

Een gelaagd semantisch metadatamodel voor langetermijnarchivering

Erik Mannens, Sam Coppens, Rik Van de Walle, IBBT-UGent-MMLab

1. Inleiding

Volgens gegevens van UNESCO [1] zou onze planeet al ettelijke miljoenen uren aan audiovisuele archieven bezitten. In Vlaanderen beschikken niet alleen de omroepen, maar ook culturele organisaties, privépersonen en overheidsinstellingen over duizenden uren aan spraak- en beeldmateriaal, dat op analoge dragers is opgeslagen. Op dit ogenblik boeten die dragers echter voortdurend aan kwaliteit in, waardoor de banden op termijn niet langer af te spelen zullen zijn en de gegevens dus ontoegankelijk zullen worden. De digitale ontsluiting en opslag van dit materiaal is slechts een tijdelijke oplossing. Het digitaal materiaal moet immers ook over 20, 50 of 100 jaar nog intact en raadpleegbaar zijn. Digitale langetermijnarchivering biedt hier uitkomst.

Het snelle verval van het audiovisuele erfgoed enerzijds en de maatschappelijke, culturele en economische waarde ervan anderzijds, vragen een oplossing op korte termijn. Het project BOM-vl (Bewaring en Ontsluiting van Multimediale Data in Vlaanderen) [2] levert een belangrijke aanzet om digitale langetermijnarchivering van het cultureel erfgoed in Vlaanderen te realiseren en de daarmee gepaard gaande, hieronder vermelde onderzoeksproblemen te analyseren. In dit artikel wordt een gelaagd semantisch metadatamodel voorgesteld dat in het kader van het project werd opgesteld en dat als mogelijke deeloplossing fungeert.

In het tweede hoofdstuk komen enkele specifieke problemen van digitale archivering aan bod, zoals het behoud van de authenticiteit van digitale informatie. In hoofdstuk drie wordt verder ingegaan op de risico's van digitale langetermijnarchivering, waaronder datacorruptie van de dragers. Om alle risico's tot een minimum te beperken, moeten de eigenlijke data op verschillende niveaus beschreven worden, vanaf de bits tot het niveau van een intellectuele entiteit. Op elk niveau bestaat er immers een mogelijk risico op gegevensverlies. De verschillende niveaus zijn gebaseerd op het referentiemodel Open Archival Information System (OAIS, [3]). In het vierde hoofdstuk wordt het gelaagd semantisch metadatamodel voorgesteld en toegelicht. Het model bestaat uit twee lagen: de bovenste laag levert de beschrijvende metadata en de onderste laag is verantwoordelijk voor de binaire metadata, de technische metadata, de structurele metadata, de conserveringsmetadata en de rechtenmetadata. Op die manier worden alle nodige metadata voor de beschrijving op de verschillende niveaus gedekt. Voor de bovenste laag wordt een Web Ontology Language (OWL, [4]) toegepast, een versie van Dublin Core [5]. Voor de onderste laag werd in het project BOM-vl een OWL-representatie van de conserveringsstandaard PREMIS 2.0 [6] ontwikkeld. Dit PREMIS OWL-schema wordt aangevuld met een thesaurus die de wettelijke taken beschrijft die een persoon, organisatie of softwaretoepassing kan hebben. Het laatste hoofdstuk vat de bevindingen samen en verwijst naar een uitvoeriger referentiewerk.

2. Problemen

Behalve conservering biedt een digitaal archief tal van voordelen: ruimtelijke grenzen vervagen, mobiliteit vormt geen obstakel meer en het doorzoeken van het archief wordt veel gemakkelijker. Maar een digitaal archief is niet onkwetsbaar. Bij bewaring op lange termijn van digitale multimediale gegevens moet het digitale archief daarom aan een aantal specifieke eisen voldoen. Allereerst moeten de software en hardware van het digitale archief gedurende een lange tijd toegang tot de informatie garanderen. Vervolgens is er nog steeds menselijke interventie nodig in de vorm van

archiefbeschrijvingen, werkprocessen en het gebruik van standaarden om de informatie zo lang mogelijk toegankelijk en interpreteerbaar te houden voor de gebruikersgemeenschap.

