Europeana en de digitale ontsluiting van cultureel erfgoed
Brussel, 16 december 2009

Workshop 3: Aggregatie

Wat is aggregatie?

> *Van Dale*:
> **Aggregaat**: vereniging of ophoping van niet scheikundig verbonden gelijk- of ongelijksoortige stoffen tot een geheel

In allerlei sectoren en vakgebieden krijgt aggregaat/aggregatie een meer specifieke betekenis. Ook in de wereld van ICT, maar daar gaat het meestal nog om een generisch gebruik van het woord. Bijvoorbeeld:

> *nl.wikipedia.org*:
> **Aggregaat (document)**: Een aggregaat is een uit verschillende delen samengesteld document of verzameling documenten die als samenstelling auteursrechtelijk behandeld wordt. Het begrip aggregaat is van belang bij het in licentie geven van documenten zoals onder de GNU Vrije Documentatie Licentie (GFDL).

> *en.wikipedia.org*
> **Aggregator**: In general internet terms, a news aggregation website is a website where headlines are collected, usually manually, by the website owner.

Het gaat dus over informatie uit verschillende bronnen die wordt verzameld om samen op een website ter beschikking gesteld te worden. De aggregatie krijgt daardoor een eigen identiteit en vaak ook een gezicht.

Aggregatoren voor Europeana

Toen Europeana op zoek ging naar een beheerbare manier om gegevens te verzamelen van duizenden verschillende bronnen, is daar voor de Europeana context een zeer specifieke definitie bijgekomen:

> *Europeana Content Strategy*
> **Aggregator**: An Aggregator is an organization that collects metadata from its group of content providers and transmits them to Europeana, helps content providers with guidance on conformance with Europeana norms and converts metadata if necessary. The aggregator also supports the content providers with administration, operations and training.
> A Content Provider (CP) is any organization that provides digital content for access via Europeana and the metadata that enables the access.

Hierbinnen zijn nog verschillende opties:

- men kan de volledige digitale objecten met hun metadata aggregeren (dan spreekt men van een repository), ofwel enkel metadata met doorverwijzingen (URL links) naar de objecten;

- men kan het aggregaat voorzien van een aparte website voor ontsluiting op internet, ofwel enkel de informatie verzamelen en ter beschikking houden voor gebruik door anderen.

Volgens de geaggregeerde inhoud spreekt Europeana verder van:
- verticale aggregatoren: met informatie beperkt tot een specifieke sector (vb. musea, archieven), maar niet geografisch beperkt
- horizontale aggregatoren: met informatie uit verscheidene sectoren (cross-domain), maar geografisch beperkt (vb. regio, land)
- thematische aggregatoren: met informatie uit verscheidene sectoren en regio's, maar gerelateerd aan een specifiek onderwerp

De verscheidene Europeana satellietprojecten onderzoeken relevante aspecten van deze opties.

Europeana organisatiemodel

Voor het verzamelen van de inhoud, verkiest Europeana te werken via de bemiddeling van aggregatoren, van welke aard ook. Rechtstreeks aanleveren van informatie door een content provider wordt echter niet uitgesloten.

Om duplicatie zoveel mogelijk te vermijden wordt ernaar gestreefd elke content provider slechts langs één aggregator informatie te laten aanleveren.

Er is een flowchart gemaakt om voor elke content provider te kunnen beslissen langs welk kanaal de informatie geleverd kan worden.

Informatie zou ter beschikking gesteld moeten worden voor harvesting in ESE formaat, via OAI-PMH (zie verder).

Content providers en aggregatoren krijgen inspraak in het Europeana beheer en beleid via de CCPA (Council of Content Providers and Aggregators), met verkozen vertegenwoordigers die deel uitmaken van de beslissingsstructuren.

Huidige beperkingen/knelpunten:
- ESE specificaties zijn geldig voor de Rhine release; voor daarna zijn nog geen specificaties beschikbaar
- Er zijn nog geen afspraken over persistente identificatoren voor aangeleverde objecten
- Er zijn nog geen afspraken voor het updaten van informatie
- Auteursrechten blijken in vele gevallen nog problematisch
- Er zijn nog maar weinig aggregatoren

Businessmodellen voor aggregatoren

Het is weinig waarschijnlijk dat iemand een aggregator zal bouwen met als enige bedoeling gegevens te verzamelen als tussenstap naar Europeana. Aggregatoren moeten hun eigen businessmodel ontwikkelen om de investeringen te verantwoorden. Dit businessmodel bepaalt welke soort aggregator er gemaakt wordt (horizontaal, verticaal, thematisch), en welke bijkomende diensten worden aangeboden aan de leveranciers van informatie (content providers) of aan de mogelijke afnemers van de geaggregeerde gegevens.

