De kracht van genetwerkte terminologiebronnen: CATCHPlus en de Cultural Commonwealth

Johan Oomen en Hennie Brugman

Door een combinatie van uiteenlopende factoren zoals de adaptatie van standaarden, massa-digitalisering, verwachtingen van eindgebruikers, alternatieve licentiemodellen en nieuwe initiatieven van marktpartijen is in relatief korte tijd een heel nieuw denken over toegankelijkheid van erfgoed ontstaan. Initiatieven als Google Books, Open Content Alliance, Wikimedia Commons en Flickr The Commons hebben er in enkele jaren toe geleid dat tientallen miljoenen objecten van erfgoedinstellingen online te vinden zijn.
Europeana, dat toegang biedt tot 10 miljoen objecten van honderden instellingen, mag als voornaamste Europese initiatief in dit rijtje uiteraard niet ontbreken. Europeana laat zien hoe collecties uit traditioneel verkokerde domeinen zoals bibliotheken, archieven, musea en audiovisuele archieven online met elkaar in verbinding gebracht worden. Nieuwe webstandaarden maken het mogelijk om de online toegang tot collecties van een nieuwe betekenislaag te voorzien, onder meer door impliciete links tussen objecten uit gedistribueerde collecties expliciet te maken. In dit artikel wordt kort ingegaan op de voorname rol die het online publiceren van terminologiebronnen hierbij speelt.

Cultural Commonwealth

Erfgoedinstellingen dragen, door het online beschikbaar stellen van collecties, bij aan een nieuw informatie ecosysteem waarbinnen digitale objecten en contextuele informatie een plaats krijgen binnen een steeds omvangrijker wordend netwerk van informatiebronnen. Op het web ontstaat zo een Cultural Commonwealth; waarbinnen erfgoedcollecties op een revolutionair nieuwe wijze bestudeerd, gecontextualiseerd en gerepresenteerd worden. (Our 2009) In een recente publicatie schrijft Stefan Gradmann over deze verschuiving in het denken over het toegankelijk maken van collecties: “This mentality shift is a big leap, since it requires cultural heritage institutions to think, not primarily within the boundaries of their particular collections, but in terms of what these collections might add to a bigger, complex and distributed information continuum coupled with various contextual resources enabling […] users to turn partial aggregations of this continuum into knowledge that is relevant in their specific context” (Gradmann 2010). In Nederland is het CATCHPlus project, als knooppunt van erfgoed en informatica, een belangrijke aanjager bij de totstandkoming van deze Cultural Commonwealth.

Het CATCHPlus project

CATCHPlus is een driejarig project (2009-2011) waarin erfgoedinstellingen, kennisinstellingen en het bedrijfsleven samenwerken aan het ontwikkelen van innovatieve technieken die de toegankelijkheid van collecties verbeteren. CATCHPlus bouwt voort op het lopende NWO-programma CATCH. Het project ontvangt subsidie van het interdepartementale Programma Implementatie Agenda ICT-Beleid, het Ministerie van OC&W en van NWO en levert de volgende technische resultaten op:

- gemeenschappelijke diensten, zoals de vocabulaire repository en een annotatie dienst
- software geënt op de praktijksituatie van de deelnemende grote erfgoedinstellingen
- implementatie van een persistente identifier-resolverstructuur

Op de website van CATCHPlus is informatie te vinden over de projectresultaten. In dit artikel wordt dieper ingegaan op één van de gemeenschappelijke diensten die zijn ontwikkeld, de Vocabulary and Alignment Repository.

Online publiceren en gebruiken van terminologiebronnen

Terminologiebronnen, zoals thesauri, zijn primair een communicatiemiddel, bedoeld om eenduidiger of met een bepaalde focus collecties en objecten te beschrijven.[1] Ze spelen een rol bij zowel collectieontsluiting als bij het zoeken en grasduinen door collecties. In het erfgoedveld zijn verschillende terminologiebronnen in gebruik voor het beschrijven van de uiteenlopende collecties.