Digitale informatie staat aan veel gevaren bloot. Sommige daarvan bedreigen ook analoge documenten, andere kenmerken in het bijzonder digitale informatie:

- In digitale vorm is informatie een conceptueel object. Digitale multimedia kunnen gemakkelijk gekopieerd en gewijzigd worden zonder onmiddellijk zichtbare effecten op de representeerbare inhoud. In vergelijking met analoge informatie is het daarom moeilijk de authenticiteit van informatie in digitale vorm te bewaren. Een van de hoofdbekommernissen bij langetermijnarchivering is dan ook de blijvende garantie van de authenticiteit van de data.
- Een andere bedreiging voor digitale gegevens komt voort uit technologische veranderingen. Dataformaten en afgeleide formaten evolueren snel. Ze kunnen achterhaald worden of in de toekomst niet meer interpreteerbaar zijn. Ook de levensduur van de opslagtechnieken zelf is eindig. Om die discrepantie tussen de korte levensduur van de digitale technologie en de noodzaak aan langdurige bewaring weg te werken en dus de toegang tot de informatie te blijven garanderen, is migratie van het oude dataformaat naar het nieuwe dataformaat nodig of moet er een blijvende emulatie van dat oude dataformaat voorzien worden. Technische metadata moeten bovendien voldoende informatie over de opgeslagen gegevens aanreiken om snelle interventies mogelijk te maken.
- Op lange termijn kan het kennisdomein van de gebruikersgemeenschap veranderen, dataspecialisten komen en gaan, de instellingen zelf kunnen wijzigen of een nieuwe taak krijgen. Dit leidt mogelijk tot interpretatieproblemen. De opgeslagen data moeten dus ook voldoende contextuele metadata bevatten, zodat ook nieuwe en toekomstige gebruikersgroepen de informatie kunnen interpreteren.

Bij het opslaan van digitaal cultureel erfgoed uit verschillende sectoren - de omroepen, de bibliotheken, de culturele sector en de archiefinstellingen - moet het digitale archief een grote hoeveelheid beschrijvende metadata verwerken. Elk specifiek toepassingsgebied bepaalt welke beschrijvende metadata noodzakelijk zijn. Digitale beelden in een bibliotheek kunnen een gescand boek betreffen, terwijl beelden in het bezit van een museum mogelijk een kunstwerk voorstellen. Beide afbeeldingen vragen bijgevolg andere beschrijvende metadata velden. Het digitale archief moet in staat zijn om in deze heel uiteenlopende dataset te zoeken. Daarom is een overkoepelend metadata-schema nodig.

3. Metadataniveaus voor langetermijnarchivering

Volgens het OAIS-referentiemodel heeft een archief verschillende types metadata nodig om informatie volledig te beschrijven met het oog op zijn langetermijnarchivering. De data moeten daarom op verschillende niveaus beschreven worden. Zoals hierboven aangehaald, bestaat op elk niveau het gevaar dat er data verloren gaan.

Op het laagste niveau is een digitaal bestand opgebouwd uit bits en bytes die op hardware-systemen 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 nodig om fouten te herstellen. Zogenaamde 'houdbaarheids'-metadata (binaire metadata) dragen hiertoe bij (zie ook het volgend hoofdstuk).

De authenticiteit van de digitale informatie is aan grotere gevaren onderhevig dan van data in analoge vorm. In het laatste geval volstaat een beschrijving van alle kenmerken van het fysieke object, maar van digitale informatie moet de hele ontstaans- en verwerkingsgeschiedenis gearchiveerd worden.

Op een hoger niveau moeten de bits en bytes in de vorm van een digitaal bestand de opgeslagen data representeren. Bestands- en compressieformaten zoals AVI [7], MP3 [8] en JPEG [9] beschrijven hoe de bits en bytes getransformeerd moeten worden naar een interpreteerbare multimediale representatie zoals video, geluid en afbeeldingen. Ook die bestandsformaten hebben een beperkte 'houdbaarheidsdatum'. Wanneer een bestandsformaat in onbruik raakt, zijn er twee manieren om de opgeslagen data in het archief te preserven, namelijk via migratie of emulatie zoals hierboven vermeld. Beide mogelijkheden hebben voor- en nadelen. Migratie kan dataverlies tot gevolg hebben en een emulatieproces kan heel complex zijn. Om de bestandsformaten in de toekomst begrijpelijk en migreerbaar te houden zonder dataverlies, zijn open standaarden onontbeerlijk. Om gesloten formaten te interpreteren, is immers bijhorende gesloten software nodig en de archivering hiervan is veel moeilijker dan bij het gebruik van open standaarden. Ook dit aspect wordt best goed overwogen bij de keuze voor een bepaald formaat.

