Gemeenschappelijke Thesaurus voor Uniforme Ontsluiting

-

Eindrapport

Matthias Priem (VIAA)
Victor de Boer (Beeld en Geluid)
Michiel Hildebrand (Spinque)
Johan Oomen (Beeld en Geluid)
Nico Verplancke (VIAA)

17 juni 2016

Inhoud

Korte samenvatting

Dit project werd aangevraagd door Beeld en Geluid en VIAA. Beide instelling beheren een groot digitaal archief dat uit diverse bronnen samengesteld is. Het materiaal is afkomstig uit de publieke omroep(en), regionale zenders en/of culturele erfgoedinstellingen. Dit archiefmateriaal wordt ontsloten naar diverse doelgroepen, zoals de klanten zelf, research of onderwijs.

Een thesaurus is één van de instrumenten die materiaal beter en uniformer doorzoekbaar kunnen maken. Wanneer lokale thesauri bovendien gelinkt worden met elkaar ontstaat de opportuniteit om op uniforme wijze door verschillende collecties te zoeken. Links tussen thesauri bieden mogelijkheden op lokaal niveau, maar evenzeer op internationaal niveau.

Dit pilootproject is een voorbeeld van zo`n samenwerking op internationaal niveau. Er werd onderzocht hoe de verschillende thesauri zich verhouden ten opzicht van elkaar en hoe ze *gemapped* kunnen worden. Een demonstrator illustreert op welke manier de collecties dan doorzoekbaar zijn. Naast de projectresultaten (code, kennisopbouw) wil dit project ook een katalysator zijn voor verdere Vlaams-Nederlandse samenwerking op dit vlak.

Project inhoud

Context en aanleiding

Archieven die collecties herbergen vanuit diverse bronnen gebruiken onder meer thesauri om de collecties uniform doorzoekbaar te maken. Thesauri zijn gecontroleerde vocabulaires van termen (of concepten) die enkele specifieke problemen kunnen oplossen:

- Desambigueren: door het toevoegen van context, of door de plaats in een hiërarchie kan een bepaalde term onderscheiden worden van een gelijknamige term. Denk bijvoorbeeld aan Kaas (het boek) of kaas (het voedingsmiddel).
- Relaties en hiërarchie toevoegen: door middel van hiërarchie of relaties kunnen termen binnen een thesaurus aan elkaar gelinkt worden (bv. Gent is een deel van Vlaanderen dat een deel is van België, etc.)
- Synoniemen: thesauri hebben typisch functionaliteit voor synoniemen, waardoor andere spellingen van een zelfde term toch als een identieke zoekterm kunnen dienen. Zo heeft de VIAF[1] tientallen varianten voor Rembrandt Van Rijn[2].
- Vertalingen: thesauri kunnen meertalig zijn waardoor termen en begrippen éénduidig in een andere taal geïdentificeerd kunnen worden.

VIAA heeft in 2014 een studie uitgevoerd die zoekt naar de haalbaarheid van een uniforme thesaurus voor de media-en erfgoedsector in Vlaanderen[3]. Daaruit bleek[4] dat de creatie van

[1] Een internationale, publieke thesaurus van personen
[2] http://viaf.org/viaf/64013650/#Rembrandt,_Harmenszoon_van_Rijn,_1606-1669
[3] Zie http://viaa.be/nieuws/de-resultaten-van-de-haalbaarheidsstudie-unified-thesaurus voor meer informatie

één ‘überthesaurus’ voor alle termen en voor verschillende sectoren geen optie is. Om uniforme doorzoekbaarheid toch te garanderen werd hier al voorgesteld om bestaande thesauri te linken aan elkaar. Links hebben het voordeel dat ze een beperkte impact hebben op de bestaande thesauri (inhoudelijk hoeft er niets gewijzigd te worden) terwijl ze toch termen relateren aan elkaar.

Dit project onderzoekt verder de piste van het linken van thesauri. Het spitst zich toe op het Nederlands en onderzoekt de mogelijk om over de landsgrenzen heen een gelinkte thesaurus te bouwen. Er wordt gewerkt met SKOS, een door het W3C gestandaardiseerde format voor thesauri.

Project opzet

Het project werd onderverdeeld in drie inhoudelijke werkpakketten, die verderop in dit document uitgebreid gedocumenteerd worden:

1. Selectie van de bronnen (collecties en thesauri), omzetting van de bronnen naar SKOS waar nodig.
2. Alignering van de thesauri
3. Bouw van een demonstrator die het praktisch gebruik van een gelinkte thesaurus illustreert.

Project partners

Beeld en Geluid

Beeld en Geluid is een van de grootste audiovisuele archieven van Europa; het unieke gebouw op het Media Park in Hilversum herbergt ruim 800.000 uur aan radio, televisie, film en muziek. Beeld en Geluid maakt zijn collectie toegankelijk voor uiteenlopende doelgroepen, waaronder mediaprofessionals, de creatieve sector, het onderwijs en het algemeen publiek. Door middel van onderzoek en innovatie heeft het instituut zich ontwikkeld tot een brede culturele instelling die door zijn opgebouwde kennis en infrastructuur een centrale functie inneemt binnen de archief-en mediasector.

Beeld en Geluid maakt haar collectie online beschikbaar door middel van diverse eindgebruikers services, met inbegrip van speciale diensten voor de creatieve industrie, onderwijs en onderzoek. Het instituut is ook een toeristische attractie voor het grote publiek en wordt bezocht door meer dan 250.000 mensen per jaar.

[4] Het eindrapport met conclusies vindt u hier: http://viaa.be/assets/files/page/downloads/Unified_Thesaurus_-_possibilities__representation_and_design.pdf

VIAA

VIAA is het Vlaams instituut voor archivering en werd opgericht in december 2012. VIAA digitaliseert, archiveert en ontsluit materiaal van meer dan 80 organisaties. Onder deze organisaties bevinden zich de Vlaamse publieke omroep VRT, regionale omroepen, archiefinstellingen en culturele erfgoedinstellingen.

De focus van het VIAA situeert zich op dit moment op audiovisueel erfgoed. Daarnaast werden in de loop van 2015 ook 55.000 W.O.-I kranten gedigitaliseerd en gearchiveerd, goed voor 270.000 pagina's in totaal. In de toekomst zal VIAA minimum 500.000 uur aan cultureel erfgoed digitaliseren en archiveren.