[1] Een uitputtend overzicht van terminologiebronnen in de erfgoedsector is te vinden op de website van DEN http://tinyurl.com/38nyclx

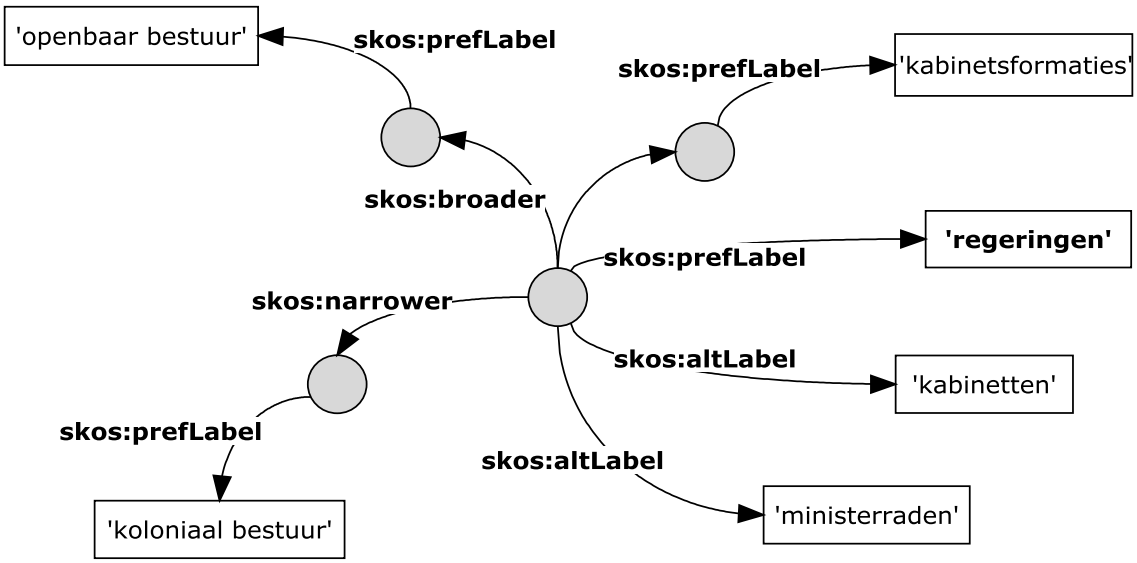
Dit maakt het lastig deze collecties gelijktijdig te doorzoeken en zoekresultaten te vergelijken. Om dit te ondervangen worden door verschillende projecten en organisaties verbanden tussen concepten in verschillende thesauri gelegd. Dit gebeurt vaak handmatig, maar er worden ook met toenemend succes pogingen tot het automatische koppelen ondernomen. Vanwege hun communicatiefunctie is het wenselijk, dat thesauri zo openbaar en toegankelijk mogelijk zijn. Teneinde bestaande technische en juridische barrières te slechten is in het kader van CATCHPlus een gemeenschappelijke 'Vocabulary and Alignment Repository' gebouwd.

Deze repository maakt het mogelijk verschillende terminologiebronnen volgens een gestandaardiseerd formaat (SKOS, zie hieronder) op te slaan. Ieder van deze bronnen is vervolgens op uniforme manier via het web te raadplegen. De Vocabulary and Alignment Repository biedt daarnaast de mogelijkheid koppelingen tussen concepten uit verschillende thesauri ("alignments") op te slaan en te doorzoeken vast te leggen en online beschikbaar te maken. De repository is te gebruiken door individuele organisaties, door consortia van instellingen of erfgoedbreed. De huidige versie van de repository bevat een twaalftal erfgoedthesauri. Dit aantal zal de komende jaren groeien.