‘Interoperability’

Interoperabiliteit(?) is de sleutel om een aggregator mogelijk te maken. Om gegevens uit verschillende bronnen te verzamelen en er een nieuw geheel van te maken, moeten ze vergelijkbaar en op mekaar afgestemd kunnen worden: technisch, vormelijk en inhoudelijk.

Technische interoperabiliteit: ‘harvesting’

De aggregator moet de digitale bronbestanden kunnen lezen en opladen.

Aan de basis van digitaal uitwisselen van gegevens ligt de algemene internet coderingstaal XML. De meeste database systemen kunnen hun informatie exporteren in XML formaat. Een XML bestand kan vrij makkelijk getransformeerd worden naar andere formaten mits ook wat inhoudelijke afspraken gerespecteerd worden.

Er wordt naar gestreefd het ophalen van de informatie bij de content providers zoveel mogelijk te automatiseren. Bij Europeana zal dit gebeuren d.m.v. de internet protocol OAI-PMH (Open Archives Initiative – Protocol for Metadata Harvesting). De meeste aggregatoren (in wording) werken ook met deze methode en protocol. Dit sluit wel de deelname uit van kleinere instellingen die niet de middelen hebben om hun database op internet ter beschikking te stellen. Voorlopig houden weinig aggregatoren hiermee rekening.

Structurele interoperatiliteit: ‘mapping’ en ‘normalisation’

De datastructuren van de brondatabase en de aggregator zullen doorgaans verschillen. Dan moet de veldstructuur van de bron gemapt worden naar de veldstructuur die door de aggregator gebruikt wordt. Dit kan gebeuren bij de bron door de content provider, bij de aggregator, of tussenin, ingebouwd in het harvesting proces. De meeste software systemen voor repositories en harvesting hebben een eenvoudige mapping functie ingebouwd, en soms zelfs enkele mappingprocessen tussen zeer gangbare standaarden voorgedefinieerd.

De datastructuur van de aggregator varieert naargelang de doelstellingen van de aggregator. In elk geval wordt informatie samengevoegd die afkomstig is uit verschillende contexten, vaak ook aangemaakt met verschillende bedoelingen. Die datastructuur is daarom meestal eenvoudiger dan die van de bronnen, met minder mogelijkheid tot onderscheiden van specifieke soorten informatie. In de aggregatie zal dus altijd wat informatie verloren gaan, omdat ze niet precies past in de nieuwe datastructuur. Aggregatoren kunnen daarom nooit de broninventarissen of catalogi vervangen, tenzij ze dit specifiek als dienst aanbieden en ervoor uitgerust zijn.

Doorgaans gebeurt in deze fase ook een normalisatie, waarbij de inhoud van bepaalde velden wordt aangepast aan een uniforme standaard; bijvoorbeeld datums.

Om mapping en normalisatie te kunnen doen moeten de brongegevens voldoen aan één fundamentele eigenschap: interne CONSISTENTIE

- geen twee soorten informatie in eenzelfde veld doorheen de database
- slechts één informatieeenheid per veld
- dezelfde soort informatie steeds in hetzelfde veld
- constant systeem van noteren van numerieke waarden, datums, en termen gekozen uit korte opsommingslijstjes
- trefwoordenlijsten consistent en zuiver houden

Inhoudelijke interoperabiliteit: 'enrichment'

De meeste aggregatoren, vooral als ze een eigen publieksinterface hebben, pogen de verzamelde gegevens ook inhoudelijk op een uniform toegankelijke manier doorzoekbaar te maken. Dit is allicht de moeilijkste opdracht bij aggregatie. De gegevens zijn meestal niet ontstaan voor uitgebreide publieke toegang, en zijn opgesteld binnen de context van elke individuele collectie. Door de aggregatie verliezen ze die band met hun oorspronkelijke context en worden ze in een nieuwe context geplaatst, waarvan het vaak niet mogelijk is te voorzien wie de gebruikers zullen zijn. De nieuwe context wordt gevormd door de confrontatie met gegevens uit andere bronnen, en de onderlinge relaties.

