



Datum
September 2010

notitie

FORUM STANDAARDISATIE

Auteurs:	prof. dr. T.W.C. Huibers MMC, ir. L.M. Punter		
Datum:	September 2010	Versie	1.0
Betreft:	Additioneel onderzoek OAI-PMH		

1 Aanleiding

Door de Vereniging Edustandaard / Kennisnet is de 'Afspraak Metadata Harvesting' aangemeld voor opname op de lijst met verplichte open standaarden onder het 'pas toe of leg uit' regime. Naar aanleiding van deze aanmelding is door het Forum Standaardisatie een expertonderzoek ingesteld.

De expertgroep, onder leiding van prof. dr. Theo Huibers MMC, adviseerde (kort gesteld) het volgende:

- Plaats achterliggende standaard, OAI-PMH versie 2.0 (Open Archives Initiative Protocol for Metadata Harvesting), op de lijst.
- Als functionele toepassingsgebied te definiëren: "*Harvesting (het volledig verzamelen en uitwisselen) van metadata van in repositories opgenomen (digitale) documenten of objecten.*"
- Als werkingsgebied te definiëren: "*overheden en instellingen uit de (semi-) publieke sector*"

De consultatieronde gaf vervolgens aanleiding om drie zaken nader te onderzoeken. Dit heeft geleid tot de volgende drie onderzoeksvragen:

1. Moeten de term 'harvesting' en 'repository' in het toepassingsgebied nader gespecificeerd worden?
Hiertoe is een sessie belegd met e-Overheid voor Burgers, die dit punt heeft inbracht in de consultatieronde.

2. Is er een overlap met het toepassingsgebied van Digikoppeling (de OSB-standaarden)?

Datum
September 2010

Hiertoe is een sessie belegd met Renoir, de regieorganisatie binnen ICTU voor NUP-projecten en verantwoordelijk voor Digikoppeling.

3. Moet het werkingsgebied worden beperkt tot de archief-/educatie-/bibliotheekwereld?

Hiertoe is een sessie belegd met de Kamer van Koophandel, één van de partijen die dit punt inbracht in de consultatieronde.

De resultaten van dit onderzoek zijn verwoord in deze notitie.

2 Resultaten van het nadere onderzoek

a De termen 'harvesting' en 'repository' (e-Overheid voor Burgers)

e-Overheid voor Burgers heeft een aantal reeds bestaande toepassingen, waarbij metadata wordt verzameld van (o.a.) websites, door middel van de zogenaamde Internet Publicatie Modellen (IPM-standaarden). Deze modellen zijn (kort gezegd) gericht op het metadateren van informatie op websites van de overheid, bijvoorbeeld via een content management systeem van een gemeente. Informatie van die bronnen kan dan verzameld worden in een landelijk systeem (zoals mijnoverheid.nl).

Nadere definitie van 'harvesting'

Samen met e-Overheid voor Burgers is onderzocht in hoeverre de term 'harvesting' nader gespecificeerd zou moeten worden. Men stelt dat er onderscheid gemaakt kan worden tussen verschillende modellen voor het uitwisselen van metadata:

- een interactief pull-model (actieve bevraging)
- een niet-interactief model ('spidering' van online content)
- een push-model waarbij het initiatief van de aanbieder van metadata uit gaat

Afhankelijk van de interpretatie zouden al deze modellen onder 'harvesting' kunnen vallen. OAI-PMH richt zich echter uitsluitend op het eerste model (interactief pull-model). Voor de andere modellen biedt OAI-PMH geen functionaliteit. Om verwarring tussen de verschillende modellen te voorkomen zou in het toepassingsgebied 'harvesting' beter gedefinieerd kunnen worden als zijnde: "*het vraaggestuurd aanbieden en ophalen van verzamelingen metadata*"¹.

