Metadata- en Technische Standaarden voor digitaliseringsprojecten

(Draft 4 02.12.2009_CDL)

Inhoud

1. Introductie tot de lezing

De digitalisering van cultureel erfgoed, met alle mogelijke voordelen maar ook uitdagingen, wordt algemeen beschouwd als een belangrijke stap in de ontsluiting van dat erfgoed binnen een regionale en Europese context. Een aantal projecten willen met de creatie van digitale bibliotheken een centraal toegangspunt creëren tot online informatie en digitale content afkomstig uit verschillende musea, archieven, bibliotheken en instituten. De creatie van de digitale bibliotheek Europeana is daarin ongetwijfeld het belangrijkste project met als doel het grootste centrale toegangspunt tot Europa's culturele en wetenschappelijke erfgoed te worden en er tevens voor te zorgen dat deze informatie bronnen eenvoudiger te gebruiken worden in een online omgeving.
Dit doel kan echter niet bereikt worden zonder de medewerking van een groot aantal individuele instellingen en musea die bereid zijn hun digitale informatie ter beschikking te stellen via de Europeana portaalsite. Voor de synchronisatie van de toelevering van al deze digitale informatie afkomstig uit verschillende types van collectiedatabases, wordt gebruik gemaakt van "aggregatoren" en best practice netwerken zoals Athena, EuropeanaLocal, EuropeanaConnect, ed., die instaan voor het creëren van standaarden en richtlijnen die de aanlevering van de metadata informatie en content door de deelnemende Europese instellingen en musea moeten mogelijk maken. Momenteel loopt deze uitwisseling van informatie vaak nog moeizaam, omdat de meeste musea en instellingen hun eigen methodologie en databaseformaat hanteren. Deze formaten zijn allemaal anders van elkaar en aangepast aan de eigen noden van het museum en de collecties. Om deze data bij een automatische export correct te kunnen interpreteren en gebruiken, moet er een algemene overeenkomst of standaard metadata schema gebruikt worden om de betekenis of de semantiek van deze data te definiëren.
Er werden en worden dan ook nog steeds verschillende metadata schema's als standaarden ontwikkeld voor specifieke domeinen zoals voor het bibliotheekwezen, archivering, of voor digitale culturele erfgoed bronnen. Er verschenen tevens een aantal officiële rapporten die een overzicht bieden van de bestaande standaarden die in gebruik of geschikt zijn voor de online uitwisseling van digitale data afkomstig van Europese musea, archieven en bibliotheken.

Deze uiteenzetting zal zich beperken tot de technische en metadatastandaarden en hun gebruik. Dit stuk is voornamelijk gebaseerd op 3 bronnen: de technische richtlijnen zoals beschreven op de website van Minerva EC, (MINERVA EC - Technical Guidelines), de technische richtlijnen van Europeana versie 1 (http://www.version1.europeana.eu) en de best practises zowel van Athena (http://www.athenaeurope.eu) als in de lokale erfgoedinstellingen.

Standaard

- een richtlijn die consistent gebruikt wordt als een regel, richtlijn of definitie
- maakt het leven gemakkelijker en verhoogt de betrouwbaarheid van data

- wordt ontworpen door ervaring en expertise samen te brengen door geïnteresseerden zoals gebruikers, producenten, regelgevers
- van toepassing op een product, object, proces of dienstverlening
- Bijkomende: verhoogt de interoperabiliteit

2. Richtlijnen technische standaarden

Basisprincipe: Altijd open standaarden gebruiken voor het creëren en aanleveren van digitale inhoud ("content").

Om
- de toegankelijkheid te verhogen
- het hergebruik te stimuleren
- de afhankelijkheid van 1 leverancier te beperken en de kostprijs te drukken (licenties)

Een advies voor het gebruik van technische standaarden kan enkel gezien worden in het licht van de gebruiksomgeving :

1. Data-archivering (bron)
2. Data-ontsluiting
3. Data-uitwisseling

De data-archivering (bron)
- de fase waarbij een digitaal object wordt gecreëerd (of gearchiveerd) vanuit de werkelijkheid
- dit kan gebeuren op basis van bepaalde technieken: voorbeeld fotografie, scannen, OCR, digitale opname. Alles wat digitaal gecreëerd wordt (digital born) wordt standaard gearchiveerd
- gebeurt meestal bij diegene die de collectie in haar bezit heeft en met hun machines
- Kwaliteit images: hoog
- Doel: bewaring
- Leveringsnelheid: laag
- Auteursrechten: hoog

