

Verslag oprichtingsbijeenkomst 'nieuwe WISH'

Datum: 28-10-2014

Locatie: SURF, Graadt van Roggweg, Utrecht.

Aanwezig:

Hans Scholte (UvA), Frank Waajen (WUR), Wouter Mettrop (NWO / CWI Amsterdam), Guido van Dongen (Utrecht), Saksia Woutersen (UvA), Arent Bosman (TU Delft), Peter Verberne (UM), Wilko Steinhoff (DANS/KNAW), Jasper op de Coul (EUR), Diny in het Groen (SURF), Martin Slabbertje (UU), Inge Hofsink (KB), Jos de Groot (RU), Michel de la Rambelje (RU), René Voorburg (KB), Chris Baars (DANS), Jeroen Hamers (Edustandaard), Ariane Goossens (Edustandaard).

0. Opening en voorstelronde

Dit is de eerste bijeenkomst van de werkgroep. De aanwezigen hebben zich kort voorgesteld en hebben aangegeven welke verwachtingen er leven ten aanzien van deze werkgroep. De volgende punten zijn daarbij benoemd:

- Een beter beeld krijgen van een verdere invulling van vervolgbijeenkomsten;
- Roep om veranderingen, mee gaan met de tijd, hopelijk een vervolg van WISH;
- Benieuwd;
- Blijkbaar geen goed idee om wish op te heffen;
- Blij dat iedereen weer bij elkaar is. Is Edustandaard met name gericht op het HO en minder op het WO?
- Hoop op vernieuwing van bepaalde technieken;
- Kennisdelingsplatform, moderniseren;
- Vanuit flinke ontwikkeling vragen over aansluiten bij standaarden;
- Duidelijkheid krijgen in het toepassingsprofiel duidelijk;
- Zicht krijgen op het ontwikkelpad en op wat weggeprogrammeerd kan worden;
- Gezichten bij e-mailcontacten. De werkgroep als een plek waar we kunnen sparren en een paar stappen verder kunnen komen.

1. Edustandaard en het standaardisatieproces

De toepassingsprofielen voor MODS, DIDL, OAI-PMH en Semantics (vocabulaires) zijn belegd bij Edustandaard. Wat is de rol van Edustandaard hierin en hoe verloopt het standaardisatieproces?

Jeroen Hamers licht de structuur van Edustandaard kort toe en belicht de rol van werkgroepen, de architectuurraad en de standaardisatieraad. Bij agenda punt 4 wordt verder ingegaan op de rol van deze werkgroep. Op <http://www.edustandaard.nl/> is meer te vinden over de structuur en werkwijze van Edustandaard.

Op <http://www.edustandaard.nl/participeren/werkgroepen/werkgroep/werkgroep-metadata-subwerkgroep-onderzoeksmetadata/> is alle informatie van deze werkgroep te vinden.

Vragen:

- Hoe verhouden zich de SURF wiki en Edustandaard.nl zich tot elkaar?
>> Edustandaard.nl is leading. De standaarden, zoals MODS, DIDL, URN en OAI-PMH staan op de website van Edustandaard. Er staan op de wiki nog wel punten die wellicht bij de website van Edustandaard zouden moeten worden ondergebracht. Overige afspraken die relevant zijn voor het onderzoeksdomein die op de wiki staan en nog niet bij Edustandaard zijn aangemeld kunnen t.z.t. door deze werkgroep worden aangedragen.

- Zijn de xsd schema's ook al gepubliceerd?
>> Jeroen zal dat uitzoeken.
- Staat NL-cerif ook op edustandaard?
>> NL-Cerif betreft een virtueel datamodel. Het wordt nu vooral intern binnen NARCIS gebruikt voor de import van onderzoeksgegevens. Het gebruik ervan is nu nog vooral in projectvorm, waarin geprobeerd wordt de toepassing ervan concreter te krijgen. NL-Cerif zou ooit een standaard kunnen worden. In november is er een bijeenkomst over NL-Cerif.
- Zou MODS/DIDL vervangen kunnen worden door NL-Cerif?
>> Het is nu niet aan de orde. Een dergelijk onderwerp zou eerst in deze werkgroep besproken moet worden. Deze werkgroep is er in ieder geval verantwoordelijk voor om hier een advies over te geven aan de Standaardisatieraad.