Toch is dit niet voldoende. Bij OAI-PMH heeft deze harvesting een specifiek doel, namelijk het opnemen van de metadata in een centrale repository. De standaard is niet bedoeld voor het stellen van een

¹ Op grond hiervan vallen de IPM-toepassingen van e-Overheid voor Burgers buiten het toepassingsgebied van OAI-PMH.

individuele zoekvraag aan een repository. Naast de eerder voorgestelde wijziging op het toepassingsgebied zou daarom ook toegevoegd moeten worden dat het doel hiervan is: *"het verzamelen van metadata in een centrale repository."*

Datum
September 2010

Nadere definitie van 'repository'

In deze context wijst e-Overheid voor Burgers ook op eventuele verwarring over het begrip 'repository'. Men heeft de vrees dat dit te breed opgevat zou kunnen worden: Is een database met persoonsgegevens bijvoorbeeld ook een dergelijke repository? Of een website, waarbij metadata is opgenomen in de header van een HTML-pagina?

Binnen OAI-PMH wordt met een 'repository' een digitale bibliotheek bedoeld. Een gangbare definitie hiervoor is: *"De verzameling van diensten en informatie-objecten die de gebruiker helpen in het omgaan met informatie-objecten en de organisatie en bewaring van deze objecten die direct of indirect beschikbaar zijn via digitale/elektronische wegen."*²

Belangrijk hierin zijn twee elementen: *diensten* en *informatie-objecten*:

- *"Informatie-objecten"* impliceert dat er sprake is van méér dan alleen *data*. Het moet gaan om *informatie*, in objectvorm. Voorbeelden daarvan zijn boeken, muziekstukken, rapporten, webpagina's of series meetwaarden. De standaard OAI-PMH maakt expliciet onderscheid tussen het informatie-object en de daarvan afgeleide of bijbehorende metadata.
- *"Diensten"* impliceert dat er bepaalde mechanismen zijn voor het omgaan, bewaren en beschikbaar stellen van deze informatie-objecten.

De aangehaalde voorbeelden zijn daarmee geen repository of digitale bibliotheek:

- Een database met persoonsgegevens bevat *gegevens* en geen informatie-objecten.
- Een website biedt weliswaar informatie-objecten aan (nl. webpagina's), maar biedt doorgaans geen diensten voor het organiseren of bewaren daarvan in de zin van de definitie.

Als digitale bibliotheek zouden wel kunnen worden beschouwd:

- Een document management systeem waarin brieven en rapporten worden bijgehouden.
- Een online bestand met videomateriaal
- Fotoarchieven
- Een index van papieren dossiers
- etc.

² Barry Leiner, *The Scope of the Digital Library*, 1998.

De term 'repository' is dan ook voldoende duidelijk. Het is niet nodig dit in het toepassingsgebied nader te specificeren. In voorkomende gevallen kan deze notitie dienen als nadere toelichting.

Datum
September 2010

b Overlap met Digikoppeling

Binnen OAI-PMH wordt gebruik gemaakt van berichten die via het HTTP-protocol worden uitgewisseld. Digikoppeling (eerder bekend als de OSB-standaarden) schrijft het gebruik van webservices (WUS) of ebMS voor. Dit is een 'rijkere' vorm van gegevenstransport, die meer mogelijkheden biedt voor o.a. beveiliging en gegarandeerde gegevensoverdracht. Een aantal partijen stelde in de consultatie de vraag in hoeverre dit met elkaar in tegenspraak is. Hierover is overleg geweest met de Digikoppeling beheerorganisatie. In dit gesprek is de interferentie tussen beide standaarden aan de orde gekomen.

Digikoppeling is ontworpen voor berichtuitwisseling tussen overheidsorganisaties en organisaties met een publiekrechtelijke taak. De standaard kent twee varianten, met ieder een eigen toepassingsgebied: WUS voor (geautomatiseerde) bevraging van informatiesystemen en ebMS voor meldingen tussen informatiesystemen. De standaard is opgenomen op de lijst voor 'pas toe of leg uit' met als werkingsgebied "sectoroverstijgend berichtenverkeer binnen de publieke sector, inclusief het verkeer met de basisregistraties".

