Vinden en verbinden met taal

Meer waarde uit data en informatie

Danny Greefhorst

Data is de nieuwe haarlemmerolie. Organisaties willen datagedreven worden en hun beslissingen baseren op feiten. Ze willen kunstmatige intelligentie, machine learning en deep learning inzetten om diepere inzichten te halen uit hun data. Dit alles veronderstelt dat duidelijk is wat de data betekent en juist daar gaat het vaak fout. Het geven van betekenis vraagt om taal. Het vinden van de juiste woorden en de daarbij behorende definities is de basis. Door deze woorden te verbinden aan data en informatie kan er naast betekenis ook meer waarde worden gecreëerd. Dit whitepaper geeft inzicht in deze waarde van taal voor data en informatie en beschrijft een visie om er invulling aan te geven.

Bruikbaarheid van data en informatie

Data is de digitale grondstof en informatie is de betekenis van deze grondstof in een bepaalde context. Er ontstaat **steeds meer data en informatie** in organisaties en de variëteit van data neemt ook steeds meer toe. Naast traditionele administraties wordt ook steeds meer gebruik gemaakt van data uit externe bronnen. Het is inmiddels duidelijk dat veel data en informatie buiten traditionele databases te vinden is. Het zit in documenten, e-mails, sociale media, API's en een grote diversiteit van opslagmechanismen. Het is logisch dat organisaties proberen meer waarde te creëren uit al dit soort bronnen.

Tegelijkertijd is er in veel organisaties nog relatief weinig tijd besteedt aan het op orde brengen van data. De aandacht voor data governance en data management neemt toe, maar veel organisaties staan nog aan het begin. De praktische consequentie is dat van veel data **geen definities** aanwezig zijn en onbekend is wat de kwaliteit ervan is. Kennis over de data zit in de hoofden van allerlei mensen, maar dit levert bij elkaar geen consistent, volledig en actueel beeld op. Het leidt vooral tot verwarring.

Met de toename van data en informatie is er ook een grotere verscheidenheid aan bronnen. Al deze bronnen geven een deelperspectief op de werkelijkheid van de organisatie. Het vinden en bij elkaar krijgen van deze data blijkt in de praktijk vaak lastig. Doordat **soortgelijke data op meerdere plaatsen is vastgelegd**, met andere en ontbrekende definities blijkt het ook lastig om een consistent beeld te creëren van al deze data. De kwaliteit van de data blijkt laag.

Er wordt wel steeds meer informatie gebaseerd op al dit soort data. Management wordt graag voorzien van allerlei dashboards en rapportages. De data wordt ook getoond in allerlei digitale uitingen zoals websites, persoonlijke portalen, mobiele apps en via geautomatiseerde berichtuitwisselingen. Het is zorgwekkend hoeveel vertrouwen wordt gegeven aan **data en informatie die eigenlijk onbetrouwbaar is**. De kans is groot dat klanten, samenwerkingspartners en medewerkers verkeerd worden geïnformeerd en verkeerde keuzes maken. Van een datagedreven organisatie is al helemaal geen sprake.

De kracht van taal

Alles wat we denken neemt de vorm aan van taal. Taal geeft betekenis aan mensen. Het is onze manier om de wereld te begrijpen en te beïnvloeden. Mensen organiseren door te communiceren. Data krijgt betekenis door er woorden aan te geven. Datadefinities zijn vooral zinnen die duidelijk uitleggen wat de maker van de data heeft bedoeld. Het creëren van informatie over de data heeft dus ook alleen zin als je deze taal goed begrijpt. Als je andere aannames doet over de data is de kans groot dat je misinformatie creëert.

Data is een weergave van de werkelijkheid. Het is daarmee essentieel om eerst de werkelijkheid goed te begrijpen voordat je deze in data probeert te verpakken. Dat doe je door woorden te geven aan de werkelijkheid en deze woorden vervolgens duidelijk te definiëren. Dit soort definities zitten dus op een taalniveau en hebben nog niets te maken met data. Dat doe je pas als je datamodellen opstelt die aangeven hoe de werkelijkheid gerepresenteerd moet worden in datastructuren. Deze datastructuren kunnen dicht aanliggen tegen de werkelijkheid, maar je kunt er ook voor kiezen andere ontwerpkeuzes te maken. Je creëert dan een nieuwe datawerkelijkheid.