2. Stand van zaken COAR vocabulaires

De info:eu-repo-namespace is overgedragen aan COAR (Confederation of Open Access Repositories). Wat gebeurt daar nu mee en welke gevolgen heeft dat?

Wilko Steinhoff geeft een korte toelichting.

- De info:eu-repo-name-space is overgedragen aan COAR. Het domein info:eu:repo is belegd bij COAR.
- Binnen COAR is een Vocabulairy Interest Group die zich bezig houdt met het beheren van de namespace. DANS is gevraagd lid te worden van de Vocabulairy Interest Group. Wilko doet dit namens DANS.
- Wat is de relatie met de Semantics afspraak? En hoe borg je dat de COAR-lijst niet uit de pas gaat lopen met de Semantics afspraak?
>> De werkgroep spreekt de wens uit dat er een goede terugkoppeling blijft bestaan tussen de Semantics afspraken en de begrippen in de COAR lijst. COAR zou ook iemand vanuit Edustandaard in de Interest Group willen hebben. De werkgroep vindt het niet noodzakelijk om nog iemand vanuit deze werkgroep lid te laten worden van de Interest Group. Wilko zit al in deze Group vanuit DANS. En hij zal ook als vertegenwoordiger vanuit de werkgroep Wish van Edustandaard deelnemen aan deze Interest Group.
>> Bureau Edustandaard (Jeroen) biedt aan om de rol van Edustandaard en Wilko als inhoudelijk vertegenwoordiger aan COAR toe te lichten.
- Een aantal zaken waar de Vocabulairy Interest Group zich momenteel mee bezig houdt zijn:
 - de open access repositories voorzien van gecontroleerde lijsten. Daarbij moet er rekening gehouden worden met de lokale aspecten.
 - Binnen de COAR werkgroep hebben ze het o.a. over een vocabulairelijst met oa resource types (lees publicatie types);
 - Aansluiting binnen Europa wordt groter. Ze willen het uitrollen naar alle leden. Hun doelstellingen zijn nogal groots. Niet alleen Europa, ook daarbuiten;
 - Lijsten opnemen in bestaande software pakketten;
 - Vertalen van deze lijst in een aantal talen, en verder om de lijst op te schonen. Voorjaar 2015 zal er een eerste publicatie beschikbaar zijn.
 - Mappings. Bij Edustandaard liggen de vocabulaires bij value vast. Wellicht moet dat anders.
 - Vanuit DANS staat er o.a. al een vraag uit over versiemangement. Voor COAX is het versie 1.0, voor ons is het versie 2.0.

- Net zoals in de oude WISH groep moeten we blijven uitgaan van eenvoud. COAR is tot nu toe buiten de deur gehouden i.v.m. concurrerende open air standaarden. Wegens veranderingen bij COAR lijkt dit nu minder te spelen.

Besluit:

- We blijven bij de Semantics afspraak, zoals deze bij Edustandaard in beheer is. Zodra er wijzigingen vanuit COAR worden voorgedragen, zal dit besproken worden in deze werkgroep. Andersom geldt dit ook. Deze werkgroep gaat advies uitbrengen over de door te voeren wijzigingen.
- Wilko zal Edustandaard vertegenwoordigen in de Vocabulary Interest Group van COAR.

Actie:

- Volgende bijeenkomst zal er een nieuwe update worden gegeven door Wilko.

3. Samenwerking serviceproviders en de Gemeenschappelijke Metadata Harvester (GMH) van DANS en de KB

Standaarden / toepassingsprofielen voor repositories en harvesting vergemakkelijken uitwisseling en gebruik van gegevens door bijvoorbeeld NARCIS, het KB eDepot, de nationale resolver en meer.

De KB en DANS hebben de afspraken geïmplementeerd in een Gemeenschappelijke Metadata Harvester (GMH). Wat is deze GMH, wat zijn de plannen en wat zijn de belangrijkste geconstateerde knelpunten bij implementatie van toepassingsprofielen in de GMH?

René Voorburg geeft een korte presentatie over de GMH en de daaraan gerelateerde activiteiten voor de komende tijd (bijgesloten).

KB heeft lange tijd moeite gehad met het goed harvesten van repositories DANS heeft met NARCIS minder moeite met het binnenhalen van metadata uit repositories. NARCIS is ook heel erg zichtbaar.

