Hoofdstuk 3 · Metadatastandaarden, Dublin Core en het gelaagd metadatamodel

Sam Coppens, Erik Mannens, Rik Van de Walle (Multimedia Lab UGent/IBBT)
Jan Haspeslagh, Patrick Hochstenbach, Inge Van Nieuwerburgh
(Universiteitsbibliotheek Gent)

1. *Inleiding*

Er wordt deze dagen druk gedigitaliseerd, in Vlaanderen en ver daarbuiten. Van boeken in het Google books project, over kunstvideo's uit een museum, tot de foto's van een plaatselijke heemkundige kring, we willen onze cultuur, ons erfgoed kenbaar maken aan de hele wereld, en als het eventjes kan, ook digitaal bewaren voor de toekomstige gebruikers.

Daarnaast worden dagelijks gigantische hoeveelheden digitale data aangemaakt: tekstdocumenten, rekenbladen, digitale foto, video, audio, in allerlei bekende of meer exotische formaten. En we willen ze ook delen, via het web, voor de hele wereld of alleen voor een selecte groep.

Hoe meer data aangemaakt en online beschikbaar worden gemaakt, hoe belangrijker de vindbaarheid van die data wordt, zowel voor archivering als ontsluiting van de data. Wat ben je immers met een massa gegevens op een wanordelijke hoop gegooid? Wat baat het een mooie foto van het koninklijk paleis van Brussel te 'posten' als er geen uitleg bij staat? Of een nieuwsuitzending uit 1980 aan te bieden zonder de context mee te geven? Zoals uit eerdere hoofdstukken is gebleken, maken beschrijvingen van data en context een belangrijk deel uit van de bewaring en vindbaarheid alsook de ontsluiting van digitale data. Deze metadata, data over data, helpen digitale bronnen te organiseren, ze uit te wisselen, van een digitale identificatie te voorzien en het archiveren en bewaren ervan te ondersteunen. Bij langdurige preservatie van digitale informatie zorgen metadata ervoor dat de risico's die daarmee gepaard gaan minimaal worden beschreven en eventueel beperkt. Naargelang de soort informatie die de metadata bevatten, kan men verschillende types onderscheiden: administratieve metadata (waar wordt het bewaard, tags ...), beschrijvende metadata (wat is de inhoud?), preserveringsinformatie (relocatie,

mutatie ...) en technische informatie (bestandsformaat, encryptie ...). Er zijn zelfs vele argumenten om deze metadata zelf ook te beschouwen als data, data die bovendien steeds aanpasbaar zijn. Het maakt het er niet eenvoudiger op.

Om de verschillende soorten metadata vast te leggen, gebruiken bibliotheken, musea, archieven, omroepen, etc, zeer diverse workflows en standaarden, vooral wat betreft de descriptieve metadata, die zelfs binnen de sectoren nog kunnen variëren. Dit maakt het uitwisselen en preserveren van gegevens soms heel lastig te verwezenlijken. Het is dan ook een grote uitdaging binnen het project BOM-vl een procedure vast te leggen die a) de verschillende praktijken zoveel mogelijk kan ondervangen en b) deze op een uniforme manier bewaren en ontsluiten. Daarover gaat dit hoofdstuk, het belang van (meta)datastandaarden en hoe daarmee om te gaan in uw organisatie.

2. *Verschillende sectoren, verschillende wensen, verschillende beschrijvende standaarden, nood aan een gemeenschappelijk model*

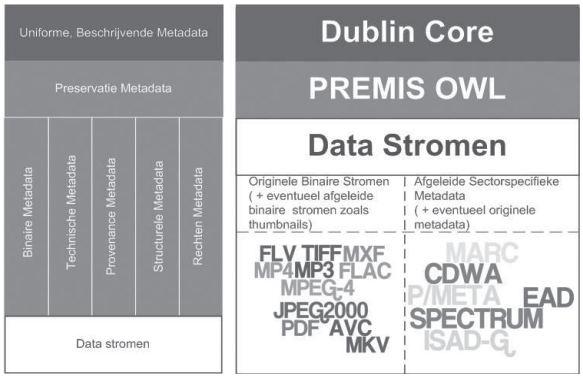
Descriptieve metadata geven een inhoudelijke beschrijving aan een werk: Wie is de auteur? Welke omroep heeft dit programma uitgezonden? Wie is er te horen in dit interview? Aan welke stad kan dit nieuwsitem gekoppeld worden?

Om data op een eenduidige manier op te stellen en uit te wisselen, worden standaarden aangemaakt, en aanbevolen te gebruiken. Standaarden brengen uniformiteit in (meta)data waardoor ze te delen zijn. De diverse sectoren in het cultuur- en erfgoedveld gebruiken verschillende descriptieve metadatastandaarden om de diversiteit aan digitale bronnen en relaties tussen deze *resources* adequaat te kunnen beschrijven. De goede standaard kiezen is lang niet evident. De metadatastandaarden verschillen namelijk in mate van detail van beschrijving, semantiek en toepassingsgebied. Een museum wil andere gegevens bewaren over een stuk uit de collectie dan een bibliotheek of een archief. Een gediversifieerde aanpak van de beschrijvingen is dus onvermijdelijk. Bovendien kunnen binnen eenzelfde sector ook nog eens verschillende standaarden worden gebruikt, en, wat ook nog veel voorkomt, eigen niet-gestandaardiseerde ordeningen.