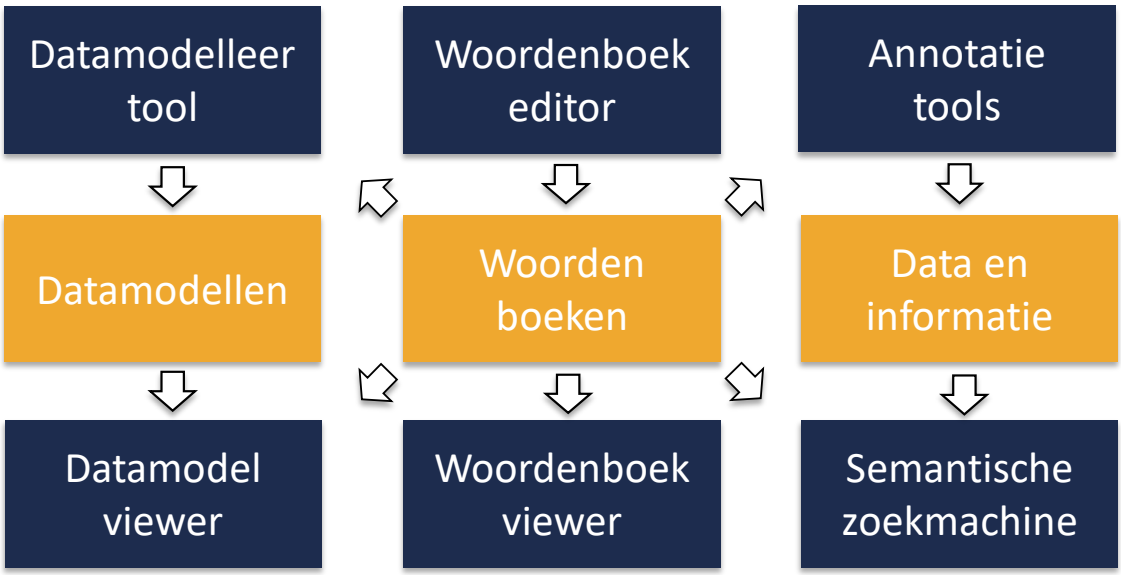
Als je de juiste woorden hebt gevonden en gedefinieerd dan ontstaan begrippen waar je van alles mee gaan doen. Je kunt deze begrippen als het ware op van alles gaan “plakken”. Denk bijvoorbeeld aan het plakken van deze begrippen op data en datamodellen, maar vooral aan allerlei uitingen ervan zoals documenten, websites, rapporten en dashboards. Als iemand data, informatie of een representatie ervan in een tekst, tabel, diagram of infographic wil begrijpen dan kan hij of zij deze begrippen gebruiken. Als je toegang krijgt tot de betekenis van de woorden dan begrijp je beter waar je naar kijkt. Als je daarnaast ziet welke andere woorden daaraan gerelateerd zijn, dan begrijp je de context ook veel beter.

Een voorbeeld van het gebruik van begrippen is weergeven in onderstaand figuur. Dit is een uitsnede van een website met informatie over het Middelbaar Beroeps Onderwijs – de Triple A Encyclopedie. Deze encyclopedie bevat allerlei informatie over processen in het MBO. Het bevat ook een begrippenlijst waarin alle belangrijke woorden zijn gedefinieerd. Deze woorden zijn ook onderstreept in de teksten op de website en hun definitie wordt getoond als de muis over het woord heen wordt bewogen. Hierdoor ontstaat meer begrip van de tekst.