De data-ontsluiting
- de fase waarbij er een betekenis wordt gegeven aan elk digitaal object. Dit houdt in: het beschrijven van de metadata (wat is metadata?) en het aanbieden van deze verrijkte inhoud op een eigen site
- Kwaliteit images: gemiddeld
- Leversnelheid: behoorlijk
- Doel: object gedetailleerd "beschrijven"
- Auteursrechten: midden

De data-uitwisseling
- gebruikers krijgen toegang tot een deel van het gedigitaliseerd materiaal
- de data-uitwisseling is meestal opgevat als het optimaliseren van zoekopdrachten en zoekresultaten, en maakt gebruik van de metadata die hiervoor geschikt zijn
- Kwaliteit images: laag
- Leveringsnelheid: hoog

- Doel: herkenbaarheid verhogen
- Auteursrechten: laag

Aanbevelingen inzake het gebruik van standaarden voor:

1. Tekstmateriaal
2. Beeldmateriaal
3. Audiomateriaal
4. Videomateriaal
5. Vector Graphics
6. Virtual reality

Tekstmateriaal	BRON	Ontsluiting	Uitwisseling
Bestandformaat	XML (voorkeur) PDF / DjVu	HTML PDF / DjVu ODF; RTF; MS Word (bijkomend)	Kan gebeuren op basis van een sample (tekst of beeld)
Kwaliteit			

Beeldmateriaal	BRON	Ontsluiting	Uitwisseling
Bestandformaat	TIFF	JPEG / PNG	JPEG /PNG
Kwaliteit Kleur	8 bit greyscale 24 bit color	8 bit greyscale 24 bit color	8 bit greyscale 24 bit color
Resolutie (dpi)	600 foto's 2400 scans	150 à 200	
Max dimensie (pixels)		600	100 à 200

Audiomateriaal	BRON	Ontsluiting	Uitwisseling
Bestandformaat	Niet gecomprimeerde: WAV; AIFF Gecomprimeerd: MP3,WMA; RealAdio; AU	Gecomprimeerd: WAV; AIFF Niet gecomprimeerd: MP3; WMA; RealAudio; AU	Relevant beeld
Kwaliteit	24bit stereo en 48/96 KHz sample rate	256 Kbps (~CDkwaliteit); 160 Kbps (goede kwaliteit)	

Videomateriaal	BRON	Ontsluiting	Uitwisseling
Bestandformaat	Uncompressed RAW AVI (voorkeur) Compressed : MPEG; WMF; ASF; Quicktime	MPEG1; AVI; WMV; Quicktime	Relevant beeld
Kwaliteit	Frame size: 720x576 pixels Frame rate : 25 frames per second 24 bit kleur PAL colour encoding	ASF; WMF; Quicktime	

Vector Graphics	BRON	Ontsluiting	Uitwisseling
Bestandformaat	SVG (voorkeur) SWF (alternatief)	SVG (voorkeur)	Relevant beeld

Virtual Reality	BRON	Ontsluiting	Uitwisseling
Bestandformaat	X3D (voorkeur) Quicktime VR (alternatief)	X3D (voorkeur) Quicktime VR (alternatief)	Relevant beeld

3. Richtlijnen inzake metadata standaarden

Principe: altijd standaarden gebruiken voor het creëren en aanleveren van digitale content

Om
- Interoperabiliteit en toegankelijkheid te verhogen
- Hergebruik te stimuleren
- Afhankelijkheid ten opzichte 1 systeem vermijden of te weinig kennis binnen het team

Een advies voor het gebruik van technische standaarden kan enkel gezien worden in het licht van de gebruiksomgeving :

In analogie met de technische standaarden zijn er zijn drie gebruiksomgevingen
1. Collectie beheer (bron)
2. Data-ontsluiting
3. Data-uitwisseling

Het collectiebeheer (bron)

Metadata wordt meestal geproduceerd door de houder van de collectie, mits heel veel menselijke inspanning en met het systeem waarover de collectiehouder beschikt.

