4v12775/fulltext.pdf {29/01/2009}.

Dekker, R. and M. Slabbertje (2003). 'e-archiving. Het duurzaam bewaren van wetenschappelijke digitale bronnen', In: *Informatie Professional* 7, no. 6, pp. 32-34, http://igitur-archive.library.uu.nl/DARLIN/2005-0526-200314/DekkerIP062003.pdf {30/01/2009}.

Dempsey, L. and B. Lavoie (2005). *DLF Service Framework for Digital Libraries. A Progress Report for the DLF Steering Committee*, DLF, http://www.diglib.org/architectures/serviceframe/dlfserviceframe1.htm {26/01/2009}.

DEN (2007). *Encoded Archival Description (EAD). Versie 1.0*, Den Haag: DEN (Digitaal Erfgoed Nederland), http://www.den.nl/docs/20070521104422/ {28/01/2009}.

DEN (2008). *Dublin Core in samenwerkingsprojecten en publieksgerichte ontsluiting. Versie 1.1*, Den Haag: DEN (Digitaal Erfgoed Nederland), http://www.den.nl/docs/20050816173630 {23/01/2009}.

Digital Formats (2007a). *Macromedia Flash FLV Video File Format*, Digital Formats, http://www.digitalpreservation.gov/formats/fdd/fdd000131.shtml {29/01/2009}.

Digital Formats (2007b). *MPEG-4 File Format. Version 2*, Digital Formats, http://www.digitalpreservation.gov/formats/fdd/fdd000155.shtml {29/01/2009}.

Digital Video (2009). *DV: Digital Video. Tools & Technology for Video Professionals – Homepage*, Digital Video, http://www.dv.com/ {29/01/2009}.

DIRAC (2008). *Dirac Video Compression*, David A. Schleef, http://diracvideo.org {19/01/2009}.

DivX (2009). *DivX*, DivX, http://www.divx.com/ {19/01/2009}.

Djuric, A. and J. Oler (ed.) (2006). *The World of True Audio*, True Audio Codec Software, http://true-audio.com/ {21/01/2009}.

Doerr, M. (2003). 'The CIDOC Conceptual Reference Module - an Ontological Approach to Semantic Interoperability of Metadata.' In: *AI magazine* 24, no. 3, pp. 75-92.

Dolby (2008). *aacPlus*, Coding Technologies, http://www.codingtechnologies.com/products/aacPlus.htm {26/01/2009 }.

Dolby (2009a). *Dolby Digital*, Dolby Laboratories, Inc., http://www.dolby.com/consumer/technology/dolby_digital.html {20/01/2009}.

Dolby (2009b). *Dolby TrueHD*, Dolby Laboratories, Inc., http://www.dolby.com/consumer/technology/trueHD.html {21/01/2009}.

Doorenbosch, P. and T. van Veen (2009). 'Nieuwe gegevensarchitectuur ondersteunt nieuwe diensten. Koninklijke Bibliotheek en Web 2.0', In: *Informatie Professional* 4, pp. 24-29, http://research.kb.nl/Publicaties/IP200704_24_29_proef_Doorenbosch.pdf {28/01/2009}.

DRAMBORA (2008). *Digital Repository Audit Method Based On Risk Assesment*, DCC en DPE, http://www.repositoryaudit.eu/ {19/01/2009}.

EBML (z.j). *EBML*, SourceForge.net, http://ebml.sourceforge.net {29/01/2009}.

EBU (2003). *EBU Subjective Listening Test on Low-Bitrate Audio Codecs*, 44 p., EBU-UER, http://www.ebu.ch/CMSimages/en/tec_doc_t3296_tcm6-10497.pdf {20/01/2009}.

EBU (2005). *The EBU Metadata Exchange Scheme - P_META. Version 1.2. Publication Release*, 241 p., EBU (European Broadcasting Union), http://www.ebu.ch/CMSimages/en/tec_doc_t3295_v0102_tcm6-40957.pdf {23/01/2009}.

EBU (2007). *P_META 2.0. Metadata Library. Version 2.0. Tech 3295-V2*, 20 p., Geneva: EBU, http://www.ebu.ch/CMSimages/en/tec_doc_t3295v2-2007_tcm6-53551.pdf {23/01/2009}.

EBU (2009). *European Broadcasting Union Homepage*, EBU, http://www.ebu.ch/ {23/01/2009}.

EDItEUR (2009). *EDItEUR. Co-Ordinating the Development, Promotion and Implementation of Electronic Commerce in the Book and Serials Sectors*, London: EDItEUR, http://www.editeur.org/ {26/01/2009}.

Eeckhaut, H., B. Schrauwen, M. Christiaens and J. Van Campenhout (2005). 'Speeding up Dirac's Entropy Coder', In: *5th WSEAS Int. Conf. on Multimedia, Internet and Video Technologies*, 17-19/08/2005, 6 p., Corfu, Griekenland, http://escher.elis.ugent.be/publ/Edocs/DOC/P105_087.pdf {19/01/2009}.

Europe's Information Society (2008). *eContentplus Programme*, Europe's Information Society. Thematic Portal, http://ec.europa.eu/information_society/activities/econtentplus/index_en.htm {28/01/2009}.

ExLibris (2008a). *Digitool. Managing and Showcasing Digital Collections and Institutional Repositories*, ExLibris, http://www.exlibrisgroup.com/category/DigiToolOverview {19/01/2009}.

ExLibris (2008b). *Rosetta. A New Way of Preserving Cultural Heritage and Cumulative Knowledge*, ExLibris, http://www.exlibrisgroup.com/category/ExLibrisRosettaOverview {19/01/2009}.