Bestandsformaten beschrijven de representatie van de opgeslagen data. Bij multimediale data is het echter niet alleen belangrijk om de informatie zelf te preserven, maar ook de originele *look and feel* moet behouden blijven. Bij de migratie van bestandsformaten kunnen de resolutie of kleurwaarden veranderen, waardoor vitale informatie verloren gaat. Rijke beschrijvingen van de *look and feel* moeten dit vermijden.

Op een nog hoger niveau is contextuele informatie onmisbaar voor de interpretatie van het digitale bestand. Op lange termijn zullen de producenten van de informatie niet meer beschikbaar zijn om de gearchiveerde dataset toe te lichten. Een datacollectie moet van voldoende contextuele data vergezeld worden zodat de informatie zonder hulp van externe experts voor een welomschreven doelpubliek of *Designated Community* [3] 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 bijgevolg noodzakelijk om ook de semantiek van de metadata op te slaan zodat de data later juist geïnterpreteerd kunnen worden. Ook het beheer van rechten speelt zich op dit niveau af. Rechten op een digitaal object veranderen in de loop van de tijd. Om deze veranderingen bij te houden, moet de betreffende rechtensituatie minutieus beschreven worden.

4. Gelaagd semantisch metadataschema

Preservering van digitale objecten moet op drie erkende conceptuele niveaus gebeuren: preservering van het medium, preservering van de technologie en preservering van de intellectuele inhoud. Een gelaagd metadatamodel is nodig om de data op deze drie niveaus volledig en nauwkeurig te beschrijven. De hieronder beschreven 6 verschillende lagen zorgen er uiteindelijk samen voor dat digitale objecten volledig op deze drie niveaus kunnen worden beschreven, maar vormen 'an sich' echter geen 1-op-1 mapping met deze drie niveaus; verschillende lagen samen vormen één enkel niveau of één enkele laag kan zich over verschillende niveaus uitspreiden:

1. Binaire metadata

Deze schema's beschrijven de data op bitniveau, de zogenaamde bitstromen. Die bitstromen zijn de feitelijke gegevens in een bestand. De metadata moeten de informatie toegankelijk houden door aan te geven op welke manier de bits omgevormd kunnen worden tot een representatie, bijvoorbeeld in een bepaald compressieformaat.

2. Technische metadata

Deze metadata beschrijven de data op bestandsniveau. Dataformaten en hun afgeleiden evolueren snel. Formaten verouderen en blijven daardoor niet interpreteerbaar. Om de informatie toegankelijk te houden, is migratie of emulatie nodig. De technische metadata, bijvoorbeeld met betrekking tot het bestandsformaat, moeten dergelijke ingrepen ondersteunen.

3. Structurele metadata

Een mogelijke representatie van de intellectuele inhoud van de data is een reeks bestanden. Structurele metadata beschrijven de verbanden tussen dergelijke sets van bestanden. Een boek kan bijvoorbeeld worden gerepresenteerd door Tagged Image File Format (TIFF, [10]) in afbeeldingen van de gescande boekpagina's. Voor een volledige beschrijving van het boek zijn structurele metadata nodig met het oog op een correcte ordening van deze TIFF-beelden.

4. Beschrijvende metadata

Bij de archivering van digitaal cultureel erfgoed uit verschillende sectoren - de omroepen, de bibliotheken, de culturele sector en de archiefinstellingen - ontstaat een bijkomend probleem met betrekking tot de beschrijvende metadata. Veel instellingen beschikken al over beschrijvende metadata volgens een eigen systeem of schema, zoals bijvoorbeeld ISAD(G). Ze bewaren die gegevens als metadata, maar eventueel ook als data. De beide strategieën hebben voor- en nadelen. Indien de beschrijvingen als metadata worden bewaard, kiest het digitaal archief ook voor een bepaald metadata-schema, maar die keuze is geen sinecure. De gebruikte metadata-schema's zijn immers heel domeinspecifiek. Voor het verliesloos opslaan van de beschrijvende metadata zou het archief voor een beschrijvend metadata-schema moeten opteren dat een kleinste gemene veelvoud is van de beschrijvende metadata-schema's die door de verschillende instellingen gebruikt worden. Dit zou leiden tot een enorm metadata-schema, dat onmogelijk te onderhouden is. Daarom is ervoor gekozen om de beschrijvende metadata samen met de data zelf te archiveren in hun oorspronkelijke metadata-formaat, bijvoorbeeld in de Machine-Readable Cataloging standard (MARC, [11], [12]), zodat geen informatie verloren gaat. Daarbovenop biedt de instelling een algemeen aanvaard beschrijvend metadata-schema waardoor het over het nodige instrument beschikt om het volledige archief te doorzoeken. Wanneer de juiste gegevens gevonden zijn, kunnen de originele metadata die als data zijn opgeslagen, alsnog aan de gebruikers worden voorgelegd.