Teneinde terminologiebronnen van verschillende instellingen op een gestandaardiseerde wijze te representeren, wordt gebruik gemaakt van het Simple Knowledge Organisation Scheme, ofwel SKOS. SKOS is een W3C aanbeveling voor het representeren van terminologiebronnen. (Miles 2009) Het is gebaseerd op bestaande ISO standaarden (zoals ISO 2788) maar, in tegenstelling tot deze standaarden, is SKOS niet term-gebaseerd, maar concept-gebaseerd. In plaats van voorkeurs- en niet-voorkeurstermen, die onderling naar elkaar verwijzen kent SKOS concepten, waaraan de verschillende termen als bundels van alternatieve tekstlabels kunnen worden gekoppeld. Dergelijke concepten hebben een unieke identifier, een zogenaamde URI, waarnaar verwezen kan worden. Traditionele hiërarchische en associatieve thesaurusrelaties zijn binnen SKOS relaties tussen concepten, niet tussen termen (zie figuur 1).

Bovendien kent SKOS al een aantal relaties, zoals broadMatch, relatedMatch of exactMatch, waarmee concept-koppelingen tussen thesauri kunnen worden weergegeven. Omdat SKOS is gebaseerd op Semantisch Web technologie is het aantal mogelijke soorten relaties tussen concepten naar behoefte uit te breiden. Voordelen van het gebruik van SKOS zijn onder meer betere mogelijkheden tot onderhoud, uitbreiding en koppeling.

Conversie van bestaande terminologiebronnen naar SKOS is technisch niet erg lastig. Wel moeten soms concessies worden gedaan om bepaalde constructies in een specifieke thesaurus af te beelden op binnen SKOS beschikbare bouwstenen.

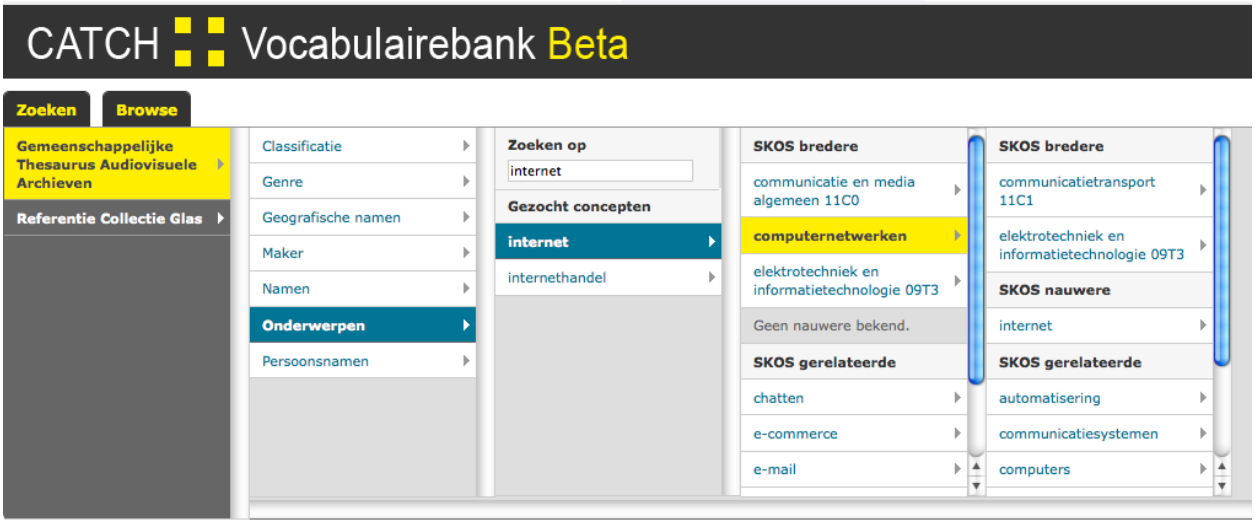


Figuur 1 Voorbeeld representatie relaties in SKOS

De huidige repository is gerealiseerd op basis van een zogenaamde RDF Store. Een eerste versie van de REST web service is binnen CATCHPlus ontwikkeld en online beschikbaar. De broncode is onder de GPL 2.0 licentie beschikbaar.