Farquhar, A. and H. Hockx-Yu (2007). 'Planets: Integrated Services for Digital Preservation', In: *International Journal of Digital Curation* 2, no. 2, pp. 88-98, http://www.ijdc.net/index.php/ijdc/article/viewFile/30/19 {28/01/2009}.

Fraunhofer IIS. (2008a). *MP3: MPEG Audio Layer III*, Fraunhofer IIS, http://www.iis.fraunhofer.de/EN/bf/amm/projects/mp3/index.jsp {29/01/2009}.
Fraunhofer IIS (2008b). *MPEG-2 and MPEG-4 Advanced Audio Coding*, Fraunhofer IIS, http://www.iis.fraunhofer.de/EN/bf/amm/projects/mpeg/index.jsp {29/01/2009}.
Friesen, N. (2005). *CanCore: Learning Object Metadata Editors*, CanCore, http://www.cancore.ca/editors.html {28/01/2009}.
GAMA (2009). *GAMA: Gateway to Archives of Media Art*, GAMA, http://www.gama-gateway.eu/ {28/01/2009}.
Getty (2004a). *Getty Thesaurus of Geographic Names Online. Full Record Display*, J. Paul Getty Trust, http://www.getty.edu/vow/TGNFullDisplay?find=brussels&place=&nation=&prev-page=1&english=Y&subjectid=7007868 {29/01/2009}.
Getty (2004b). *Union List of Artist Names Online. Full Record Display*, J. Paul Getty Trust, http://www.getty.edu/vow/ULANFullDisplay?find=le+corbusier&role=&nation=&prev_page=1&subjectid=500027041 {29/01/2009}.
Getty (2006). *CDWA: Categories for the Description of Works of Art*, J. Paul Getty Trust, http://www.getty.edu/research/conducting_research/standards/cdwa/ {26/01/2009}.
Getty (2008a). *About the Art&Architecture Thesaurus Online*, J. Paul Getty Trust, http://www.getty.edu/research/conducting_research/vocabularies/aat/about.html {29/01/2009}.
Getty (2008b). *About the Getty Thesaurus of Geographic Names Online*, J. Paul Getty Trust, http://www.getty.edu/research/conducting_research/vocabularies/tgn/about.html {29/01/2009}.
Getty (2009a). *Getty Search Vocabularies*, J. Paul Getty Trust, http://www.getty.edu/Search/ {28/01/2009}.
Getty (2009b). *Learn About the Getty Vocabularies*, J. Paul Getty Trust, http://www.getty.edu/research/conducting_research/vocabularies/ {28/01/2009}.
Giles, R. (2004). *Ogg Theora a Free Video Codec and Multimedia Platform*, Vancouver-Canada: Xiph.org Foundation, http://people.xiph.org/~giles/2004/openweekend/talk/theora.pdf {19/01/2009}.
Gladney, H.M. (2007). 'Digital Preservation in a national context', In: *D-Lib Magazine* 13, no. 1-2, http://www.dlib.org/dlib/january07/gladney/01gladney.html {28/01/2009}.
Goldman, M. (z.j.). *A Comparison of MPEG-2 Video, MPEG-4 AVC, and SMPTE VC-1 (Windows Media 9 Video)*, 20 p., TANDBERG Television, http://video.ldc.lu.se/pict/WM9V-MP4AVC-MP2V_comparison-Goldman.pdf {19/01/2009}.
Guenther, R.S. (2003). 'MODS: The Metadata Object Description Schema', In: *Portal: Libraries and Academy* 3, no. 1, pp. 137-150, http://muse.jhu.edu/journals/portal_libraries_and_the_academy/v003/3.1guenther.html {26/01/2009}.

Hacker, S. and S. Hayes (2000). *Mp3: The Definitive Guide*, 388 p., Edited by Simon Hayes, Sebastopol, CA, USA: O'Reilly & Associates, Inc.

Harmony (2009). *About Harmony*, Harmony, http://metadata.net/harmony/ {28/01/2009}.

Haslhofer, B. and R. Hecht (2005). *The Metadata Manager. Version 2.0*, Bricks Foundation, http://foundation.bricksfactory.org/deliverables/d331/ar01s04.html {29/01/2009}.

Hauttekeete, L., H. Dekeyser, R. De Sutter, F. Mathijs, P. Mechant, S. Notebaert, G. Nulens, P. Schelkens, R. Vermaut and K. Wouters (2006). *Innovative Platform on Electronic Archiving. Digitale archivering op nationaal en internationaal vlak: een stand van zaken*, 171 p., UGent (IBBT, MICT, MMLab), Videohouse, VRT, VMMa, KUL (COSIC, CUO, ICRI), VUB (ETRO, SMIT), Gent: IBBT.

Heijden, H. (2005). *Performance Comparison of Lossless Audio Compressors*, Ziggo B.V., http://members.home.nl/w.speek/comparison.htm {26/01/2009}.

Higgins, S. (2007). *Premis Data Dictionary*, Glasgow: DCC (Digital Curation Centre), http://www.dcc.ac.uk/resource/standards-watch/premis-data-dictionary/ {28/01/2009}.

Hoorens, S., J. Rothenberg, C. van Oranje and M. van der Mandele (2007). *Addressing the Uncertain Future of Preserving the Past: Towards a Robust Strategy for Digital Archiving and Presentation*, Rand Corporation, http://www.ndk.cz/dokumenty/Technologie/addressing-the-uncertain-future-of-preserving-the-past-towards-a-robust-strategy-for-digital-archiving-and-preservation {30/01/2009}.

Hopper, R. (2000). 'P/Meta. Metadata Exchange Standards', In: *EBU Technical Review*, 24 p., http://www.ebu.ch/en/technical/trev/trev_284-hopper.pdf {23/01/2009}.

