CATCH





CONTINUOUS ACCESS TO CULTURAL HERITAGE PLUS





D 3.2.3

Implementatie OAI Beeld en Geluid

Hennie Brugman

Ministerie van Onderwijs, Cultuur en Wetenschap

NWO Nederlandse Organisatie voor Wetenschappelijk Onderzoek

Ministerie van Economische Zaken

1 Documenteigenschappen

Rapportage datum: 16 september 2010
Rapportage periode: September 2009 – augustus 2010
Versie: ~~Concept~~ / definitief
Eventuele wijzigingen tov vorige versie? nvt

1.1 Document locatie

Bestandsadres:	http://www.catchplus.nl/wp-content/uploads/2010/09/D3.2.3-OAI-BenG.doc

1.2 Historie

Versie	Datum	Veranderingen (concept/definitief)	Opdrachtgever	Auteur(s)
1.0	16/9/10		Marius Snyders	Hennie Brugman

1.3 Goedkeuring

Dit document heeft de volgende goedkeuringen nodig.

Naam	Rol	Opmerkingen	Datum
projectleider	verantwoordelijke		
projectbureau	advies		
Stuurgroep	akkoord		

1.4 Distributie

De volgende personen hebben een exemplaar ontvangen.

Naam	Functie	Datum
Marius Snyders	Projectleider CATCHPlus	16/9/10

Inhoud

1 Documenteigenschappen 2
1.1 Document locatie 2
1.2 Historie 2
1.3 Goedkeuring 2
1.4 Distributie 2
Management samenvatting 4
2 Bericht van oplevering 5
2.1 Paragraaf 1 ..... **Fout! Bladwijzer niet gedefinieerd.**
3 Gedetailleerde beschrijving van de deliverable 7
3.1 Paragraaf 1 ..... **Fout! Bladwijzer niet gedefinieerd.**
3.1.1 Subparagraaf 1 ..... **Fout! Bladwijzer niet gedefinieerd.**
4 Licentie 9
5 Demonstraties, toepassingen 10
6 Performance 10
7 Conclusie 12
Bijlagen 12

Management samenvatting

OAI-PMH is een standaard protocol voor 'harvesten' van metadata records. De standaard vereist dat minimaal metadata volgens het Dublin Core formaat beschikbaar wordt gemaakt, maar meerdere metadata formaten kunnen naast elkaar worden gepubliceerd.

CATCHPlus vereist, dat collectie-metadata vanuit de diverse deelprojecten wordt ingebracht volgens standaarden als OAI-PMH. Daarom is besloten om vanuit het projectbureau actief de implementatie van z.g. OAI data providers te ondersteunen.

Eén van de deelnemers, die hiervan gebruik heeft gemaakt is het Nederlands Instituut voor Beeld en Geluid. Beeld en Geluid beheert een metadata catalogus voor zo'n 1,4 miljoen objecten, en hanteert daarbij een eigen metadata formaat: immix.

De deliverable betreft het realiseren van een OAI data provider voor de gehele collectie van Beeld en Geluid, waarbij metadata wordt gepubliceerd in de immix, qualified Dublin Core en Dublin Core formaten. In eerste instantie is daarbij slechts de Polygoon deelcollectie toegankelijk gemaakt voor iedereen.

De implementatie is gebaseerd op het open source pakket Meresco. Daaraan is maatwerksoftware toegevoegd voor conversie van immix metadata naar (qualified) Dublin Core, voor het inladen en up-to-date houden van de indexen van de data provider op basis van immix XML bestanden, voor het dynamisch creëren van zogenaamde landingspagina's (webpagina's met metadata voor ieder object in de catalogus) en voor het beheren van met de catalogusbeschrijvingen geassocieerde persistente identifiers.

Ondanks het feit dat implementatie van een OAI data provider voor de Immix catalogus van Beeld en Geluid complexer bleek te zijn dan aanvankelijk gedacht, is uiteindelijk voldaan aan nagenoeg alle wensen van Beeld en Geluid en aan de randvoorwaarden gesteld door het projectbureau van CATCHPlus. De gebouwde oplossing is robuust, performant waar hij dat moet zijn, en volledig OAI-PMH compliant.
De Beeld en Geluid OAI data provider is de eerste, die kon worden geharvest naar de portal van Nederlands Erfgoed Digitaal!, en is daarmee succesvol toegepast.

Bovendien is het interne update-mechanisme (de 'inbox') van de data provider succesvol ingezet voor het automatisch creëren, beheren en publiceren van persistente identifiers voor de gehele Beeld en Geluid catalogus.

2 Bericht van oplevering

2.1 Wat

Deze deliverable betreft de implementatie van een z.g. OAI-PMH data provider voor de gehele Immix collectie van Beeld en Geluid (1,4 miljoen objecten). Van deze gehele collectie is slechts de Polygoon deelcollectie vrij via internet te harvesten dmv OAI-PMH, de rest is alleen vanuit Beeld en Geluid zelf toegankelijk.