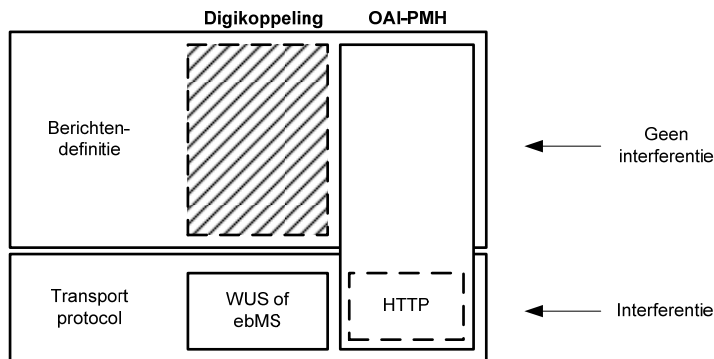
Om te bepalen of er interferentie is, moeten twee vragen worden beantwoord:

1. Is er technisch gesproken een interferentie tussen de twee standaarden?
2. Is het vastgestelde toepassings-/werkingsgebied overlappend?

Is er technische interferentie tussen OAI-PMH en Digikoppeling?

De eerste vraag is eenvoudig te beantwoorden: er is zeer zeker sprake van een technische interferentie:

- OAI-PMH beschrijft zowel de definitie van de uitgewisselde berichten als het transportprotocol dat gebruikt moet worden om deze berichten te verzenden. In het geval van OAI-PMH is dit transportprotocol 'HTTP GET/POST'; een protocol dat is opgenomen op de lijst met gangbare standaarden.
- Digikoppeling definieert enkel het transportprotocol en schrijft niet voor welke berichten uitgewisseld moeten worden. Digikoppeling gaat uit van WUS of ebMS als transportprotocol.



Datum
September 2010

Figuur: Interferentie op het niveau van berichtentransport.

Er is dus duidelijk interferentie tussen de twee standaarden op het niveau van het transportprotocol. Doordat OAI-PMH HTTP voorschrijft als transportprotocol is het niet mogelijk om voor één en dezelfde toepassing zowel Digikoppeling als OAI-PMH toe te passen.

Is het vastgestelde toepassing-/werkingsgebied overlappend?

Nu deze technische interferentie is vastgesteld, is het van belang te analyseren of er toepassings-/werkingsgebieden zijn die zowel het gebruik van Digikoppeling als OAI-PMH vereisen.

Aan het begin van deze paragraaf is het toepassings-/werkingsgebied van Digikoppeling al aangegeven. Het moet gaan om:

- a) *berichtenverkeer*
- b) dit berichtenverkeer is *sectoroverstijgend*
- c) dit sectoroverstijgende berichtenverkeer vindt plaats binnen de *publieke sector* – door Digikoppeling zelf gedefinieerd als overheidsorganisaties en organisaties met een publiekrechtelijke taak.
- d) het omvat (maar is niet beperkt tot, red.) het berichtenverkeer met de *basisregistraties*.

Er is sprake van overlap indien er een toepassingsgebied is te definiëren voor OAI-PMH dat aan deze criteria voldoet. Een analyse laat dan het volgende zien:

- De uitwisseling van metadata kan gezien worden als berichtenverkeer. Er is immers sprake van een vraag en antwoord interactie tussen twee repositories, waarbij het antwoord een set metadata is. Deze vragen en antwoorden kunnen gezien worden als berichten. Daarmee is voldaan aan punt (a).
- In de meeste gevallen zal echter niet worden voldaan aan punt (b): bij de meeste toepassingen van OAI-PMH is er sprake van uitwisseling *binnen* een bepaalde sector. Toch kan niet worden uitgesloten dat er situaties denkbaar zijn waarbij *tussen* sectoren gegevens worden uitgewisseld.
- Aan punt (c) wordt zeker voldaan – het voorgestelde werkingsgebied van OAI-PMH omvat 'overheden en instellingen uit de (semi-) publieke sector'. Dit omvat het werkingsgebied van Digikoppeling.

- Aan punt (d) wordt waarschijnlijk niet voldaan. Het verkeer met de basisregistraties is gericht op het uitwisselen van gegevensobjecten en niet op de eventuele metadata daarvan.