De terminologiebronnen in de repository zijn op twee manieren over het web toegankelijk:

1. *Via een REST web service* – Als een terminologiebron eenmaal in SKOS-vorm beschikbaar is, kan deze eenvoudig in de CATCHPlus vocabulaire repository worden geïmporteerd. Deze maakt het vervolgens mogelijk die bron op een aantal manieren te publiceren op het web. Een zgn. REST webservice API maakt het mogelijk de data op te halen en te doorzoeken volgens standaard webmethoden. Deze publicatiemethode stelt gebruikers in staat om gericht te zoeken naar concepten en relaties die aan bepaalde zoekcriteria voldoen, en om de zoekresultaten precies vorm te geven (bijvoorbeeld als gesorteerde termenlijst). Dit wordt voornamelijk gedaan vanuit software omgevingen; de API wordt primair gebruikt door programmeurs. Een voorbeeld van een dergelijke software-omgeving is de in opdracht van CATCHPlus ontwikkelde browse- en zoektool (figuur 2) Deze dient om de inhoud van de repository ook voor eindgebruikers toegankelijk te maken.
2. Als *Linked Open Data*. Linked Open Data is een snelgroeiend 'web' van naar elkaar verwijzende data-collecties op het web. (Hildebrand 2010) Deze verwijzingen hebben, net als voor SKOS concepten, de vorm van URIs, die in dit geval verwijzen naar ofwel een webpagina ofwel naar een afgebakende dataset, afhankelijk van hoe die verwijzing gevolgd wordt. Het is de bedoeling dat zowel de webpagina's als de datasets ook daadwerkelijk via het web zijn op te vragen. Door de URI van een bepaald SKOS concept in de repository te volgen, krijgt de gebruiker naar keuze de beschikking over een webpagina met informatie over dat concept, of die informatie zelf, als data voor gebruik in eigen applicaties. De dataset die hoort bij een SKOS concept omvat onder meer alle termlabels, verwijzingen naar andere concepten en de thesaurus waar het concept deel van uitmaakt.



Figuur 2. Vocabulaire Browse- en Zoektool (deel van de zoekinterface)

Casus: GTAA van Beeld en Geluid

Een goede casus van het gebruik van online terminologiebronnen is de samenwerking tussen Beeld en Geluid en het Nationaal Archief rondom de Gemeenschappelijke Thesaurus Audiovisuele Archieven, de GTAA. In de GTAA, in beheer bij Beeld en Geluid, zijn termen verdeeld over zes lijsten ofwel 'assen': Onderwerp, Genre, Maker, Persoonsnamen, Geografische Namen en Namen (overige eigennamen). In totaal omvat de GTAA 130.000 termen.

Beeld en Geluid heeft eerder dit jaar de stap genomen deze thesaurus beschikbaar te stellen onder de zgn. Open Database License; een wereldwijd gebruikte licentie waarmee onder bepaalde (beperkte) voorwaarden toestemming wordt verleend om een deel of de gehele databank te

gebruiken[2]. De GTAA is met behulp van de REST service van de Vocabulary and Alignment Repository verbonden met Memorix, het collectiebeheersysteem van het Nationaal Archief.

Het Nationaal Archief heeft voor het beschrijven van zijn fotocollectie gekozen voor de GTAA van Beeld en Geluid vanwege de brede inzetbaarheid en de mogelijkheid te interfacen met bestaande systemen. In de toekomst zullen ook andere instellingen voor hun audiovisuele collecties gebruik maken van deze op het web gepubliceerde versie van de GTAA. In samenwerking met de Vrije Universiteit en de TU Delft wordt gewerkt aan het creëren van links tussen de GTAA en andere bronnen, zoals Wordnet (semantisch lexicon voor de Engelse taal) en DBpedia (informatie uit Wikipedia). Hierdoor wordt het mogelijk om links te maken tussen uiteenlopende bronnen die onderdeel maken van de Linked Open Data principes..