De 'semantic web' technologie is ontwikkeld om dit soort relaties te kunnen behandelen. Maar de technologie is slechts een deel van de oplossing. Het herkennen van de relaties is zeer moeilijk te automatiseren, vooral omdat we met metadata niet veel contextgegevens kunnen overdragen. Vaak verschillen specifieke betekenissen grondig in andere contexten.

Thesauri

Een belangrijke manier om inhoudelijke interoperabiliteit te verwezenlijken is het gebruik van trefwoordenlijsten, of beter thesauri. Trefwoordenlijsten zijn doorgaans platte lijsten van termen waaruit kan gekozen worden. Nieuwe trefwoorden kunnen meestal ook vrij gemakkelijk aangemaakt worden binnen een systeem, wanneer de gepaste term nog niet voorhanden is. Moderne thesauri zijn (zouden moeten zijn!) meer gestructureerde lijsten van concepten die verduidelijkt worden door middel van relaties tussen de concepten en die aangeduid kunnen worden met voorkeurstermen of alternatieve termen. Een goede thesaurus is normaal ook het product van afspraken en samenwerking tussen verschillende specialisten binnen een specifiek domein. Een goede thesaurus zou echter voldoende gegevens over de concepten moeten bevatten, om ze ook buiten de context waarin ze ontwikkeld zijn begrijpbaar te maken (BT-NT hiërarchie, taxonomie, Scope Notes).

Cross-domain thesauri zijn ontzettend moeilijk te maken. Niet om mogelijke technische beperkingen, maar omwille van de vrijwel oneindige discussies tussen de specialisten uit verschillende domeinen en contexten. Doorgaans blijven zulke thesauri dan ook op een vrij algemeen niveau. In de erfgoedsector en in het semantic web verwachten we echter een grote specificiteit.

Voorbeeld: AAT / AAT-Ned

In de erfgoedsector hebben we geluk. Er bestaat een thesaurus die ontworpen is voor de ganse breedte van het erfgoedveld. De AAT (Art & Architecture Thesaurus) wordt sinds de jaren '70 ontwikkeld en onderhouden door de instellingen van de Getty Foundation. Er is een Nederlandse vertaling, AAT-Ned en een Spaanse. Delen zijn ook in Canada naar het Frans vertaald. De AAT is uiteraard niet ontstaan voor het semantic web, maar is wel geëvolueerd naar een instrument dat uitstekend past in het concept en de technologie van het semantic web.

De AAT is niet perfect, en de AAT-Ned nog iets minder. Iedereen die eraan denkt de AAT te gebruiken ziet wel fouten en heeft wel wat ideeën voor verbetering. Voorlopig is de AAT nog vrij statisch. Er wordt aan gewerkt om de AAT-Ned een veel dynamischer karakter te geven,

aangepast aan de huidige web 2.0 realiteit. De inhoudelijke specialisten zitten namelijk zeker niet bij mekaar in grote documentatiecentra, maar zijn verspreid in het ganse erfgoedveld. Het uitbouwen en onderhouden van een goede erfgoedthesaurus vraagt een bijdrage van al die specialisten. Dit proberen we op dit ogenblik te organiseren voor het Nederlandse taalgebied. Met de AAT vertrekken we daarvoor niet van nul, maar kunnen we voortbouwen op de ervaring en het werk van 30 jaar ontwikkeling.

De AAT biedt geen oplossing voor alle nodige facetten in erfgoeddocumentatie. Er is ook nood aan goede afspraken en zo mogelijk uniforme lijsten voor plaatsaanduiding en voor personen en instellingen.

Thesauri voor Europeana?

Het gebruik van thesauri krijgt een veel sterker belang in de context van aggregatoren, dan wanneer men het gebruik voor de eigen collectie beschouwt. Met de groei van behoefte aan uitwisseling en overkoepelende databases gedurende de laatste jaren, wordt daarom de noodzaak van goede thesauri sterker dan ooit.