Het gedigitaliseerde en gearchiveerde materiaal kan vervolgens ter beschikking gesteld worden van onderwijs, research en het grote publiek (via de publieke bibliotheken). VIAA heeft op deze manier het platform Het Archief gerealiseerd, waar de WO-I kranten beschikbaar zijn. Op het platform 'Het Archief voor Onderwijs' wordt materiaal specifiek voor leerkrachten ter beschikking gesteld. Op dit moment (zes maanden na de lancering) bereikt VIAA hiermee meer dan 50% van alle leerplichtonderwijs in Vlaanderen.

Andere betrokken partijen

VRT

VIAA fungeert als een *dienstenleverancier* op het vlak van zowel digitaliseren als archiveren. VIAA staat in voor de duurzame bewaring van tal van digitale archieven, maar beschikt niet over een eigen collectie. Om dit project uit te voeren werd een partnership aangegaan met VRT. VRT is de publieke omroep in Vlaanderen. Ze beschikt over een archief van meer dan 1 miljoen items aan audiovisueel materiaal.

De items bij VRT zijn door VRT archivarissen geannoteerd, onder meer op basis van een intern ontwikkelde thesaurus. Daarnaast is er een historiek op het vlak van thesauri: VIAA en VRT hebben in 2014 samen gewerkt aan een haalbaarheidsstudie rond thesauri, waar nagegaan werd of een gemeenschappelijke thesaurus binnen Vlaanderen kan opgezet worden. Voor dit project werd gebruik gemaakt van een deel van het VRT archief en de volledige VRT thesaurus.

Data Science Lab Ugent

Het Data Science Lab van de Universiteit Gent heeft -onder meer- een ruime expertise op het vlak van semantische en linked open data. Dit lab heeft in opdracht van VIAA de mapping van de bestaande VRT thesaurus naar een versie in SKOS verzorgd.

Spinque

Spinque is gebaseerd in Utrecht, NL en is een spin-off van het Centrum Wiskunde en Informatica (CWI). Spinque ontwikkelt zoektechnologie die ingezet wordt in een aantal

vakgebieden, waaronder linked data en het semantische web. Spinque heeft in opdracht van VIAA en Beeld en Geluid gezorgd voor de implementatie van de demonstrator.

Web en Media groep - Vrije Universiteit Amsterdam

De Vrije Universiteit is een algemene universiteit in Amsterdam met ongeveer 23.000 studenten. De Web en Media (W&M) groep, onderdeel van de afdeling Informatica, houdt zich bezig met onderzoek rond *(Semantic) Web* technologieen, *human computation, digital humanities* en *media.* Binnen dit project heeft de groep een consulterende rol vervuld en en leveren zij beperkte faciliteiten zoals server-ruimte.

Rapportage per werkpakket

Selectie en van de bronnen

VRT Thesaurus, omgezet in SKOS

Als test collectie bij VIAA werden de VRT thesaurus en een deel van het VRT archief geselecteerd. Het VRT archief is het grootste audiovisueel archief binnen Vlaanderen ; VRT is dan ook één van de grootste klanten binnen VIAA. Sinds de invoer van een digitale metadata systeem (1986) werd reeds gewerkt aan een centrale trefwoordenlijst. De lijst is sindsdien gemigreerd, vaak tijdens een MAM upgrade. Recent nog (in 2014) werd de thesaurus sterk afgeslankt en geïmporteerd in “Het Depot”, VRT’s MAM systeem.

De thesaurus bestaat op dit moment uit 102.173 termen. Deze zijn opgeslagen in het VRT MAM systeem om zo gebruikt te worden bij de annotatie van de archief items. De thesaurus werd geëxporteerd in XML format, wat ons een eerste zicht leverde op de structuur van de thesaurus:

```
<?xml version="1.0" encoding="utf-16"?>
<Thesauri xmlns:xsd="http://www.w3.org/2001/XMLSchema"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" Version="2"
xmlns="http://www.blue-order.com/Schema/Thesaurus">
<Thesaurus Id="VRT_THESAURUS_KEYWORDS" LastChanged="2015-04-24T11:13:11"
FixForPlatform="true">
        <Terms>
                <Term id="1" ParentId="125479">
                        <Owners>
                                <Owner>DEFAULT</Owner>
                        </Owners>
                        <Localizations>
                                <Localization Culture="nl-BE" Tenant="DEFAULT">
                                        <Label>'S GRAVENVOEREN</Label>
                                </Localization>
                                <Localization Culture="en" Tenant="DEFAULT">
                                        <Label>'S GRAVENVOEREN</Label>
                                </Localization>
                        </Localizations>
                </Term>
                <Term id="2" ParentId="98979">
                        <Owners>
                                <Owner>DEFAULT</Owner>
                        </Owners>
```

```
                <Localizations>
                        <Localization Culture="nl-BE" Tenant="DEFAULT">
                                <Label>'S GRAVENWEZEL</Label>
                        </Localization>
                        <Localization Culture="en" Tenant="DEFAULT">
                                <Label>'S GRAVENWEZEL</Label>
                        </Localization>
                </Localizations>
        </Term>
```

De thesaurus maakt geen gebruik van een inhoudelijke indeling. Het ‘wat’, ‘waar’ en ‘wie’ kan je dus allemaal in één en dezelfde lijst terugvinden. Er wordt wel uitgebreid gebruik gemaakt van een aantal relaties tussen termen en van synoniemen of andere labels voor een bestaande term. Elke term heeft bovendien een unieke ID meegekregen. Enkele relaties die in de originele file bestonden werden vertaald naar SKOS gebaseerde relaties, we lijsten deze kort op aan de hand van voorbeelden.

In een heel aantal gevallen heeft een term een ‘parentID’. Hieronder als voorbeeld de term “Amsterdam”, in het originele formaat:

```
<Term id="3661" ParentId="80076">
        <Owners>
                <Owner>DEFAULT</Owner>
        </Owners>
        <Localizations>
                <Localization Culture="nl-BE" Tenant="DEFAULT">
                        <Label>AMSTERDAM</Label>
                </Localization>
                <Localization Culture="en" Tenant="DEFAULT">
                        <Label>AMSTERDAM</Label>
                </Localization>
        </Localizations>
</Term>
```

Deze term heeft een ParentID, die wijst naar volgende record (alweer in origineel formaat):

```
<Term id="80076" ParentId="38300">
        <Owners>
                <Owner>DEFAULT</Owner>
        </Owners>
        <Localizations>
                <Localization Culture="nl-BE" Tenant="DEFAULT">
                        <Label>NEDERLAND</Label>
                </Localization>
                <Localization Culture="en" Tenant="DEFAULT">
                        <Label>NEDERLAND</Label>
                </Localization>
        </Localizations>
</Term>
```

Die op zich weer wijst naar een andere term. Op deze manier komen we tot een hiërarchie als volgt: Amsterdam => Nederland => Europese Landen => Europa => Werelddeel (een term zonder parentID, en dus een zgn.. TopConcept). Deze termen werden aan de hand van de SKOS relaties “broader” en “narrower” aan elkaar gelinkt:

In totaal zijn er van alle termen in de VRT thesaurus een relatief klein aantal (4.429 termen) die de top van de hiërarchie uitmaken.