Begrippen creëren ook zoekingangen. We weten allemaal dat je door het intypen van woorden in Google het gehele Internet kunt doorzoeken. Of je dan ook echt vindt wat je zoekt, is echter maar de vraag. Als je data en informatie vooraf voorziet van de juiste begrippen, dan is de kans veel groter dat je ook de juiste data en informatie kunt vinden. Als je de begrippen plakt op een grote diversiteit aan bronnen, dan kun je ook slimmer zoeken in veel meer data en informatie. Begrippen vormen dan ook een middel om data en informatie te integreren. Je krijgt alles wat bij de opgegeven begrippen hoort. De begrippen vormen een soort sateprikker door de data.

Een taalinfrastructuur

Om de kracht van taal in te kunnen zetten heb je middelen nodig om taal te definiëren, te ontsluiten en te gebruiken; een taalinfrastructuur. Deze infrastructuur kan allerlei vormen aannemen, afhankelijk van de context. De kern van deze taalinfrastructuur is weergegeven in onderstaand figuur.



Centraal in de taalinfrastructuur staan de woordenboeken. Een woordenboek wordt ook wel een thesaurus, begrippenlijst of business glossary genoemd. Je hebt een specifieke applicatie (editor) nodig om begrippen op een rijke manier te definiëren. Woorden en hun definities kunnen overigens ook op andere plaatsen ontstaan en aan een woordenboek worden toegevoegd. Daarnaast is het waardevol om verschillende woordenboeken aan elkaar te verbinden, zodat zicht ontstaat op het verschillend gebruik van woorden. Ook het ontsluiten van de woordenboeken vraagt een specifieke applicatie om dat gebruikers-vriendelijk te doen. Het gebruik van de begrippen in de woordenboeken is overigens veel breder; allerlei applicaties zullen de woordenboeken raadplegen en de definities op een contextspecifieke manier tonen.

Het woordenboek staat ook centraal bij het definiëren en ontsluiten van datamodellen (ook wel: informatiemodellen). Objecttypen, attributen en waardenlijsten zijn bedoeld om de werkelijkheid zoveel mogelijk te representeren. Deze elementen verwijzen daarom zoveel mogelijk naar de woorden en hun definities in het woordenboek. Alleen als er specifieke aanvullende woorden nodig zijn in het datamodel, dan kunnen zij een eigen definitie in het datamodel vragen. Denk bijvoorbeeld aan een generalisatie van een “klant”, “leverancier” of “medewerker” naar een “partij”. Dat laatste woord maakt mogelijk geen deel uit van de taal van de organisatie en kan lokaal in het datamodel worden gedefinieerd.

Het gebruik van het woordenboek zit voor een belangrijk deel ook in het toekennen van begrippen aan data- en informatie-elementen (ook wel: annoteren). Er zullen specifieke annotatietools zijn die het mogelijk maken om de begrippen op allerlei data en informatie en hun uitingen te plakken. Deze annotaties zijn in veel gevallen niet direct zichtbaar voor gebruikers, maar zullen gebruikt worden door specifieke applicaties. Die kunnen dat bijvoorbeeld gebruiken om de definities van woorden in een tekst te tonen, op een slimmere manier te kunnen zoeken of de data op een meer betekenisvolle manier uit te wisselen. Er zal ook een organisatiebrede semantische zoekmachine zijn die in alle belangrijke bronnen van data en informatie op een slimme manier zoekt.

Taalstandaarden - SKOS

Als er een serieuze keuze wordt gemaakt voor het implementeren van een taalinfrastructuur dan zijn verdere standaarden, richtlijnen en afspraken noodzakelijk. Een belangrijk aspect daarvan is het definiëren van een datamodel en uitwisselstandaarden voor de woordenboeken. Deze keuzes kunnen specifiek zijn ingegeven door eerder gemaakte keuzes in de organisatie. Als er bijvoorbeeld reeds een specifiek datamodelleertool en bijbehorende modelleerstandaarden zijn gekozen, dan is het logisch daar maximaal op aan te sluiten.

Veelvoorkomende keuzes voor standaarden zijn het gebruik van UML of het gebruik van Linked Data standaarden. Deze Linked Data standaarden zijn onderdeel van de visie om te komen tot een wereldwijd semantisch web, waarbij alle informatie op het Internet is geannoteerd met woorden. Het sluit erg goed aan bij de visie in dit whitepaper en is dus een voor de hand liggende keuze. Het sluit ook goed aan bij de open data intenties van de overheid, waarbij data zoveel mogelijk publiek beschikbaar wordt gesteld via het Internet middels open standaarden. We zullen dit scenario daarom nader uitwerken.