Het gelaagd metadatamodel

Wil men echter het sectorale overschrijden, dan is een overkoepelende standaard noodzakelijk om een gemeenschappelijke beschrijving te geven zodat de records kunnen worden bevraagd en teruggevonden. Het record opslaan volgens een standaard kan echter onmogelijk de verschillende nuances weergeven en zou tot onnodig veel (meta)dataverlies leiden. Dit metadataverlies zou trouwens tot veel verzet leiden en de aanvaardbaarheid van een gemeenschappelijk model ondergraven. Om echter een betere archivering, en op termijn dus ook ontsluiting, mogelijk te maken, is een conversie naar een gemeenschappelijk model nodig. Het voorgestelde gelaagd metadatamodel moet dit probleem ondervangen. Streefdoel is namelijk 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. Met het oog op langetermijnbewaring worden de originele metadata dan ook gearchiveerd als data, zodat deze metadata nog steeds aangeboden kunnen worden aan de eindgebruiker en er geen informatieverlies optreedt.

Het gelaagd metadatamodel bestaat uit drie verschillende lagen. De toplaag biedt een eenvoudige toegang tot het archief door de beschrijving van de onderliggende data via een gemeenschappelijke standaard, Dublin Core. Deze beschrijving verwijst naar de onderliggende, gedetailleerde preservatiemodellen die alle datastromen organiseren en beschrijven, ondermeer de datastroom met de sectoraal-specifieke beschrijvingen. Ten slotte is er de basislaag, de laag waarin alle datastromen worden gearchiveerd, de bitstromen (figuur 1).



Figuur 1: Gelaagd Metadatamodel: Uniforme metadatalaag Dublin Core (maakt alle records doorzoekbaar), de preserveringsmetadatalaag (ondersteunt de langdurige preservatie) en de datastromen, onder andere de sectorspecifieke metadatalaag (belangrijk voor bewaring: bevat de detaillistische en volledige metadata).

Het gelaagd metadatamodel voorziet dus in een gemeenschappelijke standaard, Dublin Core, om een basisbeschrijving te geven van de data, om een eenvoudige toegang tot het archief te voorzien. In de onderliggende laag worden ook de sectorspecifieke beschrijvingen mee opgenomen zodat ze niet verloren gaan.

Iedere instelling beslist zelf welke specifieke metadata van belang zijn voor haar collectie. Vervolgens worden deze metadata gemapt naar de afgesproken (en door BOM-vl voorgestelde) sectorspecifieke metadatastandaarden (bijvoorbeeld MARC voor de bibliotheeksector, EAD voor de archiefsector, P/Meta voor de omroepsector, enz.), afhankelijk van de sector waartoe ze behoort of van het materiaal dat ze bezit. Ten slotte zullen de sectorspecifieke metadata, die zoals gezegd ook in het gelaagd metadatamodel opgenomen zullen worden, geconverteerd 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 (Qualified) Dublin Core voorgesteld.

Het gelaagd metadatamodel is er niet enkel om de beschrijvende metadata te stroomlijnen, meer technisch is het ook nodig om data op de drie niveaus volledig en nauwkeurig te beschrijven (drie niveaus die belangrijk zijn voor preservering van digitale objecten: preservering van het medium, preservering van technologie en preservering van de intellectuele inhoud). 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. Om echter een efficiëntere archivering mogelijk te maken, is een conversie naar een gemeenschappelijk model nodig.

3. *Sectorspecifieke standaarden*

Uit bevragingen van verschillende erfgoedinstellingen voor het project Erfgoed 2.0[1], een IBBT-project waarin de digitale interactie tussen erfgoedinstellingen in de lijn van Web 2.0 en Library 2.0 beoogd wordt, blijkt onder meer dat in de erfgoedsector een geleidelijk proces van standaardisatie aan de gang is. Dat is echter

nog lang niet voltooid. De sector zou dan ook geholpen zijn met een suggestie voor een te gebruiken standaard. Hetzelfde geldt voor de museumsector. Ook hier kan gerefereerd worden aan digitale samenwerkingsinitiatieven zoals het project MOVE[2](Musea Oost-Vlaanderen in Evolutie), waarvoor op termijn een gemeenschappelijke museumstandaard dient afgesproken te worden, alsook de projecten 'Van Horen zeggen'[3], waarin met verschillende erfgoedinstellingen onderzocht werd hoe men mondelinge bronnen kan bewaren en ontsluiten. Onder meer op basis van de genoemde projectresultaten mag men besluiten dat het vinden van een grootste gemene deler die de beschrijvingswijze van alle mogelijke materiaalsoorten dekt, een onhaalbare opgave is. Iedere sector met zijn specifieke materiaalsoorten en data stelt immers afzonderlijke eisen met betrekking tot metadata. Omdat het BOM-vl project digitale bestanden en informatie behandelt uit diverse sectoren, werd in een eerste fase een overzicht gemaakt van de werkprocessen, de gebruikte standaarden en de te behandelen multimedia in alle deelnemende sectoren. Het State of the Art (SOTA) rapport waarin de gebruikte metadatastandaarden uitvoerig beschreven en geëvalueerd worden, werd gepubliceerd (Bastijns et al., 2009)[4] en is elektronisch beschikbaar[5]. Uit dit overzicht werd een aantal standaarden geselecteerd die zijn meegenomen in het gelaagd metadatamodel. De inhoudelijke standaarden zijn MARC21 (bibliotheeksector), ISAD(G) en EAD (archiefsector), P/META (audiovisuele sector), CDWA en SPECTRUM (kunstensector en musea). PREMIS is de gekozen standaard voor preserveringsinformatie. Hier volgt een kort overzicht van de belangrijkste punten per standaard[6].