Neem bijvoorbeeld een uitzending van Van Gewest tot Gewest over het dorp Nuenen. Op basis van het trefwoord Nuenen kan de bijbehorende Wikipedia pagina gekoppeld worden aan de uitzending. Hier staat weer meer informatie over personen die een relatie hebben met deze plaats, zoals Vincent van Gogh. Een link tussen DBpedia en de ULAN (met onder meer biografieën van kunstenaars) maakt inzichtelijk dat van Gogh van grote invloed was op schilders die deel uitmaken van het expressionisme. Door kennisbronnen semantisch aan elkaar te verbinden, kunnen verwijzingen gepresenteerd worden tussen de uitzending, en gerelateerde informatie over bijvoorbeeld het werk van Edvard Munch en Francis Bacon. Dit is een eenvoudig voorbeeld dat duidelijk maakt hoe externe kennis, gerepresenteerd volgens een gestandaardiseerd model, het doorzoeken van erfgoedcollecties verrijkt.[3]

Conclusie

Door de exponentiële groei van online erfgoed groeit ook de vraag naar gereedschap die het mogelijk maken bronnen te doorzoeken, te visualiseren en patronen te analyseren. Ook is het belangrijk dat gebruikers of groepen gebruikers eenvoudig context toe kunnen voegen en content kunnen hergebruiken.

Het online publiceren en gebruiken van gemeenschappelijke terminologiebronnen is een belangrijke bouwsteen om de vindbaarheid van objecten uit uiteenlopende collecties te verbeteren en te linken aan externe bronnen. De CATCHPlus Vocabulary Repository wordt al in verschillende gebruikersscenario's ingezet en biedt kansen voor de toekomst; voor het beschrijven van collecties en het bewerkstellingen van interoperabiliteit tussen collecties. Het spannende aan het publiceren volgens de richtlijnen van Linked Open Data is dat toegang tot de data extreem laagdrempelig wordt. (Clarc 2010) Het stelt externe partijen in staat innovatieve 'mashups' te creëren, die bestaande content in een nieuwe context presenteert. Hiermee wordt een belangrijke stap gezet van toegang tot interpretatie van grote gedistribueerde collecties erfgoed en draagt bij aan het succes van de Cultural Commonwealth.

Literatuur

Clark, Kendall. Another Reason Semantic Web Kicks Ass. Mei 2010. Online beschikbaar: http://clarkparsia.com/weblog/2010/05/26/another-reason-semantic-web-kicks-ass/

Gradmann, Stefan. Europeana White Paper 1: Knowledge = Information in Context. Europeana, April 2010. Online beschikbaar: http://version1.europeana.eu/web/europeana-project/whitepapers

Hildebrand, M., van Ossenbruggen, J. R., Hardman, L., Wielemaker, J., and Schreiber, G.. Searching In Semantically Rich Linked Data: A Case Study In Cultural Heritage. 2010. Online beschikbaar: http://oai.cwi.nl/oai/asset/15324/15324D.pdf

[2] http://www.opendatacommons.org/licenses/odbl/summary/

[3] Het Europeana Thought lab ontwikkeld door de Vrije Universiteit (http://www.europeana.eu/portal/thought-lab.html) demonstreert de kracht van Linked Open Data op een krachitige manier.

Our Cultural Commonwealth: The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Siences. American Council of Learned Societies, 2006

Miles, Alistar and Sean Bechhofer: SKOS Simple Knowledge Organization System, 2009. Online beschikbaar: http://www.w3.org/TR/2009/REC-skos-reference-20090818/

Over de auteurs

Drs. Johan Oomen is hoofd R&D bij het Nederlands Instituut voor Beeld en Geluid en onderzoeker bij de faculteit Computer Science van de Vrije Universiteit Amsterdam

Drs. Hennie Brugman is werkzaam bij het Max Planck Institute for Psycholinguistics en is de technisch coördinator van CATCHplus.