Datum
September 2010

Hoewel er in de praktijk overlap tussen de vastgestelde toepassings-/werkingsgebieden onwaarschijnlijk is, kan op basis van het door de expertgroep gedefinieerde toepassingsgebied overlap niet worden uitgesloten. Dit maakt het wenselijk toepassingen waarvoor Digikoppeling op basis van de lijst voor 'pas toe of leg uit' verplicht is uit te zonderen in het toepassingsgebied voor OAI-PMH³.

c Werkingsgebied: breder dan de archief-/educatie-/bibliotheekwereld?

Een aantal organisaties heeft in de consultatie aangegeven nog geen ervaring te hebben met OAI-PMH. Voor hen is onduidelijk wat de impact van OAI-PMH op hun eigen ontwikkelingen is. Zij vragen zich af of het werkingsgebied niet beperkt zou moeten worden tot de educatieve sector en de archiefsector, waar de ervaring al wel (breder) aanwezig is.

Dit vereist:

- inzichtelijker te maken of er risico's zijn voor bestaande ontwikkelingen buiten de archief-/educatie-/bibliotheekwereld.
- duidelijker te maken of er potentieel is voor de standaard buiten deze archief-/educatie-/bibliotheekwereld.

Risico's voor bestaande ontwikkelingen

Het is moeilijk om algemeen geldende uitspraken te doen over de impact van opname van OAI-PMH op ontwikkelingen bij individuele organisaties. Daarom is – om toch enigszins een beeld te krijgen – gesproken met de landelijke organisatie van Kamers van Koophandel. Deze partij heeft gereageerd in de consultatieronde.

De Kamer van Koophandel heeft een wettelijke taak daar waar het gaat om het bijhouden van een register van bedrijven en organisaties – het handelsregister. Dit register bevat eenvoudig gesteld twee gedeelten: een gestructureerde basisregistratie van organisaties (en daaraan gerelateerde gegevens, zoals vennoten, bestuurders, vestigingsplaatsen, etc.) en een verzameling onderliggende documenten (statuten, jaarrekeningen, etc.). Deze onderliggende documenten zijn in principe leidend en worden in papieren dossiers opgeslagen. Via een (papier) wijzigingsformulier kunnen bedrijven mutaties doorgeven. Deze wijzigingsformulieren worden toegevoegd aan het dossier; vervolgens vindt er een mutatie plaats in de gestructureerde gegevens. De gestructureerde gegevens worden

³ OAI-PMH is een internationale standaard; het gebruik kan ook een internationaal karakter hebben. Daarom ligt het in dit geval niet voor de hand te adviseren de standaard aan te passen aan het eventuele gebruik in combinatie met Digikoppeling.

ontsloten via het internet (KvK Online) en via mutatiebestanden die periodiek aan ketenpartners worden geleverd.

Datum
September 2010

Momenteel zijn er een aantal ontwikkelingen:

- Het gestructureerde gegevensbestand wordt uitgebouwd tot het Nieuwe Handelsregister (NHR). Dit is de wettelijk erkende basisregistratie van bedrijven.
- Er wordt nagedacht over de inrichting van het berichtenverkeer tussen het NHR en ketenpartners (via Digikoppeling en berichtstandaarden).
- De papieren dossiers blijven nodig (bijvoorbeeld voor statuten), maar zullen in toenemende mate gedigitaliseerd worden (substitutie van papieren documenten / ontwikkeling e-dossiers).

De vrees van de KvK, zoals geuit in de consultatieronde, is tweeledig:

1. Dat voor het genoemde berichtenverkeer mogelijk OAI-PMH moet worden gebruikt. De vraag is dan in hoeverre dit past bij andere ontwikkelingen (binnen de KvK en daar buiten).
2. Dat voor de ontsluiting van e-dossiers OAI-PMH zou moeten worden toegepast.