MARC21 - Bibliotheeksector

MARC is een acroniem van Machine-Readable Cataloging en wordt vooral in de bibliotheeksector gebruikt. MARC21 is een standaard voor de representatie en de communicatie van bibliografische en aanverwante informatie in computerleesbare vorm. De hoofdfunctie van de standaard is in oorsprong de geautomatiseerde uitwisseling van data tussen verschillende computersystemen. De MARC-dataelementen vormen dan ook de basis voor de meeste bibliotheekcatalogi. Een alternatief met eenzelfde graad van detail is er in deze sector niet. Om de records overzichtelijker te maken en bewerking van de records te vereenvoudigen is later de MARC XML-standaard ontworpen die MARC-records in XML-bestand voorstelt. Het is de Library of Congress die de standaard onderhoudt[7].MARC21 als bibliotheekstandaard ondersteunt niet enkel de beschrijving van boeken maar ook materialen als *Sound recording* dat alle soorten audio omvat met uitzondering van muziek (hieronder vallen dus ook mondelinge historische bronnen), alsook *motion picture* en *video recording*.

Een MARC (bibliografisch) record bestaat uit meerdere velden die verder kunnen worden onderverdeeld in subvelden. De tekstuele namen van de velden (zoals auteur en onderwerp) worden vervangen door *tags* die bestaan uit een driecijferige code. Subvelden worden gescheiden door middel van een karakter (bv. $) dat aangevuld wordt met een subveldcode die aangeeft welke gegevens volgen. Sommige velden worden verder gedefinieerd door indicatoren. Het MARC-formaat biedt op die manier een hoge mate van detail en de daarmee gepaard gaande complexiteit. Het formaat is desondanks compact. Veldnamen zoals 'plaats van publicatie' worden immers vervangen door een korte code.

MARC21 is een vrij complexe standaard, omdat de onderliggende metadata (bibliografische beschrijvingen) deze ingewikkelde structuur vereisen. De standaard heeft geen hiërarchische opbouw, noch een semantische zoekfunctie. Dit wil zeggen dat gezocht wordt naar gegeven sleutelwoorden in de verschillende velden, maar dat geen rekening wordt gehouden met de betekenis of het concept van de sleutelwoorden. Voor een leek is de informatie moeilijk leesbaar. Toch is MARC21 de beste keuze voor de bibliotheeksector, net vanwege de hoge verfijningsgraad en het algemene gebruik ervan in de sector. MARC21-records kunnen bovendien in XML gepresenteerd worden

ISAD(G) - Archiefsector

ISAD(G) staat voor General International Standard Archival Description en is een archiefstandaard die regels voorschrijft voor de beschrijving van archiefcollecties en -objecten[8]. Hierin wordt vooral een hiërarchische voorstelling van groot (een collectie) naar klein (een object) beoogd (zoals in EAD) waarbij telkens de relaties tot de andere niveaus worden aangegeven.

Deze set van algemene regels zorgt voor consistente archiefrecords die opzoekbaar zijn, uitwisselbaar en integreerbaar binnen een gemeenschappelijk informatiesysteem. De regels worden vastgelegd en beschreven in 26 elementen die gecombineerd kunnen worden om de beschrijving van een archiefbestanddeel te vormen. Gezien de grote toepassing van ISAD(G) in de Vlaamse archiefwereld, valt het aan te bevelen de ISAD(G)-EAD *crosswalk* (zie verder) te gebruiken om via de EAD *mapping* naar Dublin Core de Vlaamse archiefbestanden in het gelaagd model in te bedden.

EAD - Archiefsector

EAD, Encoded Archival Description, is een datamodel in XML dat is ontwikkeld voor het maken, opslaan, publiceren, koppelen en uitwisselen van archiefbeschrijvingen. EAD ontstond uit de behoefte nog meer informatie in te voeren dan mogelijk is in MARC[9]. De officiële EAD standaard wordt beheerd door de Library of Congress in samenwerking met de Society of American Archivists. De laatste versie en de *tag library* (overzicht van gebruikte EAD-componenten) is in nauwe samenwerking met Europese archiefinstellingen tot stand gekomen. De nationale archieven van het Verenigd Koninkrijk en Frankrijk hebben een belangrijke rol gespeeld. Inmiddels is EAD wereldwijd in gebruik. EAD is ook een belangrijk motor achter grote portaalprojecten als bijvoorbeeld Archive.org.