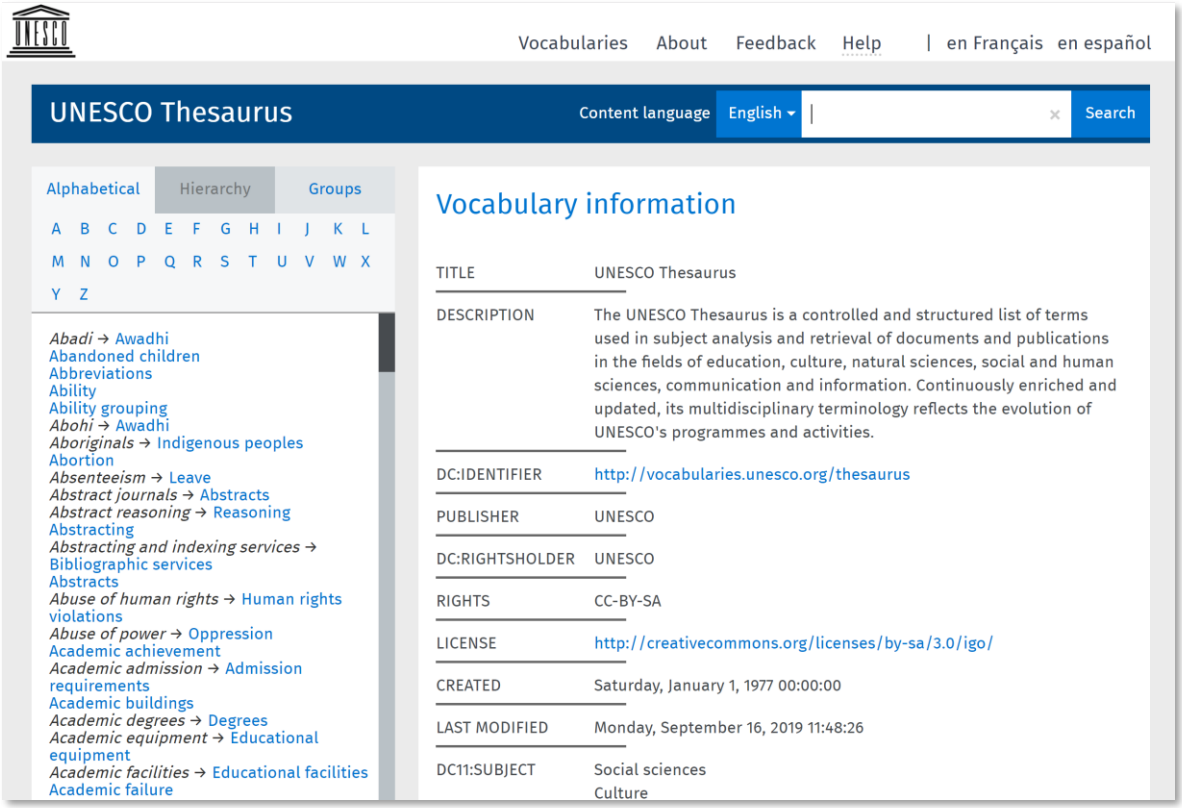
De belangrijkste Linked Data standaard voor het definiëren van woordenboeken is SKOS (Simple Knowledge Organization System). Het is een standaard van het W3C waarmee woorden in een thesaurus worden gedefinieerd. Het staat ook op de “pas toe, leg uit” lijst van standaarden van Forum Standaardisatie en is daarmee verplicht voor overheidsorganisaties. Een thesaurus is te zien als een woordenboek waaraan meer structuur is gegeven door ook relaties tussen begrippen aan te brengen. Begrippen kunnen meer specifiek of meer algemeen zijn dan andere begrippen of een meer algemene relatie hebben tot andere begrippen. Begrippen worden uniek geïdentificeerd door een URI (uniform Resource Identifier). Dat ziet er uit als het adres van een website en de definitie van een begrip is idealiter ook echt als zodanig toegankelijk via een web browser.