Deze zorgen kunnen weggenomen worden:

- Het genoemde berichtenverkeer valt buiten het toepassingsgebied. De uitwisseling met basisregistratie wordt met de voorgestelde aanpassing van het toepassingsgebied t.b.v. Digikoppeling zelfs volledig uitgesloten.
- OAI-PMH richt zich *niet* op ontsluiting van informatie-objecten (zoals dossiers) zelf, maar op de uitwisseling van metadata.
- Door de voorgestelde 'doelbinding' in het toepassingsgebied (verzameling van metadata met als doel dit op te nemen in een centrale repository), wordt voorkomen dat grote aantallen repositories (digitale bibliotheken, e-dossiers, document management systemen, etc.) moeten worden uitgerust met OAI-PMH, zonder dat deze repositories worden ingezet voor het aanbieden van metadata aan een centrale repository.

Hiermee zijn risico's voor bestaande ontwikkelingen nagenoeg uitgesloten. Pas indien er gekozen wordt voor een ontwikkeling die duidelijk binnen het toepassingsgebied van OAI-PMH valt, krijgt een organisatie hier mee te maken.

Potentieel

Het feit dat OAI-PMH momenteel vooral toegepast wordt binnen de archief-/educatie-/bibliotheekwereld, maakt nog niet dat het potentieel ook beperkt is tot deze sector.

Op steeds meer plaatsen wordt gewerkt met digitale informatie. Primair ligt de focus vaak op de uitwisseling van gestructureerde gegevens. Dit laat echter onverlet dat er daarnaast nog tal van andere informatie-

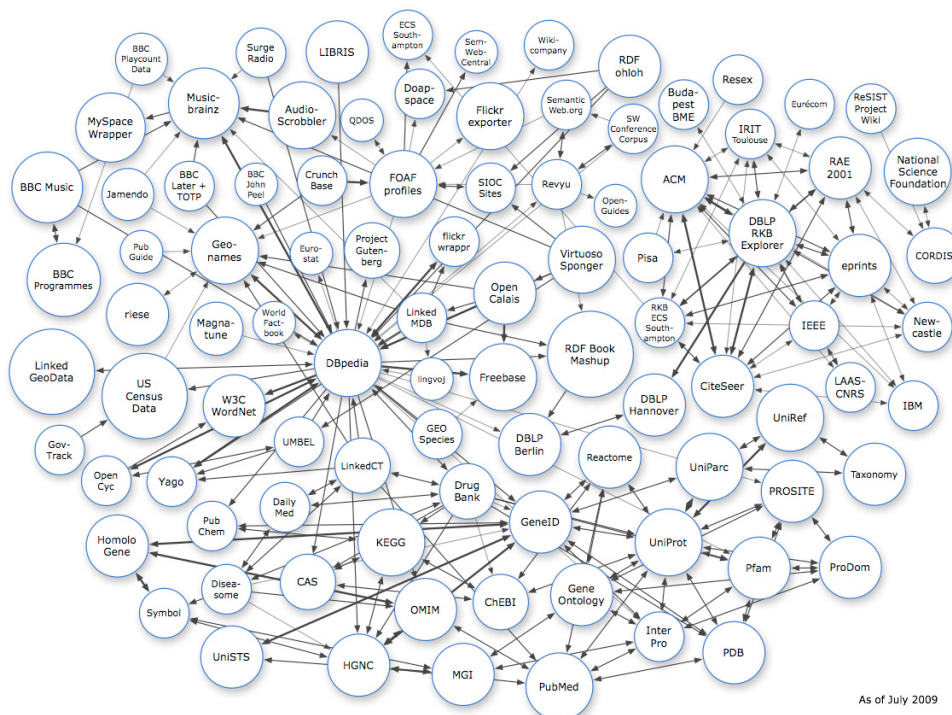
objecten zijn, die in toenemende mate gedigitaliseerd worden en online beschikbaar komen. Er zijn veel voorbeelden te noemen:

Datum

September 2010

- Kranten en nieuwsartikelen
- Fotobanken
- Vonnissen en andere gerechtelijke dossiers
- Vergunningen van overheden
- Archieven
- Wetten- en regels
- Brondossiers, zoals de aktes van de burgerlijke stand
- Online video
- etc. etc.