Omdat EAD voor en door archivarissen is ontwikkeld, is het weergeven van hiërarchie en verbanden tussen archiefstukken goed ontwikkeld. Een van de voordelen van XML is de grote flexibiliteit en migratiecapaciteit. De inhoudelijke informatie wordt in tekstformaat (bv. Unicode) opgeslagen zodat deze makkelijk gemigreerd kan worden naar nieuwe computersystemen. Ook is het mogelijk om op basis van deze XML-geformatteerde tekst meerdere zoek- en presentatietoepassingen (Word, PDF, HTML etc.) te maken, door middel van *stylesheets*.

EAD zorgt er ook voor dat samenwerking met andere archieven en deelname aan bijvoorbeeld gezamenlijke portalen mogelijk wordt, omdat inventarissen op een gestandaardiseerde wijze zijn beschreven.

EAD is een open standaard. EAD is speciaal ontworpen om de complexe hiërarchie van archiefmetadata op te vangen. De XML-gebaseerde standaard is platform- en applicatieonafhankelijk. XML maakt uitwisseling van informatie en samenwerking met andere instellingen gemakkelijker, omdat het werkt als een uitwisselformaat. Gestandaardiseerde beschrijvingen kunnen makkelijker samen doorzoekbaar gemaakt worden, en probleemloos naar Dublin Core *gemapt* worden.

In de Vlaamse archiefwereld is EAD op dit moment nog niet echt doorgedrongen. Hier wordt vooral de standaard ISAD(G) gebruikt. Er is wel een robuuste *mapping* tussen EAD en ISAD (een EAD-ISAD(G) *crosswalk*)[10]. Daarom houdt de sector best ISAD(G) als standaard voor metadatabeheer aan en, afhankelijk van hoe de partner het organiseert, gebruikt men de *crosswalk* ISAD(G)-EAD en de *mapping* naar Dublin Core, ofwel de rechtstreekse ISAD(G)-DC *crosswalk*.

P/META - Omroepsector

P/Meta[11] is een standaard die ontstaan is in de schoot van de European Broadcasting Union voor de uitwisseling van informatie van en over programma's. 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.

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. Hieraan wordt gerefereerd met *identifiers* en *names*.

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. Deze concepten worden verder beschreven met attributen (met *authority lists* voor elk soort attribuut). Verder worden in P/Meta ook P/Meta sets gedefinieerd, bestaande uit P/Meta attributen die met andere P/Meta sets gegroepeerd worden. Deze voorgedefinieerde sets vormen de bouwstenen voor de meest voorkomende informatie-uitwisselingen, maar er kunnen ook heel specifieke P/Meta sets voor speciale uitwisselingen geschreven worden.

P/Meta is hoofdzakelijk van toepassing in een business-to-businessomgeving (B2B). Ondanks die focus wordt de wisselwerking met business-to-consumermetadata(B2C) beschouwd als een elementaire eigenschap van P/Meta. Opteren voor een internationaal genormeerde specificatie als P/Meta vereenvoudigt het definiëren van de semantiek, de syntaxis en de schrijfwijze van de taal waarmee tussen verschillende actoren over programma's gecommuniceerd kan worden. Tevens wordt het eenvoudiger om op de metadata uitwisselingsschema's in XML te enten. Omdat alle partners uit de omroepsector het IPEA-model[12], dat een subset is van P/Meta, reeds gebruiken is het uiteraard logisch om deze standaard op te nemen in het gelaagd metadatamodel.

CDWA - Kunstensector

Categories for the Description of Works of Art (CDWA) beschrijft de data uit kunstdatabanken aan de hand van een conceptueel raamwerk voor het beschrijven en opvragen van informatie over kunstwerken, architectuur of ander cultureel materiaal[13]. Vooral de cultuursector hanteert daarom deze standaard.

Het CDWA bevat 512 categorieën en subcategorieën. Deze categorieën stellen de minimale informatie voor die nodig is om een werk te beschrijven en te identificeren. Daarbuiten bevat de CDWA ook *discussies*, basisregels voor het catalogiseren en voorbeelden. De categorieën leveren een raamwerk waarop bestaande informatiesystemen kunnen *gemapt* worden en op basis waarvan nieuwe systemen ontwikkeld kunnen worden. Daarbij identificeren de *discussies* in het CDWA woordenschatten en beschrijvende toepassingen die de informatie in de verschillende systemen meer compatibel en meer toegankelijk maken.

Het CDWA stelt een relationele datastructuur voor, waar records over objecten of werken aan elkaar gelinkt zijn met hiërarchische relaties. Het CDWA raadt ook aan om aparte files bij te houden voor gerelateerde visuele werken, gerelateerd tekstueel materiaal, personen en bedrijveninformatie, locaties en dergelijke.

Op basis van CDWA en CCO (Cataloguing Cultural Objects)[14] werd CDWA Lite ontwikkeld, een XML-schema dat de basiselementen bevat voor de beschrijving van een kunstwerk of cultureel materiaal. CDWA Lite-records zijn een goed instrument tot uitwisseling van data voor bibliotheken die gebruik maken van het OAI-PMH-protocol[15].