Er zijn allerlei tools beschikbaar voor het werken met SKOS. Skosmos is een open source webgebaseerde viewer voor SKOS thesauri. Het is geoptimaliseerd om thesauri die bestaan uit een groot aantal begrippen op een efficiënte en gebruikersvriendelijke manier te ontsluiten. Een voorbeeld van een Skosmos thesaurus is op de volgende pagina weergegeven. Het is de thesaurus van UNESCO die is bedoeld om onderwijs, cultuur en verschillende wetenschappen te ondersteunen. Een vergelijkbaar tool is OpenSKOS, dat ook gebruikt kan worden als editor. Semantische wiki’s zijn ook goed geschikt om thesauri in te beheren. Het is echter met een kleine inspanning ook mogelijk om SKOS te genereren uit bijvoorbeeld UML.

Uitwisselen van begrippen

Voor het uitwisselen van data is het tegenwoordig logisch om deze via API’s te ontsluiten. Dat geldt ook voor woordenboeken. Een woordenboek API kan in allerlei applicaties worden gebruikt om woorden en hun definitie op te halen. Er is op dit moment nog niet één API die breed wordt omarmd door leveranciers, waardoor de keuze vooral is om gebruik te maken van een API van een specifieke leverancier of om zelf een API te definiëren. Eerder genoemde tools Skosmos en OpenSKOS bieden API’s die direct gebruikt kunnen worden. Er kunnen echter redenen zijn om toch een eigen API te definiëren, zoals het voorkomen van toolafhankelijkheid. De GraphQL standaard voor het zoeken in data is ook veelbelovend en zou ook een basis kunnen zijn voor een woordenboek API. Wikipedia is ook een belangrijke bron voor definities. DBpedia is bedoeld om de informatie op Wikipedia op een machinetoegankelijke wijze leesbaar te maken, gebaseerd op Linked Data standaarden.

Een woordenboek is ook gewoon referentiedata en dient dan ook als zodanig gemanaged te worden. Het kan wenselijk zijn het woordenboek via een database of in de vorm van een bestand beschikbaar te stellen voor applicaties of ander gebruik. Idealiter is er dan wel ook een mechanisme dat ervoor zorgt dat wijzigingen snel ook toegankelijk zijn en waarbij afnemers zich op wijzigingen kunnen abonneren.



Kunstmatige intelligentie

Het definiëren en gebruiken van taal opent ook geheel nieuwe mogelijkheden. Deze toepassings-mogelijkheden worden vaak geschaard onder de verzamelnaam “kunstmatige intelligentie” (of “artificial intelligence” in het Engels). Een belangrijk onderdeel hiervan is het ontleden van natuurlijke taal in geschreven of gesproken teksten. Door een dieper begrip te creëren van zinnen, de daarin gebruikte woorden en hun volgorde kan ook een diepere betekenis worden toegekend aan teksten. Een populaire toepassing in deze categorie zijn chatbots, die bedoeld zijn om menselijke call center medewerkers te vervangen. De basis voor chatbots is ook een woordenboek.

De ontwikkelingen op het gebied van kunstmatige intelligentie gaan snel. De ultieme test is de Turing test waarbij je niet meer weet of je met een mens of met een computer in gesprek bent. Voorlopig kun je er in ieder geval van uitgaan dat we als mens zelf betekenis zullen moeten geven. Het is ook aan onszelf om woordenboeken te definiëren en te gebruiken.