De hiërarchie leent zich potentieel wel tot een verdere opdeling van de thesaurus. Immers, een groot aantal termen binnen 1 topconcept (bvb. Werelddeel) valt duidelijk onder termen die geografisch van aard zijn, terwijl andere topconcepten uitsluitend personen of titels bevatten. Dit zou echter een inhoudelijke wijziging van de thesaurus betekenen en valt buiten de scope van dit project.

Verder duiken een aantal niet-hiërarchische relaties op in de VRT thesaurus:

```
<Relation Term1="2005" Term2="3661" Tenant="DEFAULT"/>
   …
<Relation Term1="3661" Term2="2005" Tenant="DEFAULT"/>
```

In dit geval gaat het om de term “Amsterdam” (3661) die gerelateerd wordt aan “AJAX” (2005). Deze relaties werden vertaald op basis van skos:related in het SKOS equivalent van de VRT thesaurus.

Op die manier ziet de oorspronkelijke entry er voor Amsterdam als volgt uit na vertaling in SKOS:

```
<http://example.com/concept/3661> a skos:Concept ;
    skos:broader <http://example.com/concept/80076> ;
    skos:inScheme <http://example.com/thesaurus/VRT> ;
    skos:narrower <http://example.com/concept/102932>,
        <http://example.com/concept/10710>,
        <http://example.com/concept/11154>,
        <http://example.com/concept/53802>,
        <http://example.com/concept/57298>,
        <http://example.com/concept/65904>,
        <http://example.com/concept/69371>,
        <http://example.com/concept/78779>,
        <http://example.com/concept/81025>,
        <http://example.com/concept/84282>,
        <http://example.com/concept/85169>,
        <http://example.com/concept/92068> ;
    skos:prefLabel "AMSTERDAM"@en,
        "AMSTERDAM"@nl-be ;
    skos:related <http://example.com/concept/2005> .
```

Er werd een fictief domein naam toegevoegd (example.com) en de relaties werden aan de term toegevoegd. De narrower relaties zijn in dit geval termen die Amsterdam als ParentID hadden en zijn bijna allemaal straatnamen of pleinen in amsterdam.

Verder werden een beperkt (212) aantal verklaringen gevonden in de VRT Thesaurus. Deze werden vertaald naar scopeNotes in SKOS. Enkele voorbeelden:

Subject	Object
vrt:1492	"FILMTITEL"@en
vrt:1492	"FILMTITEL"@nl-be
vrt:AANKOMST	"AANKOMST VAN REIZIGERS, NA OPDRACHT IH BUITENLAND, ETC"@en
vrt:AANKOMST	"AANKOMST VAN REIZIGERS, NA OPDRACHT IH BUITENLAND, ETC"@nl-be
vrt:ACHTERSTAND	"OP STUDIEVLAK: LEERMOEILIJKHEID"@en
vrt:ACHTERSTAND	"OP STUDIEVLAK: LEERMOEILIJKHEID"@nl-be
vrt:ALI	"FILMTITEL !!!!"@en
vrt:ALI	"FILMTITEL !!!!"@nl-be
vrt:AMERICA	"POPGROEP !!!"@en
vrt:AMERICA	"POPGROEP !!!"@nl-be
vrt:ARBITRAGEHOF	"VANAF MEI 2007 IN BELGIE : GEBRUIK GRONDWETTELIJK HOF"@en
vrt:ARBITRAGEHOF	"VANAF MEI 2007 IN BELGIE : GEBRUIK GRONDWETTELIJK HOF"@nl-be
vrt:AZ SINT-JAN	"HEDENDAAGS ZIEKENHUIS"@en
vrt:AZ SINT-JAN	"HEDENDAAGS ZIEKENHUIS"@nl-be
vrt:BELPOP	"TITEL TV-PROGRAMMA"@en
vrt:BELPOP	"TITEL TV-PROGRAMMA"@nl-be
vrt:BILLY ELLIOT	"FILM / MUSICAL"@en
vrt:BILLY ELLIOT	"FILM / MUSICAL"@nl-be
vrt:BLOEMENHULDE	"HET LEGGEN VAN BLOEMEN OP EEN GRAF OF MONUMENT"@en
vrt:BLOEMENHULDE	"HET LEGGEN VAN BLOEMEN OP EEN GRAF OF MONUMENT"@nl-be

Tot slot werd ook gewerkt met de SKOS termen prefLabel en AltLabel, om geprefereerde schrijfwijzes te onderscheiden van alternatieven. Bijvoorbeeld:

```
<Term id="620" ParentId="57302">
	<Owners>
		<Owner>DEFAULT</Owner>
	</Owners>
	<Localizations>
		<Localization Culture="nl-BE" Tenant="DEFAULT">
			<Label>ABDULLAH (JORDAANSE DYNASTIE)</Label>
			<Synonyms>
				<Synonym>ABDALLAH</Synonym>
				<Synonym>ABDOELLAH</Synonym>
				<Synonym>ABDULLAH</Synonym>
				<Synonym>ABDULLAH BIN TALAL AL HASHEMI</Synonym>
			</Synonyms>
		</Localization>
	</Localizations>
</Term>
```

In SKOS vertaald wordt dit:

```
<http://example.com/concept/620> a skos:Concept ;
    skos:altLabel "ABDALLAH"@en,
        "ABDOELLAH"@en,
        "ABDULLAH"@en,
        "ABDULLAH BIN TALAL AL HASHEMI"@en,
        "ABDALLAH"@nl-be,
        "ABDOELLAH"@nl-be,
        "ABDULLAH"@nl-be,
        "ABDALLAH"@nl-be,
        "ABDOELLAH"@nl-be,
        "ABDULLAH"@nl-be,
        "ABDULLAH BIN TALAL AL HASHEMI"@nl-be ;
    skos:broader <http://example.com/concept/57302> ;
    skos:inScheme <http://example.com/thesaurus/VRT> ;
    skos:prefLabel "ABDULLAH (JORDAANSE DYNASTIE)"@en,
        "ABDULLAH (JORDAANSE DYNASTIE)"@nl-be .
```

In totaal werden op deze manier de termen uit de VRT thesaurus vertaald naar SKOS voor verder gebruik. We geven nog kort enkele sleutel getallen mee:

- 102172 termen in totaal
- 97744 van deze termen hebben een *broader* of *narrower* relatie
- 4429 zijn topconcepten
- Er zijn 212 scopeNotes en 6828 termen die gerelateerd zijn aan elkaar.