Hopper, R. (2002). 'P/Meta. Metadata Exchange Scheme, V1.0', In: *EBU Technical Review*, 11 p., http://www.ebu.ch/en/technical/trev/trev_290-hopper.pdf {23/01/2009}.

Houser, L. (2004). 'OCLC Digital Archive Demonstration', In: *Digital Libraries. Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries*, p. 419, 07-11/06/2004.

Hunter, J. (2002). *Combining the CIDOC CRM and MPEG-7 to Describe Multimedia in Museums*, University of Queensland, Australia, http://www.archimuse.com/mw2002/papers/hunter/hunter.html/ {23/01/2009}.

IANA (2007). *MIME Media Types*, IANA (Internet Assigned Numbers Authority), http://www.iana.org/assignments/media-types/ {23/01/2009}.

IBBT (2004-2006). *Erfgoed 2.0. Projectwebsite*, Gent: IBBT, https://projects.ibbt.be/erfgoed2.0/ {29/01/2009}.

IBBT (2007a). *IPEA: Innovatief Platform voor Electronisch Archivering. Projectwebsite*, Gent: IBBT, http://www.ibbt.be/nl/project/ipea-0 {28/01/2009}.

IBBT (2007b). *POKUMON: POdiumKUnsten Multimediaal ONtsloten. Projectwebsite*, Gent: IBBT, http://www.ibbt.be/nl/project/pokumon-0 {28/01/2009}.
IBBT (2009). *BOM-Vlaanderen. Projectwebsite*, Gent: IBBT, https://projects.ibbt.be/bom-vl/ {23/01/2009}.
IBM (2008). *DIAS Home Page*, IBM, http://www-05.ibm.com/nl/dias/ {19/01/2009}.
IBM Corporation, and Microsoft Corporation (1991). *Multimedia Programming Interface and Data Specifications 1.0*, IBM en Microsoft, http://www.kk.iij4u.or.jp/~kondo/wave/mpidata.txt {21/01/2009}.
ICA (2000). *ISAD(G): General International Standard Archival Description, Second Edition*, 91 p., Ottawa: ICA (International Council on Archives), http://www.ica.org/sites/default/files/isad_g_2e.pdf {28/01/2009}.
ICA (2003). *ISAAR(CPF). International Standard Archival Authority Record for Corporate Bodies, Persons and Families*, 94p., http://www.icacds.org.uk/eng/ISAAR(CPF)2ed.pdf {30/01/2009}.
ICA (2006). *ISAAR(CPF): Internationale norm voor archivistische geautoriseerde beschrijvingen van organisaties, personen en families. Vertaling van de tweede uitgave*, 97 p., Antwerpen/Leuven/Amsterdam: Archiefschool, VVBAD, http://www.archiefschool.nl/docs/isaar(cpf)2nl.pdf {28/01/2009}.
ICOM (2009). *ICOM: International Council of Museums*, http://icom.museum/ {28/01/2009}.
IEEE (2002). *Draft Standard for Learning Object Metadata*, 44 p., New York: IEEE (Institute of Electrical and Electronics Engineers), http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf {28/01/2009}.
IEEE LTSC (2007). *LOM XML Binding (1484.12.3)*, IEEE LTSC http://www.ieeeltsc.org/working-groups/wg12LOM/1484.12.3 {28/01/2009}.
IFLA (2008a). *Cataloguing Section. Functional Requirements for Bibliographic Records. Final Report*, IFLANET, http://www.ifla.org/VII/s13/frbr/ {28/01/2009}.
IFLA (2008b*). Cataloguing Section. Functional Requirements for Bibliographic Records. Review Group*, IFLANET, http://www.ifla.org/VII/s13/wgfrbr/index.htm {28/01/2009}.
IFLA (2008c). *Division of Bibliographical Control. Working Group on FRANAR*, IFLANET, http://www.ifla.org/VII/d4/wg-franar.htm {28/01/2009}.
IMS (2001-2008). *IMS Global Learning Consortium*, IMS Global Learning Consortium, Inc., http://www.imsglobal.org {26/01/2009}.
IMS (2003). *IMS Digital Repositories V1.0 Final Specification*, IMS, http://www.imsglobal.org/digitalrepositories/ {26/01/2009}.
ISO (2009). *ISO: International Organisation for Standardization Homepage*, http://www.iso.org/iso/home.htm {23/01/2009}.

ISO/IEC (1993). *ISO/IEC 11172-3: Information Technology - Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to About 1,5 Mbit/S - Part 3: Audio*, ISO/IEC, http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=22412 {20/01/2009}.

ISO/IEC (1997). *ISO/IEC JTC 1/SC 29/WG 1 (JPEG/JBIG): FCD 14495, Lossless and near-Lossless Coding of Continuous Tone Still Images (JPEG-LS). Public Draft*, 75 p., ISO/IEC, http://www.jpeg.org/public/fcd14495p.pdf {21/01/2009}.

ISO/IEC (2000). *ISO/IEC JTC 1/SC 29/WG 1 N1646R (ITU-TSG8): Coding of Still Pictures. JPEG 2000 Part I Final Committee Draft Version 1.0 (ITU-T Rec T800)*, 190 p., ISO/IEC, http://www.jpeg.org/public/fcd15444-1.pdf {29/01/2009}.

ISO/IEC (2002). *ISO/IEC 15444-3. Information Technology - JPEG 2000 Image Coding System - Part 3: Motion JPEG 2000*, 7 p., Zwitserland: ISO/ICE, http://www.iec-normen.de/previewpdf/info_isoiec15444-3%7Bed1.0%7Den.pdf {20/01/2009}.