5. Preserveringsmetadata

Geen enkel opslagapparaat is perfect en eeuwig betrouwbaar, waardoor het probleem van bitpreserving nog steeds niet opgelost is. Een bithalfwaardetijd van 8×10^{17} jaar geeft aan dat

de overlevingskans van een eeuw voor 1 Petabyte data 50% is [13]. Zoals vermeld, is informatie in digitale vorm een conceptueel object. Ze kan gemakkelijk worden gekopieerd en gewijzigd zonder dat dit bij de representatie zichtbaar is. In tegenstelling tot analoge informatie is het veel moeilijker om de authenticiteit van digitale informatie te bewaren. Dit probleem kan worden opgelost door de toevoeging van zogenaamde conserveringsmetadata (houdbaarheidsmetadata) aan het informatiepakket van het archiefmateriaal. Dergelijke metadata bevatten onder andere controlesommen en andere veiligheidsmaatregelen die kunnen testen of gegevens aangepast zijn zonder dat dit gedocumenteerd werd. Het gaat bijvoorbeeld om controlesommen, digitale handtekeningen, certificaten, encryptie en cyclische redundantiecontrole (CRC). Behalve conserveringsmetadata heeft een gearhiveerde dataset ook metadata nodig die de herkomst documenteren (*provenance*-metadata). Die metadata beschrijven de genese van de inhoudelijke informatie: de oorspronkelijke eigenaars van de data, de processen die de huidige vorm van de data bepalen en alle beschikbare versies. Het kan bijvoorbeeld gaan over beschrijvingen van de gebruikte scanapparatuur, metadata over het eigenlijke scanproces, de gebruikte software, verwijzingen naar de oorspronkelijke analoge data en de versiegeschiedenis van alle bestanden. Die informatie is nodig om alle veranderingen van de data te kunnen nagaan. Ook *context*-metadata moeten bewaard worden in de omvattende conserveringsmetadata. Die metadata beschrijven immers de relaties van de inhoudelijke data met andere data die niet in het informatiepakket verpakt zijn. Voorbeelden van *context*-metadata zijn gerelateerde datasets, verwijzingen naar documenten uit de oorspronkelijke omgeving op het moment van publicatie, help-bestanden en de taal.

6. Rechtenmetadata

De rechten op digitale objecten, waaronder auteursrechten, licenties en eventuele subsidies, moeten eveneens in het archief beschreven worden. Deze rechteninformatie is immers essentieel om de toegang tot data op lange termijn te blijven garanderen.

Het is voldoende aangetoond dat bij de ontwikkeling van het metadataschema voor langetermijnbewaring van digitale multimedia, er rekening is gehouden met alle mogelijke niveaus van metadatabeschrijving, van bitniveau tot intellectuele inhoud. Om tegemoet te komen aan al deze metadatabeelden, is in het kader van BOM-vl een gelaagd semantisch metadataschema ontwikkeld. De bovenste laag biedt de beschrijvende metagegevens aan. De onderste laag zorgt voor de conserveringsmetadata, rechtenmetadata, binaire metadata, technische metadata en structurele metadata die nodig zijn voor de diepe archivering. Enkel de bovenste laag zal openbaar zijn en die kan eventueel met extra gegevens van het Web verrijkt worden, voor zover dit uiteraard volgens de juridische voorwaarden toegelaten is. De onderste laag blijft uitsluitend toegankelijk voor het archief. Voor de bovenste laag werd binnen BOM-vl een OWL-representatie van Dublin Core ontwikkeld en voor de onderste laag een OWL-representatie van PREMIS 2.0. Deze standaard is gebaseerd op het OAIS-referentiemodel, waarin de gegevens op alle nodige niveaus nauwgezet beschreven worden. De gelaagde structuur van het ontwikkelde metadataschema is te zien in Figuur 1.