In de toekomst kan dit werk verder gebruikt worden om de VRT thesaurus permanent als SKOS te publiceren en bijvoorbeeld te gebruiken als onderliggend annotatiemechanisme voor het VRT en VIAA archief.

De code van dit werkpakket, die de eigenlijke omzetting verzorgt, is herbruikbaar en wordt binnenkort ter beschikking gesteld als open source via deze URL: https://github.com/viaacode/skoscreator.

GTAA

Het Nederlands Instituut voor Beeld en Geluid heeft samen met enkele andere Nederlandse organisaties die audiovisueel cultureel erfgoed beheren de Gemeenschappelijke Thesaurus voor Audiovisuele Archieven (GTAA) ontwikkeld (http://gtaa.beeldengeluid.nl). De GTAA wordt gebruikt voor het doeltreffend karakteriseren van de inhoud van audiovisueel materiaal uit het archief met labels afkomstig uit een gecontroleerde en gestructureerde lijst van termen, een thesaurus. De thesaurus wordt bij Beeld en Geluid voornamelijk ingezet in het handmatige beschrijvingsproces, maar ook meer en meer bij automatische annotatietechnieken. Met behulp van linked data principes kunnen daarnaast verbanden worden gelegd tussen de eigen collectie en andere databronnen. De GTAA is beschikbaar als SKOS en heeft rond de 180.000 termen in verschillende assen: Onderwerpen, Persoonsnamen, Geografische Namen en Genres.

GTAA maakt verder ook gebruik van een aantal skos labels en relaties:

- 184.484 termen
- 19.695 van deze termen hebben een broader of narrower relatie
- Er zijn 9 conceptSchemes, die de thesaurus inhoudelijk verdelen in geografische termen, personen, Genres, etc.
- Er zijn 90.708 scopeNotes en 33.542 termen die gerelateerd worden aan elkaar.

Mapping tussen beide thesauri

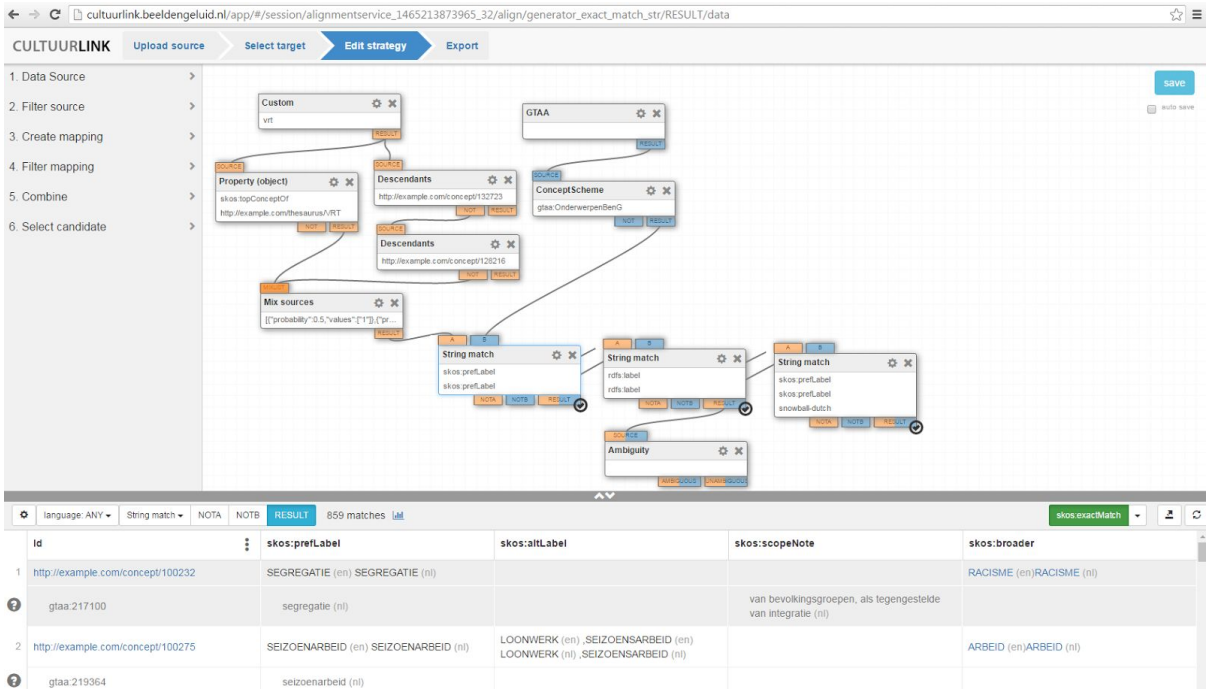
Voor het bouwen van de mapping strategie wordt gebruik gemaakt van de functionaliteiten van CultuurLINK (http://cultuurlink.beeldengeluid.nl). CultuurLINK is een '*alignment tool*' waarin strategieën kunnen worden geconstrueerd om vocabulaires aan elkaar te koppelen. Door het combineren van verschillende filters en term-matching technieken kan een hoog aantal identieke termen in twee vocabulaires geïdentificeerd worden.

Er zijn meerdere strategieën geconstrueerd voor de alignment van de VRT en GTAA thesauri die overeenkomen met verschillende typen termen. De reden voor deze opdeling is

dat verschillende typen termen verschillende eigenschappen hebben. Zo bestaan persoonsnamen uit minimaal twee delen (bv. “Achternaam, Voornaam”) en hebben onderwerpen enkelvouds- en meervoudsvormen terwijl dat bij locaties niet het geval is. Voor een optimale matching bleek een verdeling in vier strategieën optimaal:

1. vrt_onderwerpen
2. vrt_locaties
3. vrt_namen
4. vrt_personen

Als voorbeeld is hier de strategie voor vrt_onderwerpen te zien.



Het screenshot laat de gehele strategie visueel zien, waarbij verschillende filters op de originele vocabulaires worden gebruikt om de onderwerpen te selecteren. Vervolgens worden string-matchers en andere bouwblokken gebruikt om de identieke termen te identificeren. In de onderste helft van het scherm is te zien welke termen op elkaar afgebeeld zijn, ter inspectie voor de gebruiker.