De samenwerking tussen KB en DANS op dit vlak was dus een prima idee. Er zijn in Nederland 3 belangrijke harvest partijen, te weten: NARCIS, KB eDepot en de nationale resolver (ook DANS). Al die 3 partijen hebben te maken met het probleem van normalisatie, synchronisatie. Het is dus veel handiger om op 1 plek de normalisatie te laten plaatsvinden, zodat er één plek is waar alles actueel beschikbaar is.

Het model waar we samen toe zijn gekomen is 1 harvester, die alle repositories harvest. De metadata uit de repository (identificer en setnaam) wordt geprefixet met een identificer voor de repository zelf.

De GMH biedt het op 3 manieren aan:

- Origineel;
- Genormaliseerd;
- Gecombineerde records

Van die gemeenschappelijke bron maken straks de 3 serviceproviders. Voor de KB speelt dat er nog een vervolgharvest nodig is. De KB is niet alleen geïnteresseerd in de metadata van repositories, maar ook in de publicaties zelf, de objecten, om deze in het eDepot op te nemen.

De techniek van de GMH is sterk gebaseerd op de techniek van NARCIS (software van Seeker). Als er geharvest wordt, wordt er eerst gevalideerd tegen schema's (MODS, DIDL).

Vervolgens vindt er een normalisatieslag plaats. Dat wringt een beetje, want het liefst heb je normalisatie niet nodig. We streven ernaar om zoveel mogelijk samen via de standaard werken en de verschillen die er zijn via normalisatie glad strijken. Voorbeeld: datumformaten. Daar zit de nodige variatie in. Normaliseren zoveel mogelijk beperken.

Stand van zaken:

We zijn nu bezig met het finetunen van de werking van de GMH. De meeste repositories komen vrij vroeg binnen en we zijn bezig met het contact opnemen met de repositorybeheerders om aan te geven wat er nog mis gaat en waar nog naar gekeken moet worden. Er zijn nog een aantal repositories die nog wat meer problemen geven, maar dat heeft waarschijnlijk te maken met de migratie naar Pure. We hopen op korte termijn (dit jaar) in de afstemming met jullie ervoor te zorgen dat het aantal fouten flink gereduceerd wordt. KB is van plan om de GMH als basisbron voor de harvest-activiteiten van eDepot in te zetten. In 2^e instantie zullen ook de nationale resolver en NARCIS gebruik gaan maken van de GMH. Het idee is dat er uiteindelijk 3 partijen gebruik gaan maken van de GMH.

Opmerking: Belangrijk in het verder gaan gebruiken van de GMH is om eerst de content op orde te krijgen. Nu kan bijvoorbeeld de situatie voorkomen dat bv 1000 records niet worden opgenomen, omdat ze niet gevalideerd zijn. Die komen dan niet in NARCIS terecht. We moeten dit niet alleen zien als problemen, maar ook als een verbeteringsslag.

Lijstje van dingen die mis gaan (indicatieve top 10):

- Geen valide MODS genre (publicatietype). er staan meerdere lijstjes op internet, maar er is nu ook een mis-match tussen het MODS document en de Semantics document;
- Staat er expliciet in de standaard dat URL geëncodeerd moeten worden? Impliciet moet dat wel;
- Datum-veld ontbreekt;
- Geen access rights opgenomen;
- Resource location kan niet gevonden worden;
- Lege titels;
- Ongeldige uri's (niet toegestane tekens);
- Identifier die niet valide is;
- Ongeldige taalcodes;

Er zal een belronde georganiseerd om de fouten te checken. Wat niet genormaliseerd kan worden, komt in het lijstje van fouten en dan komt het niet verder.