Het gebruik van het CDWA raamwerk draagt bij tot de integriteit en de levensduur van de data en vergemakkelijkt de migratie van de data naar nieuwe systemen. Bovenal helpt het de eindgebruiker in het opzoeken van betrouwbare informatie, ongeacht het systeem waarin de data zijn opgeslagen. De bruikbaarheid van CDWA in OAI-systemen, en de beschikbaarheid van een XML-schema (CDWA-Lite) maken deze standaard onmiddellijk inzetbaar in een gelaagd metadatamodel voor de kunstensector.

SPECTRUM - Musea

SPECTRUM is een door de *British Museum Documentation Association* beheerde standaard voor het professionaliseren van de museale bedrijfsvoering. SPEC-

TRUM is ontstaan vanuit de visie dat iedere handeling rond een museaal object te bewaren informatie is.

Elke handeling in een museum geschiedt op grond van informatie en genereert informatie. Alle handelingen met betrekking tot museumstukken, van verwerving tot en met tentoonstelling, zijn vervat in 21 procedures. Naast deze procedures bevat SPECTRUM ook een overzicht van alle informatie die in het museum vastgelegd moet worden, om de procedures goed te kunnen toepassen. SPECTRUM is dus in de eerste plaats een set richtlijnen.

SPECTRUM kan toegepast worden in software voor museaal collectiebeheer maar is zelf geen softwarepakket. Het is de bedoeling dat leveranciers van museale software gebruikmaken van SPECTRUM, zodat hun software zo goed mogelijk ruimte biedt voor het vastleggen van de benodigde informatie.

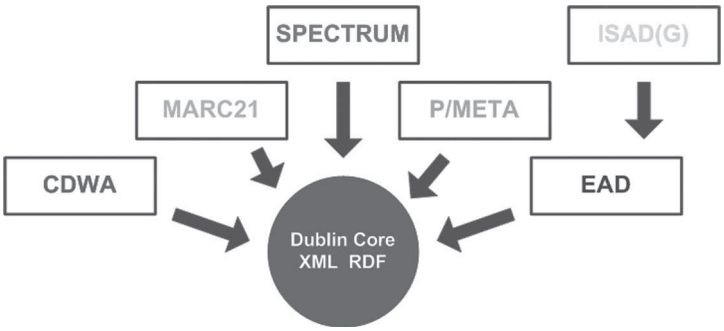
In 2007 werd SPECTRUM in Nederlandse versie zowel in Vlaanderen als Nederland geïntroduceerd. De softwareontwikkelaar Adlib en het Koninklijk Instituut voor het Kunstpatrimonium hebben de SPECTRUM standaard ondertussen gedeeltelijk in hun processen geïmplementeerd.
Zoals het geval is bij de andere behandelde standaarden, bestaat er ook voor SPECTRUM een XML-schema.

In tegenstelling tot de 5 andere standaarden is er een duidelijk contextueel verschil tussen de SPECTRUM standaard en Dublin Core. Aanvankelijk werd Dublin Core ontwikkeld als een schema dat toelaat om digitale online '*resources*' via gestructureerde metadata te documenteren en terug te vinden. Alle hiervoor behandelde metadastandaarden hadden te maken met informatie over 'fysieke' resources en hun 'content'.

SPECTRUM beheert echter de informatie over handelingen met betrekking tot 'fysieke' resources waarbij de handelingen belangrijker zijn dan de resources (museumstukken) zelf. Er zijn heel veel elementen in SPECTRUM die verwijzen naar begrippen zoals *acquisition, entry, disposal, conservation, ownership, exhibition, access, loan, location, movement, handling, place, procedure, request, use, return*, enzovoort. Dat die informatie essentieel zal zijn bij het invullen van SPECTRUM-records is evident. Voor een goede uitwisseling met Dublin Core zal het belangrijk zijn om de elementen uit het SPECTRUM-record die naar de resource (het museumstuk) zelf verwijzen over te nemen. Dit gedeelte van een SPECTRUM-record heeft grotendeels dezelfde context als een Dublin Core elementenset en kan dus perfect uitgewisseld worden.

4. *Dublin Core als sectoroverschrijdende metadatalaag*

Buiten het gelaagd metadamodel voor langdurige preservatie, levert het BOM-vl project ook een crosswalk aan die de sectorspecifieke metadatastandaarden mapt naar de Dublin Core ontologie. De uitgewerkte *mapping* van de verschillende metadatastandaarden naar Dublin Core maakt het mogelijk om het metadatagedeelte van het model heel eenvoudig voor te stellen. Figuur 2 geeft dit grafisch weer. De containers bevatten de records in hun oorspronkelijk formaat. De pijlen bevatten de crosswalks van de metadatavelden voor elke specifieke standaard naar de Dublin Core velden en vormen dus de basis om de records te ontsluiten als Dublin Core XML-records of Dublin Core RDF-records. Deze records kunnen dan online beschikbaar worden gesteld, en kunnen zo doorzocht en getoond worden. Aangezien het gelaagd metadatamodel uitgedrukt is met behulp van technieken die in het semantisch web worden gebruikt en dus beschreven is met behulp van RDF, kiest men er best voor de records als Dublin Core RDF-records te ontsluiten om ze te kunnen incorporeren in het gelaagd metadatamodel (figuur2).