[Fig 1]: Het gelaagd semantisch metadatatamodel (zie Fig1-Metadata schema NL.eps)

Toplaag: Dublin Core OWL

Beschrijvende metadata typeren de inhoud van de gegevens: onderwerp, auteur, datum van creatie, bestandsformaat, enzovoort. Deze metadatalaag maakt het mogelijk om het volledige digitale archief te beheren en te doorzoeken.

Dublin Core [5] werd als implementatiestandaard van de bovenste metadata laag gekozen om zijn wijdverbreide toepassingen en gangbaarheid. De kracht van dit metadataschema is zijn eenvoud en generaliteit. Het bestaat uit een vijftiental velden waarvan de belangrijkste ontwerper, onderwerp, bereik, beschrijving en datum zijn en biedt dus een antwoord op de fundamentele vragen: wie, wat, waar en wanneer. Alle velden in Dublin Core zijn optioneel en herhaalbaar. Daardoor is het relatief eenvoudig om bijna alle bestaande beschrijvende metadataschema's af te beelden op deze generieke standaard. Er bestaan reeds een aantal Dublin Core ontologieën, maar deze modelleren de vijftien velden allemaal als 'annotation properties', die niet altijd door 'semantic reasoners' worden ondersteund. Daarom werd een eigen implementatie voor Dublin Core ontwikkeld die de velden als 'datatype properties' modelleert.

Onderste laag: PREMIS OWL

Voor de onderste laag werd een OWL-schema van de preservingstandaard PREMIS 2.0 ontworpen. Deze laag behandelt de binaire metadata, de technische metadata, de structurele metadata, de preservingmetadata en de rechtenmetadata die het object op de vooropgestelde drie niveaus beschrijft. PREMIS is gebaseerd op het OAIS-referentiemodel en werd in 2003 ontwikkeld door een werkgroep van internationale experts op het gebied van bewaring en metadata. De leden vertegenwoordigden verschillende sectoren zoals bibliotheken, musea, archieven, overheidsinstellingen en particuliere instellingen. Het PREMIS-datamodel, zoals verduidelijkt in Figuur 2, bestaat uit vijf semantische klassen die onontbeerlijk zijn voor digitale bewaring:

- Intellectuele entiteit (*intellectual entity*): het deel van de inhoud dat beschouwd kan worden als een intellectuele eenheid voor het beheer en de beschrijving van de inhoud. Dit kan bijvoorbeeld een boek, een foto of een volledige databank zijn.
- Object (*object*): een discrete eenheid van informatie in digitale vorm.
- Gebeurtenis (*event*): een actie die impact heeft op een object of een agent.
- Agent (*agent*): een persoon, instelling of softwaretoepassing die gerelateerd is aan een gebeurtenis van een object of daarmee verbonden is via de rechten van een object.
- Rechten (*rights*): de beschrijving van een of meer rechten en/of machtigingen op een object of agent.

De object klasse bestaat uit drie subklassen: Bitstream, File en Representation. Dit laat toe om een object op alle drie de niveaus te beschrijven. Intellectuele entiteiten, gebeurtenissen en de rechten staan in een direct verband met een object. Een agent kan alleen worden gerelateerd aan een object door middel van een gebeurtenis of via rechten. Op die manier worden niet alleen de wijzigingen aan een object opgeslagen, maar wordt ook de gebeurtenis met betrekking tot de wijziging beschreven. Die relaties bieden de nodige garanties om de volledige herkomst van een digitaal object te documenteren. Verder voorziet PREMIS voor elke semantische klasse een extensiepunt. Dit extensiepunt kan gebruikt worden om extra informatie aan te bieden die niet door de PREMIS standaard wordt omvat. Dit kan door een instituut bijvoorbeeld gebruikt worden om eigen metadata betreffende de entiteit toe te voegen aan de PREMIS beschrijving. Om de rechten beter te beschrijven werd de PREMIS standaard verder uitgebreid met rollen die een agent kan aannemen met betrekking tot die rechten. Deze rollenlijst vloeit voort uit een studie die alle nodige elementen aanduidt die moeten beschreven zijn in de rechtensectie om de toekomstige risico's in verband met rechten te minimaliseren.