De bouwer (de firma SeekYouToo) heeft de data provider geïmplementeerd op basis van hun open source pakket Meresco. Relevante modules uit Meresco zijn geïnstalleerd en geconfigureerd op een door Beeld en Geluid beschikbaar gestelde server (gehost door de firma Tuxic). Vervolgens is maatwerk-software geschreven om immix metadata (aangeleverd in XML) volgens specificaties van Beeld en Geluid om te zetten naar Qualified Dublin Core, en vervolgens naar Dublin Core. Er is een 'inbox mechanisme' geïmplementeerd, dat kan worden gebruikt om metadata in Meresco te laden, of om al geladen metadata te updaten. XML bestanden in deze inbox worden door dit mechanisme regelmatig geïnspecteerd en zonodig geconverteerd en in indexen voor de drie ondersteunde metadata-formaten ondergebracht.

De gehele Beeld en Geluid immix catalogus is vervolgens via deze inbox in een grote batch operatie ingeladen.

Er is ook van het bovengenoemde inboxmechanisme gebruik gemaakt om het beheer van persistente identifiers voor de Beeld en Geluid collectie te regelen (zie deliverable D 3.3.8). Bij het detecteren van wijzigingen in de collectie door het inbox-mechanisme worden de betrokken persistente identifiers gecontroleerd, en zonodig aangemaakt of aangepast.

Zowel Meresco als de voor dit project geïmplementeerde maatwerksoftware is onder open source licentie beschikbaar gemaakt, en als een z.g. 'tarball' aan het projectbureau beschikbaar gesteld.

Documentatie met betrekking tot de Meresco installatie en –configuratie, en maatwerksoftware is meegeleverd.

2.2 Wie, wanneer en waar

De oplevering heeft op 10 augustus 2010 plaatsgevonden bij Beeld en Geluid.

Aanwezig waren:
Marius Snyders en Hennie Brugman namens het CATCHPlus projectbureau, Willem Melder namens Beeld en Geluid, en Erik Groeneveld en Jurjan-Paul Medema namens SeekYouToo.

2.3 Mate van succes

Uiteindelijk is de implementatie van deze deliverable succesvol verlopen. De requirements, zoals in detail vermeld op vier opeenvolgende planformulieren, zijn gehaald. Het bleek, dat in eerste instantie de performance bij het compleet inladen van de 1,4 miljoen metadata records van Beeld en Geluid teleurstellend was, maar daarop zijn inmiddels verbeteringen aangebracht. Bovendien is het in principe niet nodig de gehele collectie opnieuw in te laden,

dat kan stap voor stap gebeuren, waarbij de OAI data provider gewoon consistent blijft functioneren.

De performance bij het harvesten, en vervolgens zoeken in een centrale index is uitstekend.

Het koppelen van het beheer van persistente identifiers aan het up-to-date houden van de OAI data provider bleek te werken, en daarvoor een elegante geautomatiseerde oplossing te bieden.

De data provider is inmiddels gebruikt om de Polygoon-deelcollectie te harvesten naar een centrale site, die wordt ingericht in het kader van het project Nederlands Erfgoed Digitaal! Vandaaruit is de collectie efficiënt doorzoekbaar, en kan metadata voor de collectie door middel van webpagina's worden getoond.

3 Gedetailleerde beschrijving van de deliverable

OAI-PMH is een standaard protocol voor 'harvesten' van metadata records. De standaard vereist dat minimaal metadata volgens het Dublin Core formaat beschikbaar wordt gemaakt, maar meerdere metadata formaten kunnen naast elkaar worden aangeboden.

CATCHPlus vereist, dat collectie-metadata vanuit de diverse deelprojecten wordt ingebracht volgens standaarden als OAI-PMH. Daarom is besloten om vanuit het projectbureau actief de implementatie van z.g. OAI data providers te ondersteunen.

Eén van de deelnemers, die hiervan gebruik heeft gemaakt is het Nederlands Instituut voor Beeld en Geluid. Beeld en Geluid beheert een metadata catalogus voor zo'n 1,4 miljoen objecten, en hanteert daarbij een eigen metadata formaat: immix.

De deliverable betreft het realiseren van een OAI data provider voor de gehele collectie van Beeld en Geluid, waarbij metadata wordt gepubliceerd in de immix, qualified Dublin Core en Dublin Core formaten. In eerste instantie is daarbij slechts de Polygoon deelcollectie toegankelijk gemaakt voor iedereen.