ITU (1993). *CCITT Recommendation T.81. Terminal Equipment and Protocols for Telematic Services. Information Technology - Digital Compression and Coding of Continuous-Tone Still Images - Requirements and Guidelines*, 182 p., http://www.w3.org/Graphics/JPEG/itu-t81.pdf {21/01/2009}.

ITU (2004). *Recommendation T.81 (09/92). Information Technology - Digital Compression and Coding of Continuous-Tone Still Images - Requirements and Guidelines. ISO/IEC IS 10918-1*, ITU, http://www.itu.int/rec/T-REC-T.81-199209-I/en {29/01/2009}.

ITU (2008). *Recommendation H.264: Advanced Video Coding for Generic Audiovisual Services*, ITU, http://www.itu.int/rec/T-REC-H.264/e {19/01/2009}.

JPEG (2007). *JPEG 2000 - Committee Drafts*, Elysium Ltd, 2kan, http://www.jpeg.org/jpeg2000/CDs15444.html {21/01/2009}.

JPEG/JBIG (2007). *JPEG and JBIG's Official Site*, Elysium Ltd, 2kan, http://www.jpeg.org/ {21/01/2009}.

Kabal, P. (2005). *Audio File Format Specifications*, http://www-mmsp.ece.mcgill.ca/Documents/AudioFormats/AIFF/AIFF.html {21/01/2009}.

KB (2002). *KB/IBM Long-Term Preservation Study*, Koninklijke Bibliotheek. Nationale bibliotheek van Nederland, http://www.kb.nl/hrd/dd/dd_onderzoek/dnep_ltp_study.html {28/01/2009}.

Kessler, B. (2007). 'Encoding Works and Images: The Story Behind Vra Core 4.0.', In: *VRA Bulletin* 34, no. 1, pp. 20-33, http://www.vraweb.org/seiweb/readings-prep/EncodingWorksandImages-Kessler.pdf {28/01/2009}.

Klijn, E. and Y. De Lusenet (2004). *Sepiades. Cataloguing Photographic Collections*, 49 p., Amsterdam: European Commission on Preservation and Access, http://www.knaw.nl/ecpa/publ/pdf/2719.pdf {27 oktober 2008}.

Koenen, R. (2002). *MPEG-4 Overview - (V.21 - Jeju Version)*, MPEG, http://www.chiariglione.org/mpeg/standards/mpeg-4/mpeg-4.htm {19/01/2009}.

Lagoze, C. and J. Hunter (2001). 'The Abc Ontology and Model (V3.0)', In: *Journal of Digital Information, no. Special Issue - selected papers from Dublin Core 2001 Conference*, 18 p., Harmony Project, http://metadata.net/harmony/JODI_Final.pdf {28/01/2009}.

Lagoze, C. and H. Van de Sompel (2007a). *The Open Archives Initiative - Object Reuse and Exchange. Compound Information Objects: The OAI-ORE Perspective*, Open Archives Initiative, http://www.openarchives.org/ore/documents/CompoundObjects-200705.html {29/01/2009}.

Lagoze, C. and H. Van de Sompel (2007b). *The Open Archives Initiative - Object Reuse and Exchange. Ore User Guide – Primer*, Open Archives Initiative, http://www.openarchives.org/ore/1.0/primer.html {29/01/2009}.

Lagoze, C. and H. Van de Sompel (2008). *The Open Archives Initiative Protocol for Metadata Harvesting*, Open Archives Initiative, http://www.openarchives.org/OAI/openarchivesprotocol.html {26/01/2009}.

Lavoie, B.F. (2004). 'Technology Watch Report. The Open Archival Information System Reference Model: Introductory Guide', *DPC Technology Watch Series Report 04-01*, 20 p., Dublin: OCLC/DPC, http://www.dpconline.org/docs/lavoie_OAIS.pdf {19/01/2009}.

Le Boeuf, P. and M. Doerr (2007). 'Harmonising CIDOC CRM and FRBR', In: *International Cataloguing and bibliographic control* 36, no. 4, pp. 90-92.

LOC (1999). *EAD. Application Guidelines for Version 1.0. Appendix B: EAD Crosswalks*, Society of American Archivists, http://www.loc.gov/ead/ag/agappb.html {26/01/2009}.

LOC (2003). 'Available and Useful: Oai at the Library of Congress', In: *Library Hi Tech* 21, no. 2, pp. 129-139, http://memory.loc.gov/ammem/techdocs/libht2003.html#text4 {23/01/2009}.

LOC (2004). *Preservation & Digitization Actions: Terminology for MARC 21 Field 583*, 80 p. LOC (Library of Congress), http://www.loc.gov/marc/bibliographic/pda.pdf {26/01/2009}.

LOC (2006a). *Encoded Archival Description Tag Library, Version 2002 Official Site. Appendix C: Encoded Examples*, LOC (Library of Congress), http://www.loc.gov/ead/tglib/appendix_c.html {23/01/2009}.

LOC (2006b). *MODS Official Web Site: Dublin Core Metadata Element Set Mapping to MODS Version 3*, LOC: Library of Congress, http://www.loc.gov/standards/mods/dcsimple-mods.html {23/01/2009}.

LOC (2008a). *Dublin Core to MARC Crosswalk*, LOC (Library of Congress), http://www.loc.gov/marc/dccross.html {30/01/2009}.

LOC (2008b). *EAD: Encoded Archival Description. Version 2002 Official Site*, LOC (Library of Congress), http://www.loc.gov/ead/ {28/01/2009}.

LOC (2008c). *Marc 21 Format for Bibliographic Data. Version 9*, LOC (Library of Congress), http://www.loc.gov/marc/bibliographic/ecbdhome.html {26/01/2009}.