Deze data is afkomstig van diverse bronnen:
- Activiteiten gelinkt met het beheer van de collecties (voorbeeld: aankopen, uitlenen, recht op gebruik en reproductie,…)
- Beschrijving van het object zelf: naam, materiaal, dimensies, geografie,…
- Activiteiten gelinkt met de levenscyclus van het object: vb. ontstaan, gebruik, restauratie…
- Activiteiten gelinkt met personen, organisatie, groepen en plaatsen gelinkt met de levenscyclus van het object

Karakteristieken voor data afkomstig uit het collectiebeheer
- Omvang van de metadata: Hoog
- Doel: preservatie
- Domein: afhankelijk van het domein (Museum, Bibliotheek en Archief), land en organisatie
- Rechten: gebruik binnen de organisatie
- Aanlevering: traag – veel manueel werk

De data-ontsluiting

Men heeft betekenisvolle toegang tot (een reproductie van) het object voorbeeld via een site met een subset van metadata en een foto, videofragment of geluidsfragment. Zelfde fase van technische standaarden.

- Omvang van de metadata: Subset van metadata uit collectiebeheer
- Doel: aanbieden van "content" op eigen site
- Domein: afhankelijk ofwel cross domein ofwel specifiek
- Aanlevering: vlug
- Rechten: copyright (statement)

De data-uitwisseling

De gebruikers krijgen toegang tot een beperkt deel van de metadata, meestal in functie van een zoekopdracht. In deze omgeving zal men ook het object trachten te tonen met een thumbnail.

- Omvang van metadata: beperkt
- Doel: ruime verspreiding en indexering
- Domain: cross domein
- Aanlevering; vlug
- Rechten: beperkt

Aanbevelingen in functie van gebruiksomgeving

Meerdere aanbevelingen mogelijk en niet limitatief. Hierna best practising:

De collectiebeheer

Domein	**Aanbevolen standaarden**
Musea	Spectrum / CDWA
Bibliotheken	MARC
Archieven	ISAD(G); EAD

Spectrum
Standaard voor het documenteren van het collectiebeheer. Gebouwd rond 21 procedures die gewoonlijk voorkomen in musea. Gecentraliseerd rond "informatie-eenheden", dit is het geheel van data dat nodig is om de procedures te ondersteunen. Ontwikkeld in het Verenigd Koninkrijk, maar thans ook wijd verspreid in Vlaanderen en Nederland.
(Engels en Nederlandstalige beschrijving verkrijgbaar op de website van Collections Trust: http://www.collectionstrust.org.uk). FARO ondersteunt in Vlaanderen actief de Nederlandstalige versie van spectrum. (http://www.faronet.be/blogs/spectrum/download-spectrum).

CDWA (Categories for the Description of Works of Art)
Standaard op initiatief van Getty bedoeld om de gegevens in diverse kunstdatabanken zoveel mogelijk te harmonisering door de beschrijving van de objecten en foto's in concepten te gieten. Op basis van CDWA kunnen diverse compatibele subsets gemaakt worden
(http://www.getty.edu/research/conducting_research/standards/cdwa/index.html)

MARC (Machine Readable Cataloging)
Standaard voor het weergeven en het uitwisselen van bibliografische informatie op een wijze die door "machines" kan worden gelezen en geïnterpreteerd (http://www.loc.gov/marc/bibliographic/ecbdhome.html)

ISAD(G) General International Standard Archival Description
Algemene regels voor de beschrijving van archieven. Er is een set van 26 elementen waarmee men door deze te combineren een eenheid van archief kan beschrijven. (http://www.ica.org)

EAD (Encoded Archival Description)
W3C Schema gebruikt om elektronische archieven terug te vinden en te beschrijven. (http://www.loc.gov/ead)

Indien een organisatie een eigen interne standaard gebruikt moet ze wel in staat zijn om haar gegevens te mappen naar deze aanbevolen standaarden volgens domein.

De data-ontsluiting

Een echte standaard voor data-ontsluiting bestaat niet. Het publiceren op een website is meestal gebaseerd op een subset van de metadataelementen die in de eigen database zitten. Veelal zal men gebruik maken van technieken om de beelden op de website zoveel mogelijk technisch te vergrendelen onder meer door gebruik van watermarking, insluiten in Flashanimaties,… afhankelijk van het doel van de website : voorbeeld verkoop van foto's, tonen van de collecties,…

De data-uitwisseling

Hier is een formaat nodig om te exporteren en te importeren in gegevensbanken.