Figuur 2: Crosswalks sectorspecifieke metadatastandaarden naar Dublin Core

Dublin Core als uniforme beschrijvende standaard

Dublin Core (DC) is een wijdverspreide, sectoroverschrijdende standaard, een soort lingua franca van metadata. Zijn eenvoud en algemene toepasbaarheid bepalen zijn succes. Bijna iedere standaard kan gemapt worden naar DC. De eenvoud beperkt echter de mate van detail van de beschrijvingen. Daarom dat de standaard meestal gebruikt wordt naast meer sectorspecifieke standaarden.

De DC-standaard bestaat uit twee niveaus:

Simple Dublin Core bevat 15 elementen die facultatief en herhaalbaar zijn: *contributor, coverage, creator, date, description, format, language, identifier, publisher, relation, rights, source, subject, title* en *type*. *Qualified Dublin Core* voegt hier 3 elementen aan toe (*audience, provenance* en *rights holder*), en voorziet met qualifiers in een verfijning van de DC-elements. Het principe bij het specificeren van de DC-elementen stelt dat een toepassing die de specificatie, de qualifier, bij het DC-element niet 'begrijpt', deze moet kunnen negeren en het DC-element zelf moet kunnen behandelen alsof het een unqualified (breder) element zou zijn.

Qualified Dublin Core bevat bovendien een aantal 'Encoding Schemes' die een DC-element op een specifieke manier kunnen definiëren. Deze schema's kunnen woordenlijsten zijn, thesauri, notatieregels, lijsten van toegelaten waarden, enzovoort. Een waarde die in een DC-element van zo een schema gebruikmaakt, kan bijvoorbeeld een trefwoord uit een gecontroleerde woordenlijst zijn, of een formele notatie volgens welbepaalde regels (bv. '2008-12-31' volgens het schema 'YYYY-MM-DD'). Als een bepaalde toepassing het schema niet 'begrijpt', blijft de gebruikte metadatawaarde wel nog leesbaar en begrijpelijk voor mensen.

Door al deze eigenschappen is Dublin Core zeer geschikt om meer complexe metadataschema's te condenseren tot de 15 elementen die de essentiële informatie van het object bevatten. Via XML- en RDF-gebaseerde uitvoer blijft de metadata perfect opzoekbaar in alle digitale systemen.

Een voorbeeld van een Dublin Core RDF-record is:

```
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://hdl.handle.net/2147/173'
        dc:title[productiefoto] Pas de deux/dc:title
        dc:formatimage/tiff/dc:description
        dc:date2005-09-28T08:30:28Z/dc:date
dc:subjectHugo Claus; Nederlands Toneel Gent;1974- 1975/dc:subject
        dc:contributorHugo Claus/dc:contributor
dc:descriptionScan van foto uit collectie Frans Verreyt/dc:description
/rdf:Description
/rdf:RDF
```

Figuur 3: Voorbeeld Dublin Core RDF-record

Door de eenvoud en het beperkt aantal velden van Dublin Core dreigt er een verlies van detail te ontstaan. Dit wordt opgevangen door de velden van andere standaarden te mappen aan Dublin Core. De sectorspecifieke metadatastandaard kan zo verder worden ingevuld met meer detail, zoals vroeger, en het *mapping* schema beschrijft naar welk DC-element de eigen metadatavelden vertaald worden.
Onderstaand voorbeeld illustreert enkele *gemapte* velden van de ISAD(G)-standaard voor de archiefsector naar Dublin Core:

ISAD(G)-veld	*Dublin Core element*
3.2.1 Name of creator	dc:creator
3.2.2 Administrative/biographical history	dc:description
3.2.3 Archival history	dc:provenance

Een goed gedocumenteerde *mapping* zorgt ervoor dat er weinig of geen metadata-informatie verloren gaat tijdens de overdracht van een origineel record naar een Dublin Core record. Voor iedere afzonderlijke *mapping* wordt een aantal essentiële of core-elementen/velden aangeduid, die samen in het DC-record nog voldoende informatie bevatten om het record te 'begrijpen' en terug te linken naar de volledige oorspronkelijke set. Dit ziet er voor enkele velden uit de CDWA-standaard bijvoorbeeld zo uit (waarbij 7 CDWA-velden gecompileerd worden in 3 DC-elementen):

CDWA-veld	Dublin Core element	DC Qualifier
1.1. Catalog Level (core)	dc:type	
1.2. Object/Work Type (core)	dc:type	
1.3. Object/Work Type Date	dc:date	
1.3.1. Earliest Date	dc:date	Created
1.3.2. Latest Date	dc:date	
1.4. Components/Parts	dc:format	Extent
1.4.1. Components Quantity	dc:format	Extent

Dublin Core kiezen als basis voor het gelaagd metadatamodel heeft een aantal doorslaggevende voordelen:

- Dublin Core voorziet in een 'grootste gemene deler' voor de metadatastandaarden die in verschillende sectoren worden toegepast, en vereenvoudigt sterk de onderlinge informatie-uitwisseling en zoekopdrachten.
- De *qualifiers* vangen het nadeel op van het beperkt aantal DC-elementen.
- De standaard ondersteunt RDF-gebaseerde opslag.
- Ten slotte wordt Dublin Core (en de XML/RDF toepassingen ervan) nu al wereldwijd in de meeste online informatiesystemen toegepast.

