Nieuw leven voor thesauri

Naar een hogere kwaliteit van de informatiehuishouding

Wil Rombout

[Od april 2015 Special](http://od-online.sdu.rijksweb.nl/editie/od-april-2015-special)

Hoe kunnen we in de dynamiek van een digitaliserende samenleving gegevens zo ontsluiten dat we ze vanaf hun oorsprong kunnen vinden, begrijpen en kunnen vertrouwen dat ze juist en volledig zijn? Het antwoord ligt in een eenduidig beheer van begrippen en het documenteren van de betekenis en relationele context(en) van de gegevens die we ontvangen, creëren en uitwisselen. Een verhaal over de inzet van thesauritechnieken.

De overheid is voor een belangrijk deel te typeren als een kennisintensieve organisatie. Niet voor niets noemen we de medewerkers in het primair proces kenniswerkers. We werken steeds meer interactief samen, over de organisatiegrenzen heen, en verwerken daarbij grote hoeveelheden gegevens. Gedigitaliseerde communicatievormen zorgen tevens voor een makkelijke, snelle en goedkope uitwisseling van gegevens, zowel intern als extern.

Dat roept de vraag op hoe we in een de dynamiek van een digitaliserende samenleving gegevens zo ontsluiten dat we ze vanaf hun oorsprong kunnen vinden, begrijpen en kunnen vertrouwen dat ze juist en volledig zijn. En *by the way* ook zorgvuldig is omgegaan met de gevoeligheid van gegevens. Dat stelt kwalitatief hogere eisen aan de inrichting van de informatiehuishouding. Alleen dan kunnen we de interne en externe transparantie vergroten, en daarmee de tevredenheid van burgers, wat leidt tot meer vertrouwen in de overheid.

Het antwoord ligt in een eenduidig beheer van begrippen en het documenteren van de betekenis en relationele context(en) van de gegevens die we ontvangen, creëren en uitwisselen. Hierdoor kunnen we gegevens interpreteren, gebruiken en al dan niet geautomatiseerd verwerken. Semantische modellen als metagegevens, gecontroleerde vocabulaire, taxonomieën, et cetera zijn instrumenten waarmee we dit kunnen realiseren.

**Gebruik van thesauri1**
De Erfgoedsector zet hier al belangrijke stappen in door de inzet van thesauritechnieken. Een thesaurus is een gestructureerde termenlijst van concepten.2 De opbouw is hiërarchisch maar kent ook associatieve en equivalente relaties.

* Equivalentierelatie
Een equivalente relatie (ook wel gelijkwaardigheidsrelatie genoemd) betreft de relatie tussen termen met een gelijke betekenis zoals synoniemen, quasisynoniemen, afkortingen en samenstellingen. Het regelt verder het gebruik van een voorkeursterm (*descriptor*) en verwijstermen (*non-descriptor*).
* Hiërarchische relatie
De hiërarchische relatie groepeert voorkeurstermen in een hiërarchische boomstructuur. De standaarden spreken van *broader terms* (bt) en *narrower terms* (nt). Er zijn twee typen hiërarchische relaties: de generieke (genus >> species) en de partiële relaties (geheel >> deel).
* Associatieve relaties
De associatieve relaties (*related terms*) zijn relaties tussen termen die een inhoudelijk verwantschap met elkaar hebben en niet als een van de voorgaande relatie worden aangemerkt. Deze associatieve relaties zijn vooral interessant als ze leiden tot nuttige verwijzingen voor de gebruiker.

Naast deze drie relaties kan aan een term ook een zogenaamde *scope note* toegevoegd worden. Het kan gaan om een definitie van het gebruik van een term binnen de context van de thesaurus.

Een thesaurus is een goed instrument om de verschillen in taalgebruik tussen de ontsluiter of indexeerder en de in- of externe informatiezoeker te overbruggen. Redenen voor het gebruik van een thesaurus zijn:

* De natuurlijke taal van auteurs, indexeerders en gebruikers is vertaald naar een gecontroleerde woordenschat.
* Het verzekert de consistentie in het toewijzen van indextermen.
* Het geeft betekenisvolle relaties tussen termen weer.
* Het is een goede zoekleiding.