De eerste drie strategieën zijn te bekijken op http://cultuurlink.beeldengeluid.nl door de naam van de strategie als sessienaam op te geven. De vrt_personen strategie is niet recent op de publieke versie van CultuurLINK beschikbaar. Een match blok is toegevoegd waarin *reguliere expressies* kunnen worden gebruikt. Dit stelt ons in staat om GTAA namen (Achternaam, Voornaam) op VRT namen (“Achternaam Voornaam”, dus zonder comma) te mappen. Hiermee vonden we i.p.v. 400 links ruim 11.000 links tussen personen.

De geëxporteerde links zijn te vinden op https://github.com/biktorrr/gtou_taalunie. Ze zijn live te bekijken op een publiek toegankelijke triple store: http://semanticweb.cs.vu.nl/test/. Op

deze ClioPatria server[5] zijn de individuele alignments geladen als named graphs: http://semanticweb.cs.vu.nl/test/browse/list_graphs. Via deze server zijn de alignments ook met het SPARQL protocol bevraagbaar: http://semanticweb.cs.vu.nl/test/sparql/.
De tabel hieronder laat zien hoeveel links er uiteindelijk gevonden zijn tussen de GTAA en VRT thesauri. In totaal zijn 21.640 links gevonden.

Type term	**Aantal links**
Onderwerpen	4.167
Namen	2.197
Lokaties	4.011
Personen	11.265
Totaal	**21.640**

In totaal is 21% van de VRT termen gemapt op een GTAA term. Dit lijkt een laag percentage, maar hierbij moet in acht genomen worden dat de VRT thesaurus en VRT daadwerkelijk ook maar ten dele overeenkomende termen hebben. Zo bestaat GTAA voor een zeer groot deel uit persoonsnamen en namen van programmamakers die alleen in de Nederlandse context relevant zijn. Datzelfde geld voor Vlaamse personen.

Voor onderwerpen is het zelfs zo dat het aantal links bijna gelijk is aan het aantal linkbare onderwerpen. In GTAA zijn 4.683 onderwerpen (geidentificeerd door het conceptScheme: “onderwerpenBenG”) en in de VRT thesaurus 25.155 potentiele onderwerpen (geidentificeerd via uitsluiting van personen en geografische concepten). Hiervan worden 4.167 mappings gevonden, wat overeenkomt met 89% van de GTAA termen.
Voor lokaties worden 4.011 links gevonden. In de VRT thesaurus zijn 8.617 lokatietermen aanwezig, wat overeenkomt met een percentage matches van 47%. Een snelle analyse van de niet-gematchte lokatietermen laat zien dat dit kleinere plaatsen betreft, die blijkbaar wel in de ene thesaurus en niet in de andere thesaurus voorkomen. Een verdere analyse van de niet-gematchte termen zou verder inzicht geven in de exacte kwaliteit van de alignment en in hoeverre deze overeenkomt met de verwachte kwaliteit.

Gebruik van de links

De links tussen de concepten uit de GTAA en VRT thesauri maken het mogelijk om in de strategie resultaten te vinden over de twee collecties. Een voorbeeld is dat de gebruiker zoekt met een concept uit de GTAA thesaurus en hiermee ook videos uit de VRT collectie worden gevonden. De links tussen de concepten maken ook de cross-collectie suggesties van videos mogelijk. Een zoekvraag met een video uit de openbeelden collectie geeft daarom ook resultaten uit de VRT collectie.

[5] http://cliopatria.swi-prolog.org

Gebruik van de links - demonstrator

Demonstrator

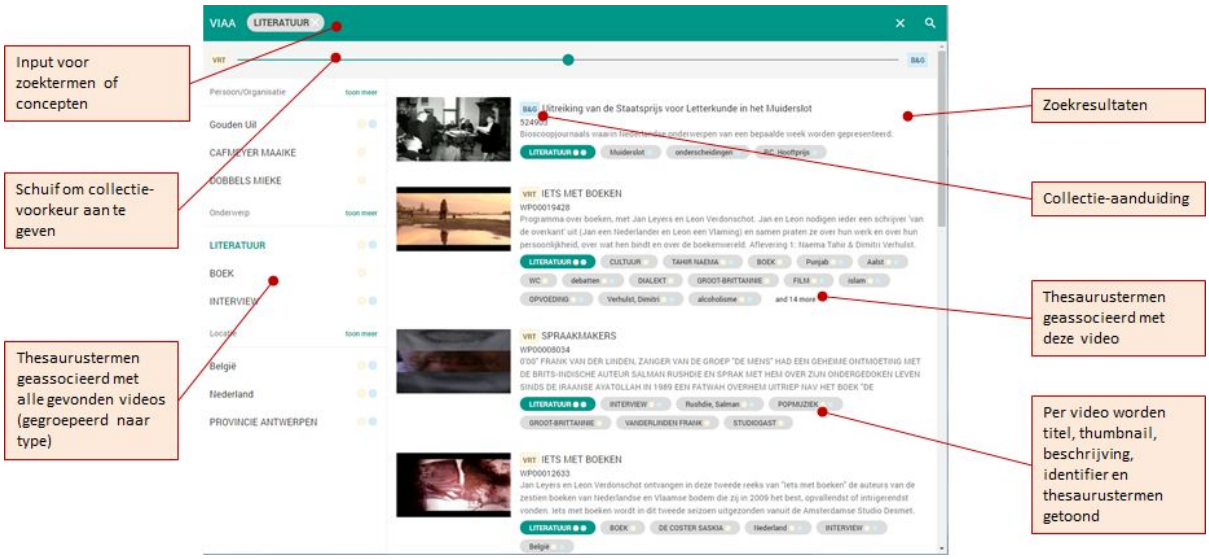
Bronmateriaal
Op basis van de links kan nu gezocht worden in beide collecties. Hiervoor werd een demonstrator gebouwd die werkt op basis van de volgende inputs:

- Voor VIAA
    - Een deel van de VRT beeldcollectie. In de huidige versie van de demonstrator zijn ongeveer 35.000 items opgenomen wat een kleine subset is van een archief dat meer dan 1 millioen records bevat.
    - De annotatie van die items, op basis van de VRT thesaurus, maar uitgebreid met beschrijvingen en datums van uitzending.
- Voor Beeld en Geluid
    - De collectie openbeelden: een set van 1700 video items die vrij beschikbaar zijn op het Web. Ook hier is dit maar een fractie van het volledige NIBG archief.
    - De annotatie van de items, onder meer op basis van de GTAA.