Aan deze keuze zijn toch ook enkele nadelen verbonden:

- Dublin Core beperkt zich tot de beschrijving van resources zoals boeken en geluidsfragmenten maar ondersteunt niet de beschrijving van personen en instellingen. Dit verlies van informatie wordt grotendeels opgevangen door een goed gedocumenteerde *mapping* naar de oorspronkelijke metadataset.
- Dublin Core beschrijft voornamelijk het voorwerp zelf en slechts in beperkte mate het uitgebeelde of beschreven onderwerp.
- Verschillende interpretaties van eenzelfde element kunnen leiden tot 'vertaalproblemen'. In het gelaagd metadatamodel zullen alle mogelijke interpretaties voor de verschillende sectoren gedocumenteerd moeten worden.

De mapping naar Dublin Core

Het bovenstaande is een lange inleiding om te komen tot de opzet, het ontwikkelen van een gelaagd metadatamodel waarbij een brug geslagen moest worden tussen de 6 gekozen standaarden en Dublin Core. In haar eenvoudigste vorm bestaat dergelijke operatie uit het selecteren en groeperen van contextueel verwante elementen uit de standaard, om dan deze groepen vervolgens te laten verwijzen naar een van de 15 DC-elementen. Bij de uitwerking van de mapping werden, voor de 6 standaarden, alle metadatavelden (en eventueel subvelden) gemapt naar een DC-element. Contextueel of inhoudelijk verwante velden komen op die manier in groepen terecht met dezelfde Dublin Core-verwijzing. Deze mapping is een crosswalk van elke besproken metadatastandaard naar Dublin Core. De crosswalk mapt de semantiek van de metadatavelden van de sectorspecifieke metadatastandaarden naar de Dublin Core elementen en dient als basis voor de mapping naar het semantisch, gelaagd metadatamodel.

Per standaard werd een set van essentiële velden geselecteerd die altijd ingevuld zijn (metadata bevatten) en *gemapt* worden naar Dublin Core. De 'essentiële veldenset' moet het record voor de eindgebruiker 'leesbaar' houden, en eenduidige, zij het minimale informatie bevatten. Om vervolgens het oorspronkelijke record met alle metadata te kunnen oproepen, moet een van de essentiële velden de unieke recordidentificatie bevatten en gelinkt worden aan dc:identifier.

5. *PREMIS OWL: de preservatielaag*

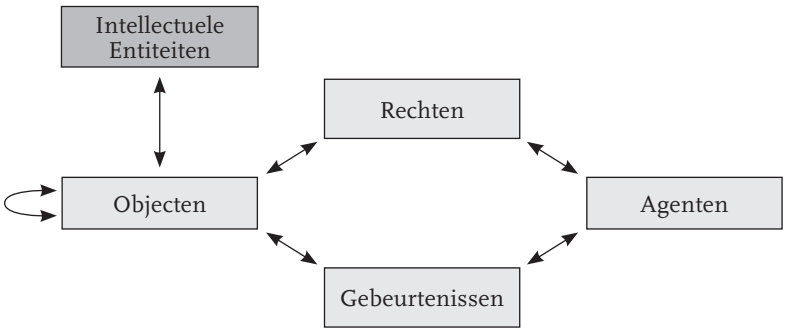
Het louter beschrijven van de intellectuele inhoud van audiovisuele data, door middel van een Dublin Core schema, is onvoldoende voor langdurige preservatie. Het archief moet de data niet alleen opslaan en opzoekbaar maken, het moet er ook voor zorgen dat de data op lange termijn toegankelijk en interpreteerbaar blijven. Hier zijn vele risico's aan verbonden die op verschillende niveaus meespelen: datadragers, zoals tapes of harde schijven slijten, bestandsformaten verouderen, de terminologieën en organisatiestructuren veranderen. Om deze risico's te minimaliseren, moeten de audiovisuele data niet alleen beschreven worden door descriptieve metadata, maar ook door binaire metadata, technische metadata, preserveringsmetadata, structurele metadata en rechten metadata. Meer technisch is het gelaagd metadatamodel nodig om data op drie niveaus volledig en nauwkeurig te beschrijven, drie niveaus die belangrijk zijn voor preservering van digitale objecten: preservering van het medium, preservering van technologie en preservering van de intellectuele inhoud. Volgende metadataschema's zijn daarin te onderscheiden:

- 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, 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.