LOC (2008d). *MARC Mapping to MODS Version 3.3*, MODS, LOC (Library of Congress), http://www.loc.gov/standards/mods/mods-mapping.html {26/01/2009}.
LOC (2008e). *MARC Standards Homepage*, LOC (Library of Congress), http://www.loc.gov/marc {26/01/2009}.
LOC (2008f). *MARC to Dublin Core Crosswalk*, LOC: Library of Congress, http://www.loc.gov/marc/marc2dc.html {23/01/2009}.
LOC (2008g). *MARCXML. MARC 21 XML Schema. Official Web Site*, LOC (Library of Congress), http://www.loc.gov/standards/marcxml/ {26/01/2009}.
LOC (2008h). *MODS Official Web Site*, LOC (Library of Congress), http://www.loc.gov/standards/mods/ {26/01/2009}.
LOC (2008i). *Outline of Elements and Attributes in MODS Version 3.0*, LOC (Library of Congress), http://www.loc.gov/standards/mods/v3/mods-3-0-outline.html {26/01/2009}.
LOC (2009a). *The Library of Congress Subject Headings*, LOC (Library of Congress), http://www.loc.gov/aba/cataloging/subject/ {28/01/2009}.
LOC (2009b). *METS (Metadata Encoding & Transmission Standard) Official Web Site*, LOC (Library of Congress), http://www.loc.gov/standards/mets/ {28/01/2009}.
LOC (2009c). *SRU: Search/Retrieval via URL*, LOC (Library of Congress), http://www.loc.gov/standards/sru/ {26/01/2009}.
Lupovici, C. and J. Masanès (2000). 'Metadata for the Long Term Preservation of Electronic Publications', *NEDLIB Report series 2*, 22 p., Bibliothèque nationale de France, NEDLIB Consortium, http://nedlib.kb.nl/results/NEDLIBmetadata.pdf {28/01/2009}.
Malvar, H.S. (2007). 'Lossless and near-Lossless Audio Compression Using Integer-Reversible Modulated Lapped Transforms', In: *Data Compression Conference (DCC'07)*, 27-29/03/2007, pp. 323-332.
Mannens, E., T. Paridaens, L. Hauttekeete, Evens T. and J. Gysels (2007). *Onderzoeksproject 'Van Horen Zeggen Fase III'. Haalbaarheidsstudie naar een innovatieve applicatie voor de ontsluiting van mondelinge bronnen*, 171 p., Gent: UGent-MMLab/MICT, Universiteit Gent/IBBT, http://www.faronet.be/files/pdf/pagina/van_horen_zeggen_III.pdf {23/01/2009}.
Marpe, D., V. George, H.L. Cycon and K.U. Barthel (2004). 'Performance Evaluation of Motion-JPEG2000 in Comparison with H.264/Avc Operated in Pure Intra Coding Mode', In: *Proceedings of SPIE* 5266, pp. 129-137, http://www.f4.fhtw-berlin.de/~barthel/paper/spie03_marpe_et_al.pdf {20/01/2009}.
Martinez, J.M. (ed.) (2004) *MPEG-7 Overview (Version 10). ISO/IEC JTC1/SC29/WG11N6828*, Palma de Mallorca: ISO/IEC, http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm {28/01/2009}.

Mauthe, A. and P. Thomas (2004). *Professional Content Management Systems: Handling Digital Media Assets*, John Wiley and Sons, 314 p., http://books.google.com/books?id=Oqd5iOrv8PIC&dq=SMEF+P/Meta&hl=nl&source=gbs_summary_s&cad=0 {23/01/2009}.

McCallum, S.H. (2002). 'MARC: Keystone for Library Automation', In: *IEEE Annals of the History of Computing* 24, no. 2, pp. 34-49.

McCallum, S.H. (2004). 'An Introduction to the Metadata Object Description Schema (MODS)', In: *Library Hi Tech* 22, no. 1, pp. 82-88.

McCrory, A. and B.M. Russell (2005). 'Crosswalking EAD: Collaboration in Archival Description', In: *Information Technology & Libraries* 24, no. 3, pp. 99-107, http://mysite.pratt.edu/~croach/images/crosswal.pdf {30/01/2009}.

McGowan, J.F. (1996-2004). *Avi Overview*, http://www.jmcgowan.com/avi.html {29/01/2009}.

MeSH (2008). *MeSH: Medical Subject Headings*, National Library of Medicine, http://www.nlm.nih.gov/mesh/ {28/01/2009}.

METS Editorial Board (2007). *METS. Metadata Encoding and Transmission Standard: Primer and Reference Manual*, 129 p., METS, Digital Library Federation, http://www.loc.gov/standards/mets/METS%20Documentation%20draft%20070310p.pdf {28/01/2009}.

MIDI (2004). *Midi Manufacturers Association. Printed Documents & Online Information*, MIDI Manufacturers Association Inc., http://www.midi.org/about-midi/specshome.shtml {21/01/2009}.

MITLibraries (2007). *Metadata Reference Guide: VRA Core*, MITLibraries: Metadata Advisory Group, http://libraries.mit.edu/guides/subjects/metadata/standards/vra.html {26/01/2009}.

MovE (2009). *MovE: Musea Oost-Vlaanderen in Evolutie*, Gent: Provincie Oost-Vlaanderen, http://www.museuminzicht.be/public/index.cfm {28/01/2009}.

MPEG (2008). *The Mpeg Home Page*, MPEG, http://www.chiariglione.org/mpeg/ {19/01/2009}.