Verspreiding gebeurt optimaal via Europeana. Het principe van Europeana is zeer eenvoudig. De beschikbare metadata wordt optimaal voor zoekopdrachten geïndexeerd en toont de gestructureerde gegevens waarover Europeana beschikt met als het kan een relevante thumbnail. Een hyperlink verwijst naar de eigen website van de inhoudleverancier (of "aggregator"). Het formaat dat Europeana gebruikt is ESE (dit is in hoofdzaak Dublin Core, aangevuld met een aantal specifieke velden).

Dublin Core (DC) is niettegenstaande zijn gigantische beperkingen evenwel één van de enige standaarden die door quasi ieder gekend is. De **Dublin Core Metadata Element Set (DCMES)** is ongetwijfeld een van de meest bekende metadataschema's. Deze ISO-standaard geeft een eenvoudig schema voor beschrijvende metadata (resource discovery metadata), ontwikkeld door een interdisciplinair project en ontworpen om het ontdekken van bronnen te ondersteunen die afkomstig zijn uit een brede waaier van domeinen. Het gaat om een set van beschrijvende metadata die een aantal conventies voor digitale beschrijvingen definieert en informatie geeft over een databron en zijn URI. Het beschrijft 15 elementen om op eenvoudige manier bronnen uit verschillende domeinen te ontdekken zoals: Titel, Auteur, Onderwerp, Beschrijving, Uitgever, Medewerkers, Datum, Type, Formaat, Correlatief, Bron, Taal, Relatie, Bereik en Rechten. Elk van deze elementen is optioneel en herhaalbaar.

Dublin Core (DC) wordt wijdverspreid gebruikt als een 'minimum' metadata standaard voor verspreiding, ook binnen Europeana zelf. De standaard wordt echter soms beschouwd als slecht georganiseerd wegens de moeilijkheid om definities te vinden voor individuele metadata elementen. (http://dublincore.org)

Hiernaast bestaan er een aantal standaarden voor metadata structuren die in het bijzonder gericht zijn op de registratie en export van digitale data eigen aan een specifiek cultureel domein of gericht op een bepaald doel.

Zowel in het bibliotheek als het archiefwezen zijn de gebruikte standaarden voor collectiebeheer ook de standaarden voor de ontsluiting. In de meeste toepassingen worden deze standaarden dan ook als exportformaat geprogrammeerd en gehanteerd. Indien men zich tevreden kan stellen met beperktere informatie kan het ESE model van Europeana gebruikt worden (zie verder).

In de musea bestaan er geen echte uniforme standaarden die alom gebruikt worden, wel integendeel.
Enkele belangrijke museumspecifieke standaarden voor dataontsluiting en interoperabiliteit zijn **Categories for the Description of Work of Arts Lite XML (CDWA Lite)** en **MuseumdatXML.**

Museumdat blijkt een aanrader voor musea als best practise. Het idee groeit om dit ook **uit te breiden tot andere domeinen**. Op dit moment is het nog steeds een project: Light Information for Describing Objects (LIDO).

Museumdat: is een XML formaat dat zich laten inspireren heeft op het CRM model. Alles draait rond events, voorbeeld "restauratie" dat 3 componenten heeft: datum, plaats en actor. (http://Museumdat.org).

Er is ook nog een alternatief voor museumdat als harvesterformaat: Europeana zelf gebruikt een standaard, dat ESE (Europeana Semantic Elements) heet. De versie 3.2. is de laatste versie, maar wordt verder ontwikkeld. Erfgoedinstellingen kunnen rechtstreeks naar ESE mappen, maar verliezen hierbij wel de rijkere metadata die ze in hun eigen systeem hebben.