Naast dit algemene gebruik in *information retrieval*, stellen Aitchison et al. (2000, p. 1) dat een thesaurus ook kan helpen bij het verwerven van een algemeen begrip van een kennisdomein, bij het opstellen van semantische netwerk3 en bij het definiëren van termen.

Thesauritechnieken helpen informatieobjecten te duiden, te ordenen, op te slaan, te verbinden en te presenteren. Zonder volledig te zijn kunnen de volgende functies aan een thesaurus worden gehangen:

* trefzekere contextgeoriënteerde bevraging van bronnen;
* inzicht in de beschikbare informatie (informatieregister);
* mogelijk maken van onverwachte zinvolle zoekresultaten (zowel door synoniemtoepassingen als door koppeling van verwante thesauri);
* in kaart brengen van kennisleemtes, structuren van werk, onderzoek en informatieverzameling;
* wetenschappelijke kennisontwikkeling door bijdrage van collectief erkend begrippenapparaat;
* verbonden expertise, mensen en organisaties door verbonden informatie;
* thesauri zijn kennisbronnen op zichzelf, zeker wanneer scope notes, afbeeldingen en bronnen zijn toegevoegd: geïllustreerd verklarend woordenboek.

**Hoe zijn gegevens uit bronnen verbonden met thesauri in een semantische laag?4**
Uitgangspunt voor de verbinding is een driedeling (zie figuur 1)

1. presentatielaag (websites, kaarten, dashboard, Mijn digitale werkplek, …) NB. Ediensten zijn vormen van een presentatielaag;
2. semantische verbindingslaag met thesauri en geografische koppelmogelijkheden;
3. bronen doelsystemen.

Nietgestructureerde bronnen (bijvoorbeeld teksten) bevatten gegevens en metagegevens, met – als het netjes is opgezet – een ID (bijvoorbeeld per webpagina of document). Gestructureerde bronnen (databases) hebben naast gegevens per record ook een unieke ID. De koppeling vindt plaats in de verbindingslaag. Daar koppelen de ID’s van de metagegevens van de *content*items aan de concepten in de thesauri en hun ID’s.

Voor de koppeling wordt gebruik gemaakt van standaardtechnologie die daarvoor is ontwikkeld in de internetwereld. Die standaarden heten SKOS en RDF.5
De manier waarop relaties weergegeven worden, is6:

* Er wordt een relatie aangegeven tussen twee contentitems via een eigenschap.
* Er is bijvoorbeeld een tekst waarin melding wordt gemaakt dat gebouw X is opgetrokken uit baksteen. Gebouw X hasItemType = listedBuilding, hasMaterialbaksteen. In formele termen wordt dat in de verbindingslaag opgelost via de door SKOS voorgeschreven manier: [object – predicate – subject]

Het maken van deze koppeling is redactiewerk. Iemand doet dat. Je kunt dat semiautomatiseren, het systeem helpt je een handje door koppelingen voor te stellen, maar iemand neemt de beslissing, bijvoorbeeld op basis van de mate waarin baksteen kenmerkend is voor het gebouw.

**Koppelen en zoeken met thesauri/verbindingslaag**
De semantische laag bevat een groot aantal van deze termenlijsten met onderlinge verwijzingen. Connemaramarmer bijvoorbeeld is als bouwmateriaal gekoppeld aan een land van oorsprong, aan bouwkundige termen zoals vloeren, plinten, wanden en meubilair. De focus hier is van toepassing in monumenten, maar dit type marmer wordt ook gebruikt voor beeldhouwwerken. Door deze nieuwe koppeling kan ook worden gezocht in kunstwerken.

In RDFstructuur wordt dit als volgt opgeslagen:

|  |  |
| --- | --- |
| prefLabel | Connemara marmer |
| itemType | materiaal |
| useForBuilding | vloertegel |
| useForBuilding | plint |
| useForBuilding | wand |
| hasApplication | beeldhouwwerk |
| hasCountryOfOrigin | Ierland |

**Wat kun je er nog meer mee doen?**
Als we de semantische laag verbinden met het register van rijksmonumenten (ODB), met de beeldbank (in Pictura) en de bibliotheek (in AdLib), kunnen we met één term zoeken binnen verschillende bronnen.