MPEG/ITEC (2005). *Information Technology - Mpeg-21 Multimedia Framework. Mpeg-21 Vision, Technology, and Strategy*, Klagenfurt: ITEC, http://mpeg-21.itec.uni-klu.ac.at/cocoon/mpeg21/ {28/01/2009}.

Multimatch (z.j.). *Schema \Multimatchmetadata-1.1.Xsd*, http://www.dcs.shef.ac.uk/~nsi/xsd1.1/default.html {28/01/2009}.

MultiMatch (2009). *Multimatch. Multilingual/Multimedia. Access to Cultural Heritage*, MultiMatch Consortium, http://www.multimatch.eu/ {28/01/2009}.

Nelson, M.L., J. Bollen, G. Manepalli and R. Haq (2005). 'Archive Ingest and Handling Test. The Old Dominion University Approach', In: *D-Lib Magazine* 11, no. 12, http://www.dlib.org/dlib/december05/nelson/12nelson.html {28/01/2009}.

Nilsson, M., M. Palmér and J. Brase (2003). 'The LOM RDF Binding - Principles and Implementation', *Paper presented at the 3rd annual ARIADNE conference*, 9 p., Leuven, november 2003, Centre for user oriented It Design, http://cid.nada.kth.se/pdf/CID-243.pdf {28/01/2009}.

Noordermeer, T. (1998). 'Depot Van Nederlandse Elektronische Publicaties', In: *Informatie Professional* 2, no. 2, pp. 22-24, Den Haag: De Koninklijke Bibliotheek, http://igitur-archive.library.uu.nl/DARLIN/2005-0520-200304/Noordermeer%2002.98.pdf {30/01/2009}.

Nussbaumer, P. and B. Haslhofer (2007). 'Cidoc Crm in Action - Experiences and Challenges', In: *Lecture notes in Computer Science*, no. 4675, pp. 532-533, http://www.springerlink.com/content/y44m33524r341776/fulltext.pdf {28/01/2009}.

OAI/ORE (2009). *Open Archives Initiative. Object Reuse and Exchange*, OAI, http://www.openarchives.org/ore/ {28/01/2009}.

OCLC (2008). *Digital Archive Service. Secure, Managed Storage for Digital Preservation*, OCLC, http://www.oclc.org/digitalarchive/ {19/01/2009}.

OCLC/CRL (2007). *TRAC: Trustworthy Repositories Audit & Certification: Criteria and Checklist*, 88 p., Dublin and Chicago: OCLC and CRL, http://www.crl.edu/PDF/trac.pdf {19/01/2009}.

Oltmans, E. and A. Lemmen (2006). 'The E-Depot at the National Library of the Netherlands', In: *The Journal for the Serials Community* 19, no. 1, pp. 61-67, http://www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/Serialsmarch2006.pdf {30/01/2009}.

Onthriar, K., K.K. Loo and Z. Xue (2006). 'Performance Comparison of Emerging Dirac Video Codec with H.264/Av', In: *Proceedings of the International Conference on Digital Telecommunications*, IEEE Computer Society, 22 p., http://ieeexplore.ieee.org/Xplore/defdeny.jsp?url=/stamp/stamp.jsp?arnumber=1698469&isnumber=35811&code=2&code=2 {19/01/2009}.

Oomen MA, O. and H. Smulders (2006). *Multimatch. D2.1 First Analysis of Metadata in the Cultural Heritage Domain*, 118 p., Nederlands Instituut voor Beeld en Geluid, http://www.multimatch.org/docs/publicdels/D2%201-FINAL-2006-23-10.pdf {28/01/2009}.

Patton, G.E. (2005). 'FRAR: Extending FRBR Concepts to Authority Data', *Paper presented at the World Library and Information Congress: 71th IFLA General Conference and Council: "Libraries - A voyage of discovery"*, 14 p., 14-18/08/2005, Oslo, OCLC Inc., http://www.ifla.org/IV/ifla71/papers/014e-Patton.pdf {29/012009}.

PREMIS Editorial Committee (2008). *Premis Data Dictionary for Preservation Metadata. Version 2.0*, 217 p., http://www.loc.gov/standards/premis/v2/premis-2-0.pdf {28/01/2009}.

PROV (2005). *Public Record Office Victoria*, Government of Victoria, http://www.prov.vic.gov.au/ {28/01/2009}.

RAMEAU (2008). *RAMEAU: Répertoire D'autorité-Matière Encyclopédique Et Alphabétique Unifié*, BnF (Bibliothèque nationale de France), http://rameau.bnf.fr/ {28/01/2009}.

RealNetworks (2009). *Realnetworks. Technology and Services That Help People Enjoy Digital Entertainment Whenever and Wherever They Want*, RealNetworks Inc., http://www.realnetworks.com/ {29/01/2009}.

RLG (2002). *RLG Best Practice Guidelines for Encoded Archival Description*, 24 p., Mountain View, California: RLG EAD Advisory Group, http://www.oclc.org/programs/ourwork/past/ead/bpg.pdf {28/01/2009}.

SantaCruz, D., T. Ebrahimi, J. Askelof, M. Larsson and C. Christopoulos (2000). 'Iso/Iec Jtc 1/Sc 29/Wg1 (Itu-T Sg8) Coding of Still Pictures (Jpeg/Jbig): An Analytical Study of Jpeg 2000 Functionalities. Jpeg 2000 Still Image Coding Versus Other Standards', In: *Proceedings of SPIE* 4115, 10 p., http://www.jpeg.org/public/wg1n1816.pdf {21/01/2009}.

Saur, K.G. (1998). *Functional Requirements for Bibliographic Records. Final Report*, 136 p., München: IFLA Study Group on the Functional Requirements for Bibliographic Records, IFLA-IFLANET, http://www.ifla.org/VII/s13/frbr/frbr1.htm {28/01/2009}.