Het semantisch, gelaagd metadatamodel, voorgesteld door BOM-vl, bestaat uit twee lagen: de toplaag beschrijft de intellectuele inhoud van het object via Dublin Core, de onderste laag verzorgt de specifieke metadata die nodig zijn voor langdurige opslag van audiovisuele informatie. Voor deze preservatielaag werd een ontologie ontworpen op basis van de PREMIS 2.0 metadatastandaard, PREMIS OWL. PREMIS is een preserveringsstandaard gebaseerd op het OAIS-referentiemodel en is beschreven via een datamodel dat bestaat uit vijf semantische eenheden:

1. Intellectuele entiteiten: Deze beschrijven de intellectuele inhoud van een audiovisueel object, bijvoorbeeld een boek of foto. Dit wordt beschreven door descriptieve metadata, hier de Dublin Core ontologie.
2. Objecten: Dit is een discrete eenheid van informatie in digitale vorm. Deze discrete eenheid kan een bestand zijn, een *bitstream*, die de eigenlijke data bevat binnen een bestand, of een representatie, een verzameling van bestanden die samen de intellectuele entiteit vormt. Bijvoorbeeld een digitaal boek kan bestaan uit een verzameling geordende TIFF-beelden. Deze entiteit verzorgt de binaire metadata, de technische metadata en de structurele metadata.
3. Gebeurtenissen: Deze entiteiten beschrijven de acties die een impact kunnen hebben op een object of agent, bijvoorbeeld de migratie van een bestandsformaat naar een ander formaat.
4. Rechten: Deze klasse beschrijft een of meer rechten van een object of agent.
5. Agenten: Dit kan een persoon, organisatie of software applicatie zijn die gerelateerd is aan een gebeurtenis, een object of de rechten van een object.

Intellectuele entiteiten, gebeurtenissen en rechten zijn direct gerelateerd aan een object. Een agent kan enkel gerelateerd zijn aan een object via een gebeurtenis of via rechten. Op deze manier worden niet alleen de veranderingen opgeslagen van een object, maar ook de gebeurtenis die de verandering tot stand bracht. Deze relaties zorgen ervoor dat de oorsprong van een object kan worden nagegaan, en leveren de preserveringsmetadata. De onderstaande figuur verduidelijkt het datamodel van PREMIS.



Figuur 4: Datamodel van PREMIS

Daarbuiten zijn er ook nog descriptieve metadata nodig om de intellectuele inhoud van een object te beschrijven. Deze toplaag van het gelaagd metadatamodel maakt de informatie opzoekbaar en beheerbaar. De Dublin Core ontologie neemt deze rol op zich.

6. *Besluit*

Als Vlaanderen een optie wil nemen op het beschermen en langdurig bewaren van het Vlaamse erfgoed, dan zijn centralisatie en expertdeling belangrijke elementen om in dit opzet te slagen. BOM-vl heeft met het gelaagd metadatamodel een van de basiselementen opgesteld waarmee gegevensuitwisseling kan worden verzekerd, de gedetailleerde beschrijving van een digitaal object kan worden bewaard en een preserveringslaag wordt aangereikt die de data opent voor toekomstige gebruikers.

EINDNOTEN

1. Zie projectwebsite IBBT: http://www.ibbt.be/nl/project/heritage-20-0 (24 jan. 2010)
2. Zie http://musea.oost-vlaanderen.be/public/index.cfm (24 jan. 2010)
3. Mannens E., Paridaens T., Hauttekeete L., Evens T., Gysels J. (2007). *Onderzoeksproject `Van Horen Zeggen fase III`. Haalbaarheidsstudie naar een innovatieve applicatie voor de ontsluiting van mondelinge bronnen*. Universiteit Gent – IBBT. http://hdl.handle.net/1854/LU-622761
4. Zie het boek dat het resultaat is van dit onderzoek: Bastijns, P. e.a. (2009*). (Meta)datastandaarden voor digitale archieven*. http://hdl.handle.net/1854/LU-480734
5. Zie het boek dat het resultaat is van dit onderzoek: Bastijns, P. e.a. (2009). *(Meta)datastandaarden voor digitale archieven*. http://hdl.handle.net/1854/LU-480734
6. Uitgebreide uitleg: Zie het boek dat het resultaat is van dit onderzoek: Bastijns, P. e.a. (2009). (*Meta)datastandaarden voor digitale archieven*. http://hdl.handle.net/1854/LU-480734
7. MARC standards: http://www.loc.gov/marc/
8. Zie ICA (de referentiepagina van ISAD(G)). http://www.ica.org `
   Voor crosswalks, zie LOC (1999) (met EAD) en Shepherd (vergelijking met SPECTRUM).
9. Zie LOC voor de homepage van EAD, LOC voor voorbeelden van EADrecords in XML, DEN (toelichting door DEN bij EAD).
10. Zie website Library of Congress: http://www.loc.gov/ead/tglib/appendix_a.html
11. Zie ook de documenten op de website van EBU: http://tech.ebu.ch/metadata/p_meta
12. Zie de projectwebsite IBBT (2007a) en het rapport Hauttekeete, L., Dekeyser, H. e.a (2006). *Digitale archivering op nationaal en internationaal vlak: een stand van zaken*. http://hdl.handle.net/1854/LU-622897
13. Zie Getty (2006) voor een overzicht van de elementen van CDWA: http://www.getty.edu/research/conducting_research/standards/cdwa/
14. CCO en CDWA Lite: http://www.vraweb.org/ccoweb/cco/about.html
15. Voluit The Open Archives Initiative Protocol for Metadata Harvesting. http://www.openarchives.org/