Gebruikersinterface
De demonstrator is te bereiken op http://link.spinque.com/VIAA-1.0/ . Aangezien een deel van de collectie niet publiek toegankelijk gemaakt mag worden is een wachtwoord nodig. Om de functionaliteit van de demonstrator zichtbaar te maken is ook een *screencast* van de demonstrator publiek beschikbaar gemaakt op youtube: https://youtu.be/iOJvcHRfvDY

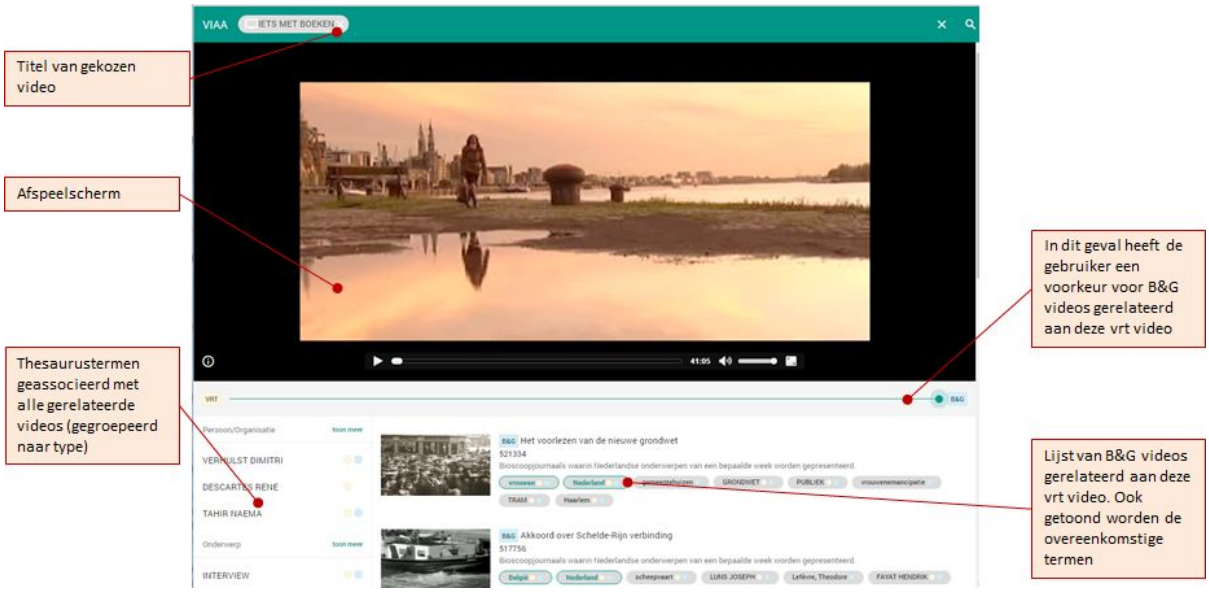
Als ingelogd is kan een gebruiker beginnen met zoeken, gebruikmakend van een zoekterm. De interface toont matchende resultaten op basis van titels en beschrijvingen.
Ook de concepten waarmee de videos geannoteerd zijn worden getoond. Voor deze concepten is ook te zien of ze in een enkele, of in beide thesauri voorkomen (geel / blauw stipje). De concepten zijn te selecteren waarna ze als zoekterm worden gebruikt. Het screenshot hieronder laat een voorbeeld zien waarbij het concept “Literatuur” gekozen is. Dit onderwerp is gelinkt in de thesauri en de interface laat ook verschillende videos uit beide collecties zien.

Bovenin het scherm is ook een ‘*slider*’ bedienbaar, die de gebruiker in staat stelt om meer of minder gewicht te geven aan een van de twee collecties. Ook op die manier kan de gebruiker de resultaten sturen.



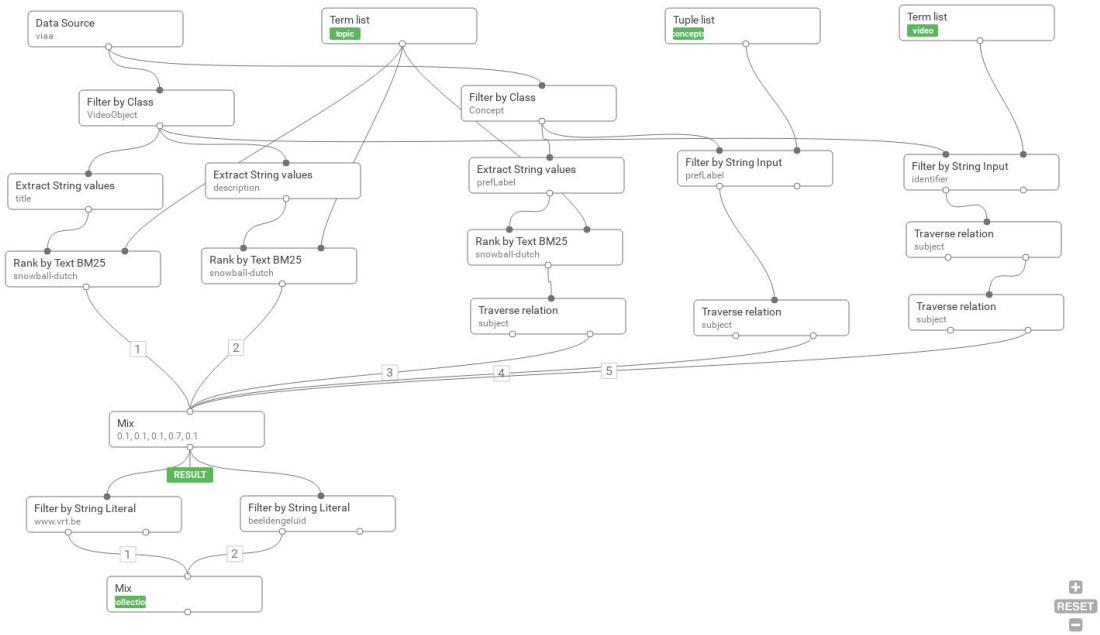
Het geannoteerde screenshot hierboven laat de verschillende eigenschappen van de demonstrator zien.

Wanneer een gebruiker een video kiest zal deze getoond worden. Ook worden gerelateerde videos getoond. Deze worden bepaald aan de hand van de overeenkomstige thesaurus concepten. Een voorbeeld staat hieronder weergegeven.