[Fig 2]: Het PREMIS datamodel (zie Fig2-PREMIS 2 NL.eps)

5. Samenvatting:

Voor de preservering van digitale informatie op lange termijn zijn verschillende soorten metadata belangrijk. Beschrijvende metadata zijn nodig voor de beschrijving van de intellectuele entiteiten. Binaire, technische en structurele metadata zijn dan weer essentieel voor de beschrijving van de gegevens op verschillende andere niveaus: van de fysieke bitstroom, het eigenlijke bestand tot de werkelijke representatie. Preserveringsmetadata zijn nodig om de herkomst van de digitale gegevens te beschrijven en de authenticiteit ervan te waarborgen. Ten slotte dienen ook rechtenmetadata over de gegevens te worden opgeslagen.

Het gelaagd semantisch metadataschema dat in dit artikel bondig is toegelicht en dat in het project BOM-vl grondig wordt uitgewerkt, waarborgt de inclusie van alle noodzakelijke metadata-soorten. Door dit semantisch metadataschema ter beschikking te stellen, worden niet alleen de metadata zelf, maar ook de semantiek van de metadata opgeslagen, wat uiterst belangrijk is aangezien terminologie aan voortdurende veranderingen onderhevig is. De bovenste laag, een OWL-representatie van Dublin Core, behandelt en gebruikt de beschrijvende metadata. De onderste laag behandelt de binaire, technische, structurele, preserverings- en juridische metadata. Voor deze laag werd een OWL-representatie van PREMIS 2.0 ontwikkeld. Die standaard heeft namelijk betrekking op alle genoemde soorten metadata. Om de juridische aspecten op een nog meer gedetailleerde wijze te beschrijven, werd het PREMIS OWL-schema verder uitgebreid met een thesaurus die de verschillende wettelijke instanties, organisaties en software definieert.

Door alle data door middel van het gelaagd metadataschema te beschrijven, worden alle risico's van langetermijnarchivering geminimaliseerd. Door de opdeling van het semantisch schema in twee lagen, kan de bovenste laag met de beschrijvende metadata openbaar gemaakt worden en verweven worden met data in het globale Web, voor zover aan de juridische voorwaarden voldaan is. De onderste laag blijft voor het grote publiek gesloten en is noodzakelijk voor de langetermijnbewaring van alle gegevens.

Het uitgebreid rapport waarin deze materie binnen het project BOM-vl behandeld wordt, (*Meta*)datastandaarden voor digitale archieven, kan besteld worden via de Universiteitsbibliotheek Gent.

[Foto boek + ISBN-Nr - 9789052230009] zie BB-cover-HighRes-Front.pdf van boek)

Referenties:

1. EU FP6 Project, *PrestoSpace*, 2004-2007, <http://prestospace.org/project/index.nl.html>.
2. IBBT, *BOM-Vlaanderen. Preservation and Disclosure of Multimedia Data in Flanders*, 2004-2006, <https://projects.ibbt.be/bom-vl/>.
3. Consultative Committee for Space Data Systems (CCSDS), *Reference Model for an Open Archival Information System. Blue book. Issue 1*, 148 p., Washington: CCSDS, 2002, <http://public.ccsds.org/publications/archive/650x0b1.pdf>.
4. McGuinness, D., Van Harmelen, F., et al., *OWL web ontology language overview. W3C Working Draft*, W3C, 2004, <http://www.w3.org/TR/owl-features/>.
5. DCMI, *The Dublin Core Metadata Initiative*, DCMI, 2009, <http://dublincore.org/>.
6. Higgins, S., *PREMIS Data Dictionary*, Glasgow: Digital Curation Centre (DCC), 2007, <http://www.dcc.ac.uk/resource/standards-watch/premis-data-dictionary/>.
7. McGowan, J.F., *AVI Overview*, 2004, <http://www.jmcgowan.com/avi.html>.

8. Fraunhofer IIS, *MP3: Moving Pictures Expert Group Audio Layer III*, 2008, <http://www.iis.fraunhofer.de/EN/bf/amm/projects/mp3/index.jsp>.
9. ITU, *Recommendation T.81. Information Technology – Digital Compression and Coding of Continuous-Tone Still Images – Requirements and Guidelines*, ISO/IEC IS 10918-1, ITU, 2004.
10. Adobe Systems Inc., *TIFF*, 2008, <http://partners.adobe.com/public/developer/tiff/index.html>.
11. Library of Congress, *MARC Standards homepage*, LOC, 2008, <http://www.loc.gov/marc>.
12. Library of Congress, *MARC 21 Format for Bibliographic Data, version 9*, <http://www.loc.gov/marc/bibliographic/ecbdhome.html>.
13. Rosenthal, D. S. H., *Bit Preservation: A Solved Problem?* Stanford University Libraries, CA, 2008.