Bij ODB-monumentnummer 2064 gelegen aan Ligusterlaan 14 te Bilthoven vinden we de volgende beschrijving:
*Hoekhuis (1912) ontworpen door J. De Gooier met rijke interieur detaillering in jugendstil. Opvallend is de rijk versierde tegelvloer in hal, uitgevoerd in Connemara-marmer*.

Via de zoekterm Connemaramarmer verwacht je dus dit monument te kunnen vinden. In dit voorbeeld is de term Connemaramarmer via zijn conceptID verbonden met het monument in de Ligusterlaan 14 Bilthoven, waar Connemaramarmer als vloertegel is toegepast.
Maar je vindt meer door alle verbindingen.

Wat je zou kunnen vinden (versimpeld voorbeeld):

* de beschrijving van het monument;
* een geografische positie (kaartje met stip);
* het adres;
* de architect;
* bouwtekeningen en monumentale panddossiers;
* de bouwstijl waarin het ingedeeld is;
* verschillende foto’s, waaronder een foto van de vloer;
* een lijstje boeken en artikelen over marmer, de architect of de bouwstijl.

**Toegevoegde waarde informatiehuishouding**
De sector cultureel erfgoed heeft de ambitie de beschikbare gegevensbronnen actief beschikbaar te stellen voor de inen externe omgeving. Dat doen zij, zoals hiervoor beschreven, door hun kennisdomein met thesaurustechnieken semantisch en contextueel te modelleren. Dat doen zij met het Linked Dataconcept, met als doel metagegevens *repository* te realiseren, waarmee gegevensobjecten op een gestandaardiseerde en gestructureerde ontsloten worden. In het domein van de informatiehuishouding hebben we de opdracht zoveel als mogelijk generiek te ontsluiten en de hiervoor beschreven ontwikkeling biedt een handelingsperspectief. Achterliggende gedachte is de rijksbrede generieke (en verplichte) metagegevens7 met dit concept te modelleren tot een administratieve thesaurus met als doelen:

* de relevante (verplichte) elementen van het toepassingsprofiel te modelleren (de portfolio van gecontroleerde woordenlijsten);
* de administratieve thesaurus te positioneren als de bron voor de gecontroleerde metagegevens (eenmalige opslag);
* de metagegevens via een webservice beschikbaar stellen aan de onderliggende applicaties (meervoudig gebruik);
* als bron te dienen voor het contextueel zoeken van informatieobjecten.

Hiermee zijn we in staat het beheer van de portfolio van metagegevens (vocabulaires) vanuit een centrale bron te beheren. Spreekt voor zich dat dit de kwaliteit van de informatiehuishouding en de interoperabiliteit van gegevensobjecten te goede komt. Niet onbelangrijk ook te benoemen dat de administratieve thesaurus ook kan dienen als het door de politiek gewenste ‘informatieregister’.

Van belang is het te verbinden met het werken onder architectuur, zoals hergebruik van metagegevens, bedrijfsregels, taalherkenning et cetera. Daar zal in een volgend artikel uitgebreider op worden ingegaan.

warom@ziggo.nl, *Wil Rombout is redactielid van Od*.

**Noten**
1 Thesauri, interne discussienotitie, Rijksdienst Cultureel Erfgoed 2013.
2 Thesaurus: gestructureerde lijst van concepten (ISO 2788). Concepten zijn volgens deze standaard ‘eenheden van begrip’.
3 Essentiële factor in een ondersteunende webgebaseerde kennisinfrastructuur.
4 Hier wordt enkel ingegaan op (indirecte) semantische relaties, waarmee de indruk wordt gewekt dat verbinden op basis van semantiek alleen afdoende is. Gegevens zijn ook (direct) verbonden door geometrische en administratieve relaties.
5 <http://www.frankwatching.com/archive/2008/06/09/de-klassieke-benadering-van-web-30/>
6 Voorbeeld uitwerking van de Rijksdienst voor Cultureel Erfgoed.
7 Toepassingsprofiel Metagegevens Rijksdienst.