Shepherd, E. and R. Pringle (2002). 'Mapping Descriptive Standards across Domains: A Comparison of Isad(G) and Spectrum', In: *Journal of the Society of Archivists* 23, no. 1 pp. 17-34.

Sierman, B. (2007). 'Enhancing Our Data Model with Premis', *DigCCurr 2007*, 18-20/04/2007, 7 p., Chapel Hill: Koninkijke Bibliotheek Nederland, http://ils.unc.edu/digccurr2007/papers/sierman_paper_4-1.pdf {28/01/2009}.

Simko, V. (2008). *Gama- Gateway to Archives of Media Art. D2.3 Content and Metadata Analysis Report*, GAMA.

SMPTE (z.j). *SMPTE W25.10 - Mxf Implementers Working Group*, SMPTE (Society of Motion Picture and Television Engineers), http://www.smpte-mxf.org/ {29/01/2009}.

SMPTE (2003). *SMPTE 330m-200x. Proposed SMPTE Standard. Unique Material Identifier (Umid) Version 5.e*, 21 p., SMPTE (Society of Motion Picture and Television Engineers), http://www.irmaproject.net/Members/egoray/thesaurus-dictionnaire-metadata/s330m-umid.pdf/ {26/01/2009}.

SMPTE (2005). *SMPTE Draft Recommended Practice for Television. VC-1. Bitstream Transport Encoding. SMPTE RP227*, 29 p., White Plains: SMPTE (Society of Motion Picture and Television Engineers), http://neuron2.net/misc/rp227.pdf {29/01/2009}.

SMPTE (2006). *Decoder and Bitstream Conformance. SMPTE RP228,* SMPTE (Society of Motion Picture and Television Engineers).

SMPTE (2009). *Society of Motion Picture and Television Engineers*, http://www.smpte.org/home {19/01/2009}.

SourceForge (2009). *Dirac*, SourceForge.net, http://sourceforge.net/projects/dirac {19/01/2009}.

Spicher, K.M. (1996). 'The Development of the Marc Format', In: *Cataloging and Classification Quaterly* 21, no. 3-4, pp. 75-90, http://books.google.be/books?id=R7Vu6xYwoZIC&pg=PA75&lpg=PA75&dq=Cataloging+and+Classification+Quaterly+spicher&source=bl&ots=C4HlVnmkFy&sig=dap7VAbXOnyB6qI6p0L6cHrxo6A&hl=nl&sa=X&oi=book_result&resnum=1&ct=result#PPA80,M1 {26/01/2009}.

Steenbakkers, J. F. (2000). 'Setting up a Deposit System for Electronic Publications. The NEDLIB Guidelines', *NEDLIB Report Series nr 5*, 25 p., Den Haag: Koninklijke Bibliotheek, http://nedlib.kb.nl/results/NEDLIBguidelines.pdf {30/01/2009}.

Steenbakkers, J. F. (2004). 'Treasuring the Digital Records of Science: Archiving E-Journals at the Koninklijke Bibliotheek.', In: *RLG DigiNews* 8, no. 2, http://worldcat.org/arcviewer/1/OCC/2007/08/08/0000070511/viewer/file3645.html {29/01/2009}.

Steenbakkers, J. F. (2005). 'Digital Archiving in the Twenty-First Century. Practice at the National Library of the Netherlands', In: *Library Trends* 54, no. 1, pp. 33-56, http://muse.jhu.edu/journals/library_trends/v054/54.1steenbakkers.pdf {28/01/2009 }.

Stein, R., J. Gottschewski, R. Heuchert, A. Ermert, M. Hagedorn-Saupe, H-J. Hansen, C. Saro, R. Scheffel and G. Schulte-Dornberg (2005). 'Das Cidoc Conceptual Reference Model: Eine Hilfe Für Den Datenaustausch?', In: *Mitteilungen und Berichte aus dem Institut für Meseumskunde*, no. 31, 35 p., http://www.museumsbund.de/cms/fileadmin/fg_doku/publikationen/CIDOC_CRM-Datenaustausch.pdf {30/01/2009}.

Svítek, J. (2006). 'Ogg Vorbis: Subjective Assessment of Sound Quality at Very Low Bit Rates', *CESNET technical report*, CESNET, http://www.cesnet.cz/doc/techzpravy/2006/vorbis/ {30/01/2009}.

The National Archives (2008). *ERA: Electronic Records Archives*, http://www.archives.gov/era/ {19/01/2009}.

Theora.org. (1994-2008). *Theora.Org. Theora Video Compression*, Xiph.Org, http://www.theora.org/ {29/01/2009}.

Thibodeau, K. (2007). 'If You Build It, Will It Fly? Criteria for Success in a Digital Repository', In: *Journal of Digital Information* 8, no. 2, http://journals.tdl.org/jodi/article/view/197/174 {19/01/2009}.

Unisys (2009). *About Unisys. LZW Patent Information. License Information on Gif and Other LZW-Based Technologies*, http://www.unisys.com/about__unisys/lzw {21/01/2009}.

Van der Werf-Davelaar, T. (1999). 'Long-Term Preservation of Electronic Publications. The NEDLIB Project', In: *D-Lib Magazine* 5, no. 9, http://www.dlib.org/dlib/september99/vanderwerf/09vanderwerf.html {28/01/2009}.

Van Diessen, R. J. and J. F. Steenbakkers (2002). 'The Long-Term Preservation Study of the DNEP Project: An Overview of the Results', *IBM/KB Long-term Preservation Study Report Series*, 54 p., Amsterdam: IBM / Koninklijke Bibliotheek, http://www.kb.nl/hrd/dd/dd_onderzoek/reports/1-overview.pdf {28/01/2009}.