Volledigheid van metadata	**Aanbevolen standaard**
Hoog	Museumdat / LIDO
Laag	ESE

Een aparte noot over **CIDOC Conceptual Reference Model (CIDOC - CRM).** Deze ISOstandaard is dan weer een conceptueel object-georiënteerd model dat via extensieve ontologische beschrijvingen culturele erfgoed en museum documentatie van betekenis of semantiek voorziet. Deze standaard is tevens geschikt voor de interoperabiliteit met andere domeinen zoals de bibliotheken en archieven. Het is evenwel een standaard die enige studie vergt om deze te kunnen gebruiken en niet voor iedereen zo toegankelijk. **(http://cidoc.ics.forth.gr) is**. Ze kan evenwel geschikt zijn om te voldoen aan de vraag voor uitwisseling van "rijke" metadata tussen de diverse domeinen.

4. Metadata mapping

Mapping is de overgang tussen de ene datastructuur naar een andere. Hierbij wordt gekeken naar de inhoud van de metadata. Welke element in de ene structuur komt overeen met de andere.
In het Engels wordt ook het woord crosswalk gebruikt.

Wijzig in geen geval de standaard bij de creatie van een eigen ontworpen standaard zelfs al is de verleiding zeer groot.

1. De organisatie bezit meer datavelden dan er elementen zijn in een standaard: gebruik alternatieve velden om deze data niet te verliezen
2. De organisatie is te klein om alle elementen uit een standaard te gebruiken: gebruik toch de standaard maar vul de velden niet in die je nu niet nodig hebt. Misschien is er ook documentatie of opleiding nodig

Indien er een eigen standaard wordt gebruikt (omwille bijvoorbeeld uit historische redenen): map deze structuur naar een export standaard. Zorg ervoor dat de betekenis van de velden zo goed mogelijk gelijk blijven. Hiervoor zijn tools ter beschikking.

Mapping zal hoogstwaarschijnlijk nodig zijn om bijvoorbeeld de gegevens samen te brengen bij een “aggregator”.

Voorbeeld van mapping

EAD example:
<controlaccess>
<geogname role=”country of coverage” source=”tgn”>United States</geogname>
<geogname role=”state of coverage” source=”tgn”>California</geogname>
<geogname role=”city of coverage” source=”tgn”>San Francisco</geogname>
</controlaccess>
Becomes
<dcterms:spatial>United States</dc:spatial>
<dcterms:spatial>California</dc:spatial>
<dcterms:spatial>San Francisco</dc:spatial>

Voorbereiding Mapping technisch

1. Zorg dat je de betekenis/concept van jouw metadata kent;
2. Check de mogelijkheid om XML te genereren vanuit jouw database of collectiebeheer toepassing;
3. Zorg dat je de betekenis/concept van de metadata van het standaard schema kent;
4. Verricht een manuele mapping van jouw metadata naar het standaard schema

Enkele scenario’s

1. Pas jouw collectiebeheer toepassing aan.

Implementeer een exportformaat in jouw collectiebeheer toepassing
De mapping wordt door een vendor of de open source community in jouw systeem vertaald.

2. Gebruik tools om een mapping te verrichten

Voorbeeld voor beperkte datastructuur:
- exporteer naar XLS (Excel) en geef de kolommen de naam van de elementen van de standaard (ESE);
- Tool vertaalt XLS naar XML met de juiste XMLheadertags;
- Bewerken van XML (ESE headers)

3. Vraag hulp aan informatici of aan instellingen die bijstand bieden aan erfgoedinstellingen;

4. Sluit je aan bij een Europees of lokaal project die u ondersteunt met de juiste methodologie en techniek

5. Conclusie

Ondanks dat het overzicht hierboven slechts een klein deel van de bestaande standaard formaten omvat, zijn er toch maar een beperkt aantal belangrijke 'best practice' standaarden die extensief gebruikt worden op een Europese en Internationale schaal en die dienst kunnen doen om digitale gegevens met betrekking tot het culturele erfgoed op een georganiseerde en gestandaardiseerde manier uit te wisselen. Uit de moeilijkheden die de verschillende Europese digitaliseringsprojecten ondervinden, blijkt dat we nog ver verwijderd zijn om een goede interoperabiliteit te bereiken. Met deze lezing is het dan ook de bedoeling om een scherper beeld te schetsen van het bestaande en aanbevolen standaarden voor musea, archieven en bibliotheken in Europa.

Contact
Chris De Loof
Projectmanager ICT
Koninklijke Musea voor Kunst en Geschiedenis
Jubelpark 10
1000 Brussel
+32 2 741 73 06
c.deloof@KMKG.BE
http://www.kmkg.be