De implementatie is gebaseerd op het open source pakket Meresco. Daaraan is maatwerksoftware toegevoegd voor conversie van immix metadata naar (qualified) Dublin Core, voor het inladen en up te date houden van de indexen van de data provider op basis van immix XML bestanden, voor het dynamisch creëren van zogenaamde landingspagina's (webpagina's met metadata voor ieder object in de catalogus) en voor het beheren van met de catalogusbeschrijvingen geassocieerde persistente identifiers.

De data provider is geïnstalleerd en geconfigureerd op een door Beeld en Geluid beschikbaar gestelde server. De volledige catalogus van Beeld en Geluid is vervolgens in deze data provider ingeladen, waarbij als neveneffect ook 1,4 miljoen persistente identifiers zijn aangemaakt.

3.1 Onderdelen van de deliverable

De diverse onderdelen van de deliverable zijn online te vinden (stand 15 september 2010):

Techische documentatie:
http://www.catchplus.nl/wp-content/uploads/2010/09/OAI-BenG_DNA.pdf

Installatie-handleiding:
http://www.catchplus.nl/wp-content/uploads/2010/09/OAI-BenG_installatie.pdf

Mapping document Immix metadata naar (qualified) Dublin Core:
http://www.catchplus.nl/wp-content/uploads/2010/09/beng_conversie_iMMix_oai_qdc_0.1_20100204.doc

Source code nog niet online beschikbaar, maar via het CATCHPlus projectbureau te verkrijgen.

De data provider service is online beschikbaar op: http://oai.tuxic.nl

Geharveste Polygoon metadata kan worden doorzocht en bekeken via de portal van Nederlands Erfgoed Digitaal! (onder constructie): http://neddev.cq2.org

3.2 Bugs en omissies

Er zijn momenteel geen bugs bekend. De deliverable is volledig conform de overeengekomen planformulieren, dus de deliverable betreft een volledig afgerond product.

3.3 Toekomstplannen

Er zijn geen plannen voor verbeteringen of uitbreidingen van de functionaliteit. Waar in het kader van CATCHPlus Beeld en Geluid catalogus-informatie wordt gebruikt, zal dat zijn op basis van metadata harvesting via deze OAI data provider.
Momenteel kan nog niet worden voorzien in hoeverre Beeld en Geluid ook buiten de context van CATCHPlus van deze OAI implementatie gebruik zal gaan maken.

4 Licentie

Meresco (http://meresco.org) is beschikbaar via SourceForge onder GNU Public License. De maatwerkcode uit dit project valt eveneens onder GPL, en is op dit moment op aanvraag beschikbaar via het CATCHPlus projectbureau. In een later stadium zal ook deze source code via SourceForge beschikbaar worden gemaakt.

De in het kader van deze deliverable door Beeld en Geluid gepubliceerde metadata is vrij toegankelijk. Het betreft hier vooralsnog alleen de Polygoon-deelcollectie.

5 Demonstraties, toepassingen

Twee toepassingen zijn nu al operationeel of voorzien:

- harvesting van de Polygoon metadata in het kader van het project Nederlands Erfgoed Digitaal!
- gebruik van collectie metadata in het kader van de cases van diverse CATCHPlus deelprojecten

6 Performance

Bij het evalueren van de performance van deze deliverable moet naar een tweetal verschillende aspecten worden gekeken:

Performance tijdens het laden en updaten van metadata in de data provider.
Vanwege de complexe hiërarchische structuur van de immix metadata moeten vele bestanden herhaaldelijk worden geopend, en moet veel informatie tijdelijk worden opgeslagen. Omdat het ook nog een groot aantal bestanden betreft, is het compleet inladen van de Immix catalogus een tijdrovende aangelegenheid (in de orde van weken). Echter, eenmaal ingeladen kunnen wijzigingen worden doorgevoerd, zonder dat de data provider zijn diensten moet onderbreken. Daarbij blijft de toestand van de data provider steeds consistent.

Performance tijdens harvesten en zoeken.
Meresco is hiervoor geoptimaliseerd. Daardoor is de performance tijdens harvesten en zoeken uitstekend te noemen.

7 Conclusie

Ondanks het feit dat implementatie van een OAI data provider voor de Immix catalogus van Beeld en Geluid complexer bleek te zijn dan aanvankelijk gedacht, is uiteindelijk voldaan aan nagenoeg alle wensen van Beeld en Geluid en aan de randvoorwaarden gesteld door het projectbureau van CATCHPlus. De gebouwde oplossing is robuust, performant waar hij dat moet zijn, en volledig OAI-PMH compliant.
De Beeld en Geluid OAI data provider is de eerste, die kon worden geharvest naar de portal van Nederlands Erfgoed Digitaal!, en is daarmee succesvol toegepast.

Bovendien is het interne update-mechanisme (de 'inbox') van de data provider succesvol ingezet voor het automatisch beheren en publiceren van persistente identifiers voor de gehele Beeld en Geluid catalogus.