In het concept van Europeana worden thesauri zeker ook als essentiële voorwaarde onderkend. In het datamodel worden twee lagen onderscheiden: de 'surrogaten'-laag, waarin de verzamelde metadata over digitale objecten terecht komen, en de 'semantische' laag, waar de samenhang tussen de objecten wordt uitgebouwd, en die de eindgebruiker zal toelaten op een betekenisvolle (semantische) manier te zoeken en navigeren in de informatie. Die semantische laag bestaat eigenlijk nog niet in het huidige prototype en in de Rhine release. Maar de Danube release zou wel moeten beginnen ze te gebruiken. Voorlopig is er echter nog geen duidelijk zicht op hoe de invulling ervan zal gebeuren., maar dit is waar alle gebruikte thesauri gebruikt door de content providers aan mekaar gekoppeld zouden moeten geraken.

Ondertussen hebben we er alle voordeel bij zo breed mogelijk afspraken te beginnen maken over het gebruik en de opbouw van onze thesauri. Een breder draagvlak verbetert de kwaliteit van onze eigen informatie en vermindert aanzienlijk de inspanning die later zal moeten gebeuren om aan te sluiten bij grotere semantische initiatieven.

Voorbeelden

Nationale aggregatoren:

- Frankrijk: http://www.culture.fr
- Italië: http://www.culturaitalia.it
- Duitsland: http://www.bam-portal.de
- Oostenrijk: http://www.kulturpool.at

In België:

- Oost-Vlaanderen: MovE http://www.museuminzicht.be
- Limburg & Vlaams-Brabant: http://www.erfgoedplus.be
- Vlaamse Kunstcollectie: http://www.vlaamsekunstcollectie.be
- Religieus erfgoed (CRKC): http://www.religieuserfgoed.be

Hoe deelnemen in Europeana?

Europeana verkiest dat individuele content providers hun gegevens ter beschikking stellen via aggregatoren. Er zijn wel enkele algemene voorwaarden die moeten vervuld zijn om bruikbare informatie te kunnen leveren aan Europeana. Vaak vragen die wat voorbereidend werk van de content provider.

Algemene voorwaarden:

- inhoud geschikt voor Europeana
    - digitale objecten zoals voorzien in Europeana
    - voldoende metadata
    - metadata converteerbaar naar ESE
- vrij van auteursrechten
- digitale objecten bereikbaar op internet d.m.v. een URL
- geschikte aggregator zoeken
    - Musea: Athena
    - Oost-Vlaanderen – Limburg – Vlaams-Brabant: EuropeanaLocal via MovE en Erfgoedplus.be
    - Andere: er zijn nog andere Europeanaprojecten: zij hebben hun website en vaak ook partners in België (lijst van projecten op http://group.europeana.eu)

Checklist voor voorbereiding van gegevens

Er zijn een aantal dingen die potentiële content providers in hun eigen database kunnen verifiëren en eventueel verbeteren om de opname in Europeana te vergemakkelijken en met goed resultaat te laten verlopen.

Dit zijn aandachtspunten en acties die niet alleen in functie van Europeana van belang zijn, maar die in het algemeen de kwaliteit van de databasegegevens bevorderen voor intern gebruik zowel als voor mogelijke publicatie en hergebruik.

Sleutelvragen

- Welke regels worden toegepast?
- Worden de regels gevolgd op consistente wijze?
- Is er voldoende informatie om de objecten te identificeren en vindbaar te maken?
- Kan de informatie begrepen worden buiten de context van mijn collectie?

Systemen

- Gegevens exporteerbaar in een XML formaat, dat het gebruikte datamodel getrouw weergeeft
- Gecontroleerde invoer van relevante velden d.m.v. trefwoordenlijsten (bij voorkeur met thesaurus functionaliteit)

Datastructuren

- Breed gedragen standaard gebruiken
- Standaard strikt en consistent toepassen
- Toegepaste opties of afwijkingen goed documenteren, vastleggen in richtlijnen en consistent toepassen

Thesauri

- Bestaande breed gedragen thesuari/trefwoordenlijsten kiezen
- Consistent toepassen en afwijkingen of toevoegingen documenteren
- Termen de-contextualiseren door de betekenis expliciet toe te voegen (hiërarchie, scope notes)
- Termen/concepten moeten identificeerbaar zijn, ook buiten de context van de eigen collectie

Afbeeldingen

- Logische folderstructuur toepassen
- Eigen regels afspreken en consistent toepassen
- Gangbare bestandformaten gebruiken