Zoekstrategie

De applicatie haalt video's van de twee collecties op aan de hand van zoektermen of geselecteerde concepten uit de twee thesauri. Hieronder wordt de zoekstrategie beschreven die door de applicatie wordt gebruikt om relevante video's op te halen.



De afbeelding toont de strategie voor de zoek en suggestie functionaliteit voor de VIAA applicatie.

Data Source
De strategie begint links boven met het ‘Data Source’ blok. Hierin wordt de volledige database doorzoekbaar gemaakt. De database bevat de VRT collectie en thesaurus, de openbeelden collectie, de GTAA thesaurus en verschillende linksets die de VRT en GTAA thesauri met elkaar verbinden.

In de applicatie zoeken we alleen naar video's. In de data zijn zowel de video's uit de VRT collectie als de openbeelden collectie gemodelleerd als instanties van het type “http://schema.org/VideoObject”. In de strategie maken we hier gebruik van door de *Data source* te filteren op objecten van het type VideoObject. Dit is het blok direct onder het Data source blok, Filter by Class.

Inputs
In de applicatie kan de gebruiker een complexe zoekvraag samenstellen. De zoekvraag kan bestaan uit keywords, uit thesaurus concepten en een video. In de strategie zijn deze drie inputs gerepresenteerd door de blokken bovenaan, met een groen label.

Zoeken
De drie type inputs uit de zoekvraag worden op verschillende manieren gebruikt om videos te vinden. De keywords worden gebruikt om te zoeken in de titels en beschrijvingen van de videos. De vier blokken aan de linkerkant zijn hiervoor verantwoordelijk. Ook worden de keywords gebruikt om thesaurus concepten te vinden, die vervolgens via de subject relatie leiden tot relevante videos.

De concepten in een zoekvraag kunnen direct aan een video worden gerelateerd via de subject relaties.

Als een video onderdeel is van de zoekvraag dan kan de rol van de strategie gezien worden als *recommendation* (vindt video's gerelateerd aan deze video). Om dit te realiseren worden de thesaurus concepten gebruikt. Dit gebeurt met de blokken aan de rechterkant van de strategie. Eerst worden de concepten gevonden van de gegeven video en vervolgens worden video's gevonden die beschreven zijn met dezelfde concepten. De gerelateerde video's worden gewogen op het aantal concepten dat overeenkomt. Hoe meer concepten hetzelfde hoe relevanter.

Resultaten combineren
Uiteindelijk worden de sub-resultaten van de verschillende *takken* in de strategie bij elkaar gebracht door een *Mix* blok. Dit blok heeft 5 inputs. Elk input heeft een gewicht dat aangeeft hoeveel moet worden meegenomen. Zo zijn resultaten gevonden via concepten belangrijker dan de resultaten gevonden via keywords.

Collecties wegen
De resultaten bevatten videos van de VRT collectie als de openbeelden collectie. In de laatste drie blokken onderaan wordt de prioriteit van de collectie bepaald. In de zoekapplicatie kan de gebruiker aangeven (met een slider) welke collectie relevanter is. In het laatste *Mix* blok wordt deze input gebruikt om de gewichten te zetten. Als de slider in de applicatie in het midden staan tellen bij collecties evenredis mee, de gewichten zijn allebei 0.5. Als de gebruiker de slider naar de VRT collecties beweegt wordt dit gewicht hoger, bijjvoorbeeld 0.75 en het gewicht van de resultaten uit de openbeelden collectie wordt dan 0.25.

Disseminatieplan

Los van de resultaten uit dit project is de disseminatie ervan minstens even belangrijk: dit project heeft zich ook tot doel gesteld om te dienen als katalysator voor samenwerkingen binnen het taalgebied. Daarnaast heeft dit project ook een voorbeeldfunctie en moet het mensen inspireren om data te publiceren, linken en hergebruiken. De erg praktische insteek van dit project werkt hopelijk drempelverlagend voor andere partijen in het archief en erfgoedveld.

Tijdens de looptijd van het hele project werd het project werd voorgesteld op de thesaurus werkgroep vergaderingen. In de toekomst wordt het eindresultaat en de next steps eveneens naar deze groep meegedeeld. De thesaurus werkgroep bestaat uit een twintigtal Nederlandse en Vlaamse spelers uit de archief en erfgoedsectoren (RKD, DEN, NCDD, VIAA, Packed, MoMU, Provincie Oost-Vlaanderen, Limburg, Brabant, …) die actief zijn op het vlak van thesauri en thesaurus management. Onder meer dankzij dit project maakt VIAA ook deel uit van de stuurgroep van het ModeMuze project, waar verschillende mode-thesauri opgeschoond worden, omgezet worden naar SKOS en gelinked worden aan elkaar.

In het najaar wordt het project op minstens twee externe events voorgesteld:

- het zal gepresenteerd worden tijdens een aankomende “Themadag Linked Data”, die zal worden georganiseerd in samenwerking met Platform Linked Data Nederland.
- daarnaast wordt het onderdeel van het de COnnected Data studiedag die door Meertens, Beeld en Geluid, RCE en DEN wordt georganiseerd in het najaar.

VRT is een belangrijke speler in dit project geweest. VIAA plant een workshop met VRT medewerkers om de resultaten van dit project te bespreken. Mogelijk kan dit een nieuwe fase betekenen in het leven van de VRT thesaurus, waarbij de data beschikbaar is voor annotering over de VRT grenzen heen en meteen ook gelinkt kan worden aan nederlandse data.

In het landschap van het Nederlands digitaal erfgoed sluit het project aan bij de Netwerk Digitaal Erfgoed (NDE)[6]. GTAA is een belangrijke digitale bron binnen de infrastructuur van dit netwerk en door middel van de links gecreëerd binnen dit project wordt zichtbaar wat de impact van het Netwerk binnen de Taalunie is. Daarnaast zal de demonstrator gebruikt worden als state-of-the-art voorbeeld van ontsluiting van gelinkte collecties.

Verder wordt het project (een web-versie van dit rapport) online ontsloten op de websites van Beeld en Geluid en VIAA. Het project en de demonstrator zullen verder gedissemineerd via de Beeld en Geluid nieuwsbrief (3000 leden), de VIAA nieuwsbrief (2000 leden) en door middel van een aantal blog posts. Deze zullen in ieder geval verschijnen op de pagina’s van VIAA (www.viaa.be), Beeld en Geluid (http://labs.beeldengeluid.nl), de Web en Media groep van de VU (http://wm.cs.vu.nl).