Het is dan van groot belang om zinvolle structuren te kunnen opzetten voor het kunnen ordenen en verbinden van content uit verschillende bronnen.



Figuur: Linked Open Data – een voorbeeld van verbonden informatiebronnen op het internet. Van Flickr.com tot US Census Data tot BBC Music

Standaarden spelen hierbij een cruciale rol. OAI-PMH maakt het mogelijk bronnen aan elkaar te verbinden en gebruikers in staat te stellen op een centrale plek te zoeken in meerdere bronnen. Deze centrale plek kan bovendien worden benut om context aan te brengen in de grote verzameling onderliggende data (bijvoorbeeld: educatieve toepassingen, juridische dienstverlening, kunst, etc.).

Naar verwachting kan opname van OAI-PMH op de lijst bijdragen aan de ontwikkeling van deze toepassingen, juist ook buiten de sectoren waar de standaard nu al veel wordt gebruikt.

Datum

September 2010

3 Conclusies

Op basis van dit nadere onderzoek kunnen de volgende conclusies getrokken worden:

Moeten de term 'harvesting' en 'repository' in het toepassingsgebied nader gespecificeerd worden?

Ja, de term 'harvesting' moet verder gespecificeerd worden. Er vallen nu nog vormen van metadata uitwisseling tussen die niet door OAI-PMH worden ingevuld.

In het toepassingsgebied moet worden gespecificeerd dat het gaat om "*het vraaggestuurd aanbieden en ophalen van verzamelingen metadata*". Daarbij is vervolgens een doelbinding noodzakelijk, het moet gaan om: "*het verzamelen van metadata in een centrale repository*". Indien een toepassing daar niet op gericht is, dan zou het gebruik van OAI-PMH niet verplicht moeten worden.

Het is niet nodig de term 'repository' nader te definiëren. In deze notitie wordt nader ingegaan op dit begrip.

Is er een overlap met het toepassingsgebied van Digikoppeling (de OSB-standaarden)?

Ja, er is potentieel een interferentie met Digikoppeling. OAI-PMH maakt gebruik van HTTP. Dit is een ander transportprotocol dan voorgeschreven door Digikoppeling. HTTP is opgenomen op de lijst met gangbare standaarden.

Hoewel naar verwachting in de praktijk de kans op interferentie gering is, zou in het toepassingsgebied expliciet gedefinieerd moet worden dat OAI-PMH niet toegepast hoeft te worden indien voor diezelfde toepassing Digikoppeling al verplicht is. Daarmee wordt interferentie in de praktijk effectief uitgesloten.

Moet het werkingsgebied van OAI-PMH worden beperkt tot de archief-/educatie-/bibliotheekwereld?

Aan de hand van de case van de Kamer van Koophandel is toegelicht dat het risico op interferentie met lopende ontwikkelingen te verwaarlozen is, indien deze ontwikkelingen niet expliciet betrekking hebben op het (verscherpte) toepassingsgebied van OAI-PMH.

Tegelijkertijd wordt gesignaleerd dat de ontwikkelingen op het gebied van digitalisering van content zich niet beperken tot de archief-/educatie-/bibliotheekwereld spelen. Standaarden, waaronder OAI-PMH, spelen een belangrijke rol bij deze ontwikkelingen.

Datum
September 2010

Dit maakt het wenselijk het werkingsgebied niet te beperken.

4 Advies

Op basis van dit additionele onderzoek wordt geadviseerd:

1. OAI-PMH versie 2.0 op te nemen op de lijst met open standaarden.
2. Het functionele toepassingsgebied (in afwijking van het eerdere voorstel) te definiëren als:
"Het vraaggestuurd aanbieden en ophalen van verzamelingen metadata uit repositories met (digitale) documenten of objecten, met als doel het opnemen van deze metadata in een centrale repository. Uitgezonderd zijn die toepassingen waarvoor op basis van de lijst voor pas-toe-of-leg-uit het gebruik van Digikoppeling verplicht is."
3. Het organisatorische werkingsgebied te handhaven, zijnde *'overheden en instellingen uit de (semi-) publieke sector'*.