Vragen / Wensen uit de werkgroep:

- Er is behoefte aan gebruikersvriendelijke foutenlijsten;
- Is er een on-line tool om metadata weer op te halen?
 - o Deze is er, nl oai-pmh-repository: <http://oai.gharvester.dans.knaw.nl/>
- Is de software open source? Ja, in principe wel. Deze is wel wat aangepast. Wilko geeft aan dat dit inzichtelijke gemaakt kan worden.
- Wanneer fouten genormaliseerd worden, kun je als instelling niet zien welke fouten je wellicht herhaaldelijk maakt. Kan dit inzichtelijk gemaakt worden?
 - o Hier is een RSS feed voor. Deze kan eventueel opengesteld worden.
- Er wordt nog bekeken op welke wijze er gezorgd kan worden voor 1 loket, bv door een gemeenschappelijk email adres. Nu heeft iedereen nog te maken met René en Ishan.
- Het lijstje met de 10 meest voorkomende fouten kan mooi dienen als input voor nieuwe standaarden / aanpassingen bestaande standaard;
- Wordt er gekeken naar de nieuwe standaard Resource Sync?

- Er wordt wel naar gekeken. Maar omdat er al zoveel is fout gegaan, lijkt het beter om eerst te proberen om met de bestaande afspraken zo ver mogelijk te komen. Dit is zeker wel een ontwikkeling die gevolgd wordt.
- Zijn er al mensen die benaderd zijn door Google? Google wil de metadata zoveel mogelijk verrijken voor Google Scholar. De meningen over de bruikbaarheid zijn verdeeld. Dit is wel een onderwerp waar we meer van willen weten, maar op een later moment.
- Wordt de geaggregeerde data straks doorgestuurd (zoals nu door NARCIS) en welke set wordt dat?
 - Niet alle repositories die in NARCIS zitten, zitten nu in de GMH. Wat er uitgeleverd gaat worden moet nog uitgezocht worden.

Actiepunten:

- René komt de volgende bijeenkomst met een lijstje als input.
- Ishan en René nemen contact op met een overzicht van wat er nu mis gaat. om fouten helder te kunnen krijgen.

4. Voorstel start werkgroep metadata

Voorstel is om een werkgroep Metadata te starten, à la 'WISH' met een tweeledig doel:

- 1) voorbereiden van wijzigingsvoorstellen voor de bij Edustandaard belegde toepassingsprofielen en
- 2) kennisdeling ten aanzien van de implementatie en toepassing van de standaarden en bijbehorende toepassingsprofielen.

Besluit: Doorgaan met de werkgroep.

Werkwijze:

- Om te beginnen 4 keer per jaar bij elkaar komen.
- LinkedIn groep (zoals gedaan wordt in andere ES werkgroepen) nog niet nodig.
- Agendapunten in kleine werkgroepjes uitwerken/voorbereiden.

De groep geeft aan dat er, naast een technische groep zoals deze ook behoefte is aan een groep a la WRM (= werkgroep repositories managers, klankbord groep). De WRM kan gezien worden als een functionele werkgroep. De WRM gaat bijvoorbeeld tijdens een bijeenkomst over 2 weken een voorstel doen om open access tellingen voortaan bij NARCIS te laten doen. Dit heeft impact op wat er in deze werkgroep op technisch niveau besproken moet worden. Er is afstemming nodig met de WRM.

Actiepunt:

- Jeroen neemt contact op met Elly Dijk van DANS om afstemming te zoeken met de WRM.

5. Discussie: hoe gaan we verder, wo. inventarisatie aandachtspunten korte en lange termijn

Er volgt enige discussie om grip te krijgen op wat de volgende bijeenkomst geagendeerd kan worden. Waarschijnlijk zullen er ook een aantal zaken uit de bel- of mailronde komen, die René en Ishan gaan houden. Deze zaken dienen geprioriteerd te worden. Ook komt het aantal keren dat deze werkgroep jaarlijks bij elkaar gaat komen ter sprake.

- Vraag: Zijn er mensen uit deze groep die de bijeenkomsten van de Edustandaard Vocabulaires werkgroep bezoeken? Dat is op zich een interessante werkgroep, maar

deze is heel erg gefixeerd op het HO en niet op het WO. Dit geldt voor meerderen van de aanwezigen.

>> Jeroen geeft als toelichting dat de werkgroep Vocabulaires wel degelijk over alle sectoren heen gaat, maar zich richt op de onderwijskant. Alle onderzoeks-gerelateerde standaarden, inclusief vocabulaires, horen bij deze werkgroep. Daarom is de semantics standaard (wat eigenlijk vocabulaires is) niet ondergebracht bij de werkgroep Vocabulaires, maar bij de werkgroep Wish.