Tenslotte zijn de projectmedewerkers voornemens om de resultaten van het project te publiceren in de vorm van een paper voor een workshop of conferentie in het gebied van Linked Data.

Lessons learned en toekomstig werk

Omzetten VRT thesaurus

De vertaling naar SKOS, vertrekkende van een gestructureerde lijst woorden is een zeer haalbare kaart. Dit werk leverde een goed inzicht op de inhoud van de thesaurus en leerde ons dat een op het eerste zicht vrij eenvoudige lijst toch vrij goed verrijkt kan worden met hiërarchie, relaties en scopeNotes.

De omzetting zelf gebeurde door de mensen van het Data Science Lab van de Universiteit Gent. We stippen ook even de randvoorwaarden aan voor dit soort taak: de omzetting zelf is een technische zaak (affiniteit met databanken, SKOS, XML, CSV is een noodzaak), maar kennis van de interne structuur van de bron-thesaurus levert ook een belangrijke tijdswinst. Die kennis hebben we in de loop van het project opgebouwd, waardoor het een kleine tien dagen duurde om tot een cleane SKOS versie te komen. Mits kennis van de thesaurus kan

[6] http://www.den.nl/pagina/511/netwerk-digitaal-erfgoed/

dit wellicht tot de helft herleid worden. Op dit moment kan het gedane werk op dezelfde (maar inhoudelijk meer uitgebreide) VRT-databron zonder problemen hernomen worden op minder dan een halve dag en op termijn zelfs volledig geautomatiseerd verlopen.

Het uitgevoerde werk kan dus dienen als basis voor een gepubliceerde en gelinkte thesaurus, waarmee de items bij VIAA geannoteerd kunnen worden in de toekomst. Verder zijn de inhoudelijke inzichten in de thesaurus erg nuttige pointers naar een verbeterde versie van de thesaurus: zo kan de data nog beter gestructureerd worden en verder verrijkt worden. Voor dat laatste denken we aan scopeNotes maar ook links naar bestaande vocabulaires zoals VIAF, GeoNames of andere thesauri.

De ervaring die opdeden tijdens dit project toont ons niet alleen een aantal mogelijke verdere stappen, maar kan ook als use case gebruikt worden om andere organisaties te overtuigen om huidige niet-gestandaardiseerde en/of gepubliceerde thesauri te ontsluiten als linked data.

Mapping strategieën en geproduceerde links

In het algemeen is het *alignen* van thesauri een lastige taak, zelfs wanneer de labels van de termen in dezelfde taal beschikbaar zijn (in dit geval het Nederlands). Een van de redenen is dat er om allerlei redenen verschillende structuren zijn aangebracht in de thesauri. De verschillende filtermogelijkheden die de CultuurLINK tool bied stelt ons in staat om in de vrij platte GTAA, die op het hoogste niveau verdeeld is in een aantal assen en de VRT thesaurus, waar veel structuur inzit is het daarom lastig de overeenkomende delen te identificeren. Dit maakt het mogelijk om verschillende label matchers die van toepassing zijn op verschillende typen termen effectief toe te passen. Hoewel de thesauri een andere oorsprong, structuur en gebruik kennen is toch een zeer grote hoeveelheid links gevonden. Daarnaast is het waarschijnlijk dat de vier strategieen die binnen dit project ontwikkeld kunnen dienen als blauwdrukken voor andere mapping strategieen.

Ook hier is een gemengd team het best geplaatst om de links te vinden. In het ideale geval zitten hier personen in die de thesauri (of hun eigen thesaurus) inhoudelijk kennen en zicht hebben op de structuren. Zij kunnen pointers geven om de mapping zo vlot mogelijk te laten lopen. Het team wordt idealiter aangevuld door een medewerker met wat technische achtergrond, die in staat is de mapping strategie uit te tekenen op basis van vergelijking van *strings, fuzzy matching,* etc. Een tool zoals cultuurlink wordt bij elke iteratie gebruiksvriendelijker en vereist steeds minder technische kennis. Naar verwachting zou dit soort tools mettertijd in die mate gebruiksvriendelijk moeten worden dat de linking kan gebeuren door personen zonder technische kennis.

De links zelf kunnen verder worden onderzocht. Met name door te kijken naar wat er wel en niet gemapt is, kan verder inzicht worden verkregen in de keuzes die gemaakt zijn in het ontwerp van de beide thesauri en eventuele fouten opgespoord worden.

Verder linken

Naast de bestaande links tussen VRT en GTAA, die bij wijze van piloot gelegd werden uit beide thesauri zijn mogelijks nog een aantal andere pistes interessant. GTAA is op dit moment bijvoorbeeld ook gelinked aan andere thesauri en datasets zoals Wordnet[7] en DBPedia[8]. Deze thesauri zijn zelf ook gepubliceerd op het web. Mits de bestaande links is nu ook VRT gelinked aan de bestaande thesauri.

Het linken aan derde thesauri heeft potentieel nog een voordeel. Mochten beide thesauri gelinked zijn aan een derde thesaurus (bvb. VIAF voor personen), dan kunnen deze links op hun beurt gebruikt worden als match tussen de GTAA en VRT thesaurus. Daarnaast heeft een dergelijke link ook potentieel om de bestaande thesaurus te verrijken. In beide gevallen leidt dit tot rijkere source databases wat uiteindelijk zal leiden tot een betere doorzoekbaarheid van de assets zelf.

Uitbreiding collectie en herbruikbaarheid van de demonstrator

De demonstrator zoals die nu gebouwd werd illustreert mooi het gebruik van thesauri om binnen de collecties of over de collecties heen de doorzoekbaarheid en suggesties te verbeteren. Op dit moment is de demonstrator gevuld met een vrij klein aantal media-items, die bovendien inhoudelijk vrij ver uit elkaar liggen.

Zowel VIAA als Beeld en Geluid zijn nog dit jaar van plan om een deel van de metadata van de collecties te publiceren voor test en demo-doeleinden. Indien deze metadata mee opgenomen zou worden in de demonstrator, kan dit enkel leiden tot mooiere resultaten op het vlak van de demonstrator, waardoor we overtuigd zijn dat deze gedurende 1 a 2 jaar kan dienen als een van de piloot en illustratie van de mogelijkheden en haalbaarheid van dit soort technologie.

[7] Malaisé, Véronique, et al. "Anchoring dutch cultural heritage thesauri to wordnet: two case studies." *ACL 2007* (2007): 57.
[8] Bouma, Gosse. "Cross-lingual Ontology Alignment using EuroWordNet and Wikipedia." *LREC*. 2010.