VRA (2007a). *VRA Core 4.0*, VRA (Visual Resources Association), http://www.vraweb.org/projects/vracore4/ {26/01/2009}.

VRA (2007b). *VRA Core 4.0 Element Description*, 37p., VRA (Visual Resources Association), http://www.vraweb.org/projects/vracore4/VRA_Core4_Element_Description.pdf {26/01/2009}.

VRA (2007c). *VRA Core 4.0 Outline*, 1 p., VRA (Visual Resources Association), http://www.vraweb.org/projects/vracore4/VRA_Core4_Outline.pdf {26/01/2009 }.

Wallace, G.K. (1991). 'The JPEG Still Picture Compression Standard', In: *Communication of the ACM* 34, no. 4, pp. 31-44, http://white.stanford.edu/~brian/psy221/reader/Wallace.JPEG.pdf {30/01/2009}.

Walterus, J. (2009). *Presentatie onderzoeksresultaten project 'Van Horen Zeggen'*, FARO. Vlaams instituut voor cultureel erfgoed, http://www.faronet.be/blogs/presentatie-onderzoeksresultaten-project-van-horen-zeggen {28/01/2009}.

Weinberger, M. J. and G. Seroussi (1999). *From LOCO-I to the JPEG-LS Standard. HPL-1999-3*, 19 p., Minneapolis: Hewlett-Packard Company, http://www.hpl.hp.com/techreports/1999/HPL-1999-3.pdf {21/01/2009}.

Weinberger, M. J. , G. Seroussi and G. Sapiro (1998). *The LOCO-I Lossless Image Compression Algorithm: Principles and Standardization into JPEG-LS. HPL-98-193*, 31 p., Minneapolis: Hewlett-Packard Company, http://www.hpl.hp.com/techreports/98/HPL-98-193.pdf {21/01/2009}.

Wiegand, T., G. Sullivan, G. Bjontegaard and A. Luthra (2003). 'Overview of the H.264/AVC Video Coding Standard', In: *IEEE transactions on circuits and systems for video technology* 13, no. 7, 17 p., http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=01218189 {19/01/2009}.

Windows Hardware Developer Central (2007). *Multiple Channel Audio Data and Wave Files*, Microsoft Corporation, http://www.microsoft.com/whdc/device/audio/multichaud.mspx {21/01/2009}.

Windows Media (2004). *Advanced Systems Format (ASF) Specification*, Microsoft Corporation, http://www.microsoft.com/windows/windowsmedia/forpros/format/asfspec.aspx {29/01/2009}.

Windows Media (2009). *Windows Media Audio Codecs*, Microsoft Corporation, http://www.microsoft.com/windows/windowsmedia/forpros/codecs/audio.aspx {21/01/2009}.

Wolfgang, R. (z.j.). *JPEG Tutorial*, http://cobweb.ecn.purdue.edu/~ace/jpeg-tut/jpegtut1.html {21/01/2009}.

Wollschlaeger, T. (2006). 'ETD's as Pilot Materials for Long-Term Preservation Efforts in Kopal', In: *Paper presented at the 9th International Symposium on Electronic Theses and Dissertations*, 07-10/06/2006, Quebec, http://www6.bibl.ulaval.ca:8080/etd2006/pages/papers/SP10_Thomas_Wollschlaeger.pdf {28/01/2009}.

Xiph.Org (1994-2008a). *The Ogg Container Format*, http://www.xiph.org/ogg/ {29/01/2009}.

Xiph.Org (1994-2008b). *Speex: A Free Codec for Free Speech*, http://www.speex.org/ {29/01/2009}.

Xiph.Org (2008). *Vorbis Audio Compression*, http://www.xiph.org/vorbis/ {20/01/2009}.

Xiph.org Foundation (2008). *Theora Specifiation*, 206 p., Theora, http://www.theora.org/doc/Theora.pdf {30/01/2009}.

Xvid (2006). *Xvid*, http://www.xvid.org {19/01/2009}.

In het derde werkpakket van het project BOM-vl (Bewaring en Ontsluiting van Multimediale data in Vlaanderen, 2008-2009) staat de technische problematiek van langetermijnbewaring van digitaal erfgoed centraal. Het OAIS-model, een ISO-standaard sinds 2002, geldt hierbij als conceptueel referentiemodel dat richtlijnen biedt bij de opzet van een digitaal archief. Aan de hand hiervan werd in een eerste deliverable aangegeven met welke representatiewijzen van de data en soorten metadata men rekening dient te houden om de preservering van digitaal materiaal te garanderen en hoe men mogelijk dataverlies kan tegengaan door grondige technische overwegingen. In een uitvoerig overzicht, een state-of-the-art, komen de gangbare opslagformaten met betrekking tot verschillend audiovisueel materiaal aan bod. Vervolgens worden ook de meest courante standaarden in het bibliotheekwezen, de omroepsector, de culturele sector en de erfgoedsector besproken, in het bijzonder metadatastandaarden (descriptieve, technische, administratieve), thesauri of ontologieën en containerformaten. Ten slotte worden twee representatieve praktijkvoorbeelden toegelicht, namelijk de ontwikkeling van het e-Depot in de Koninklijke Bibliotheek van Nederland en de opzet van een Europese meertalige zoekmachine voor cultureel erfgoedonderzoek. Dit boek is de neerslag van deze deliverable en is bedoeld als referentiewerk voor alle betrokken projectpartners en spelers in het veld.

Met steun van de Vlaamse overheid

ISBN 9789052230009