Begrippenlijst

- **Annoteren**: Het toekennen van begrippen aan (delen van) data en informatie.
- **Begrip**: Een term voorzien van een definitie.
- **Data**: Weergave van een feit, begrip of aanwijzing, geschikt voor overdracht, interpretatie of verwerking door een persoon of apparaat.
- **Datamodel**: Een formele definitie van objecttypen, attributen, relaties en regels.
- **Informatie**: De betekenis van gegevens in een specifieke context.
- **Informatiemodel**: zie *datamodel*.
- **Kunstmatige intelligentie**: De wetenschap die zich bezighoudt met het creëren van een artefact dat een vorm van intelligentie vertoont.
- **Objecttype**: Een type van gelijksoortige data.
- **Semantische zoekmachine**: Een zoekmachine die gebruik maakt van de begrippen die zijn toegekend aan (delen van) data en informatie bij het zoeken.
- **Simple Knowledge Organization System (SKOS)**: Een standaard voor het definiëren van thesauri.
- **Taalinfrastructuur**: Een verzameling middelen die ondersteuning bieden bij het definiëren en gebruiken van begrippen.
- **Term**: zie *woord*.
- **Thesaurus**: Een verzameling van begrippen en hun onderlinge relaties.
- **Unified Modeling Language (UML)**: Een standaard voor het definiëren van softwaresystemen.
- **Uniform Resource Identifier (URI)**: Een manier om objecten wereldwijd uniek te identificeren op het web.
- **Woord**: Het kleinste zelfstandig gebruikte taalelement op het niveau van de spreektaal.
- **Woordenboek**: Een verzameling van begrippen.

Over ArchiXL

ArchiXL is een onafhankelijk adviesbureau, gespecialiseerd in enterprise- en informatie-architectuur. Wij adviseren organisaties bij het operationaliseren van hun strategie. De naam ArchiXL is een samenvoeging van "Architectuur" en "XL", waarbij XL staat voor “excelleren”. Wij helpen organisaties om hun doelen te bereiken waardoor zij kunnen excelleren. Onderscheidend daarbij is onze pragmatische en doelgerichte werkwijze. Dat zorgt dat we sterk gericht zijn op het leveren van toegevoegde waarde, passend bij de context van de organisatie. Als specialist op het gebied van architectuur kennen we alle relevante methoden en technieken en weten we als geen ander wat de valkuilen zijn. Onze medewerkers onderscheiden zich door hun communicatieve vaardigheden, resultaatgerichtheid, en hun abstractie- en inlevingsvermogen.

Het is onze passie om de doelmatigheid en effectiviteit van veranderingen en de wijze waarop architectuur en kennis daarbij worden toegepast te verbeteren. Wij denken dat mensen en hun kennis daarin een centrale rol spelen. Het is belangrijk om de specifieke kennis, vaardigheden en talenten van mensen te zien en maximaal in te zetten voor de doelstellingen van de organisatie. De basis daarvoor is een goed gesprek en een goed luistervermogen. In onze visie wordt architectuur nog onvoldoende effectief ingezet om de organisatie te ondersteunen. Symptomen hiervan zijn ontoegankelijke architectuurdocumenten, abstracte modellen die niet aansluiten bij de praktijk en architecten die zich afzonderen van de organisatie. Door kennis te mobiliseren zet je anderen in hun kracht en kom je samen tot grote hoogte.

Onze principes

- Standaard methode – onze aanpak is gebaseerd op standaard methoden en technieken zoals ArchiMate en TOGAF, en daarmee op uitgebreide kennis en ervaring van anderen.
- Hergebruik – organisaties lijken in veel opzichten op elkaar en hergebruik van kennis en architecturen is daarom verstandig.
- Iteratief werken – het is belangrijk om snel antwoord te geven op vragen vanuit de organisatie; dit hoeft niet altijd een volledig antwoord te zijn.
- Concrete en bruikbare resultaten – architectuurproducten moeten direct bruikbaar zijn en waarde opleveren voor de organisatie.
- Samenwerking – veranderen doe je samen, daarmee bundel je ook de kennis en denkvermogen en ontstaat draagvlak voor de verandering.
- ”Just enough” architectuur – architectuurdocumenten moeten bijdragen aan de doelstellingen en niet meer beschrijven dan noodzakelijk.
- Mobiliseren kennis – architecten moeten zich vooral richten op het verzamelen, analyseren, genereren en verspreiden van kennis.

Meer weten?

telefoon: 033-2585545

e-mail: info@archixl.nl

website: http://www.archixl.nl