>> In de beginfase van Edustandaard bestond er één metadata werkgroep. Deze is gesplitst in een werkgroep metadata voor onderwijs en een werkgroep metadata voor onderzoek. Jeroen kan wel een keer presenteren wat er besproken wordt in de werkgroep Vocabulaires en welke hulpmiddelen daar gebruikt worden. Er bestaat ook het OBK (onderwijsbegrippenkader), waar een namespace is voor onderwijsbegrippen. Deze zou je in principe ook voor onderzoeksbegrippen kunnen gaan gebruiken. Mocht dat met COAR niet lukken. Maar dat is niet de hoogste urgentie.

Actiepunt Jeroen: presentatie werkgroep Vocabulaires.

- Het top-10 lijstje van René.
>> René geeft aan dat we niet alles in één keer moeten willen bespreken. Er is één ding dat bovenaan het lijstje staat, namelijk het Mods-genre publicatietype. Er is hierin nu onduidelijkheid bij de Edustandaard afspraken. Vragen zijn bv wat precies de publicatietypes zijn die we accepteren en hoe we ermee omgaan wanneer we iets tegenkomen dat niet op het lijstje staat. Ook de mapping van Metis naar Semantics wordt genoemd. Tevens zou het handig zijn een samenvatting te hebben van wat er te vinden is aan publicatietypes: ze zijn verschillend, maar waarom is niet duidelijk.
Actiepunt René: Voor volgende bijeenkomst een uitgewerkt voorstel hoe hier meer duidelijkheid in te scheppen.
- Is wat op de wiki staat in voldoende mate bij Edustandaard ondergebracht?
>> De documenten bij Edustandaard lijken niet compleet, de documenten in de surfwiki lijken verouderd. Hoe gaan we om met de borging van de documenten die er zijn en de update van de documentatie?
>> In de documentatie van de standaarden bij Edustandaard kunnen dingen staan die afwijken of verschillend geïnterpreteerd kunnen worden van wat in de schema's staat bij SURF. Dat zou besproken moeten worden. Ook de schema's dienen uiteindelijk bij Edustandaard ondergebracht te worden. Edustandaard stelt hier wel eisen aan.
>> Bv de DAI. Het DAI xsd schema is niet bij Edustandaard ondergebracht. DAI staat wel in profiel. Peter doet wat voorwerk, Jeroen vult aan.
Aktiepunt Peter: overzicht maken van te gebruiken xsd's. Jeroen vult dit eventueel aan.
- Oris heeft een aantal subwerkgroepen gehad. Die hebben beschreven wat er allemaal ligt. Tevens hebben zij een aantal aanbevelingen gedaan ten aanzien van wat er nog ligt aan werkpakketten. Het zou goed zijn om te zien of we dit kunnen updaten en bij Edustandaard kunnen onderbrengen.
Aktiepunt iedereen: bekijken van de ingangsdocumenten ORIS voor de standaardisatieraad van Edustandaard en prioritering in aandachtspunten/aanbevelingen aangeven. De Link naar deze documenten is:
Aktiepunt iedereen: eventuele eigen issuelijstje erbij betrekken.
- 1 loket voor vragen:

>> Normaal komen vragen rond werkgroepen binnen bij het emailadres: info@edustandaard.nl. Dit is een prima werkwijze, alleen is dan wel iemand met verstand van zaken, die dat soort vragen kan oppakken voor deze werkgroep.

>> Als reactie daarop wordt aangegeven dat het aanmelden voor harvesten lastig is. In de praktijk mailen mensen naar NARCIS, waarna zaken eventueel moeten worden doorgezet. Ook hier zou het fijn zijn wanneer er één email adres is. Dit punt zou je dan ook kunnen gebruiken als punt voor vragen. Er is duidelijk een meerwaarde voor één loket. Hier zal intern binnen DANS over gesproken worden. Daarna zal er een voorstel voor de werkgroep komen.

Actiepunt Chris: voorstel van 1 loket intern bespreken en maken van voorstel.

6. Wvttk & sluiting

Iedereen wordt bedankt voor zijn/haar aanwezigheid. Er zal door Edustandaard een miltje worden verstuurd om een nieuwe datum te prikken. Eerst volgende bijeenkomst zal in februari gepland worden. In deze maand valt de voorjaarsvakantie, waar rekening mee gehouden zal worden.

Sluiting: 16:30 uur.