Datamining zonder data

Jurriaan van Rijswijk
LEI-DLO
Postbus 29703, 2502 LS Den Haag
telefoon (070) 33 08 330
e-mail: j.h.vanrijswijk@lei.dlo.nl

De Dienst Landbouwkundig Onderzoek onderzoekt welke methode er ontwikkeld moet worden om aan de hand van data inzicht te krijgen in de kennis die op de verschillende instituten aanwezig is over de groene ruimte. Dit wordt gedaan in het project Datamining Groene Ruimte, dat valt binnen het Strategisch Expertise Onderzoek (SEO) van DLO. Het onderzoek wordt uitgevoerd op 5 instituten, te weten: Instituut voor Bos en Natuur (IBN); Landbouw Economisch Instituut (LEI); Staring Centrum (SC); Rijks Instituut voor Visserij Onderzoek (RIVO) en Agrobiologisch en Bodemvruchtbaarheidsonderzoek (AB). Dit artikel gaat in op de ontwikkelde methodologie en de toepassingen voor marketing.

Datamining en Metamining

De titel van het project, Datamining Groene Ruimte, is misleidend. Datamining suggereert analyse van reeds aanwezige grote hoeveelheden gegevens om hierin bijvoorbeeld verborgen relaties te vinden. Bekend is het voorbeeld van een grote supermarktketen in de Verenigde Staten die met datamining een verband ontdekte tussen de verkoop van gereedschap en meubels. Op het moment dat consumenten bijvoorbeeld een zaag kochten werd na een periode van een maand door dezelfde consument een bed gekocht. Wat bleek, men gaat verbouwen en koopt daarna nieuwe meubels. Marketingtechnisch erg waardevolle informatie. Of de relatie tussen de verkoop van bier en luiers. Het bleek dat als bier in de nabijheid van luiers stond, dat er meer bier werd omgezet. Onderzoek wees uit dat vrouwen die luiers kochten, denkend aan hun kind, ook meteen bier meenamen, denkend aan hun man. Mannen die luiers kochten wilden zichzelf belonen met deze prestatie en namen als beloning daarom voor zichzelf bier mee. Ik vrees dat de relatie luiers en parfum nooit onderzocht is.

Datamining is dus figuurlijk 'graven' in grote bergen data. Voor het project Datamining Groene Ruimte is dat niet het geval en daarom is de titel misleidend. Voor dit project moet datamining geïnterpreteerd worden als het vinden van waardevolle informatie in beschrijvingen van data, dus niet de data zelf. Metamining dus eigenlijk, het 'graven' in beschrijvingen van data.
Een andere opmerking is dat de 'grote bergen' aan data er wel moesten zijn, maar niet een-twee-drie te vinden waren. Dit vormde één van de gedachten die geleid heeft tot het onderzoek, namelijk dat intelligent hergebruik van aanwezige data mogelijk moet zijn, maar niet optimaal gebeurt. Te vaak nog wordt data verzameld voor onderzoek terwijl het, bij wijze van spreken, een deur verder al verzameld is. Dit komt misschien door een gebrek aan communicatie of informatie. Het blijkt dat data vaak maar bij enkele personen bekend is, wanneer het niet voor een monitoring of wettelijk dienstverlenende taak verzameld is, maar voor bijvoorbeeld een (toegepast) onderzoeksproject. Dit geeft aan dat het dus naast het hebben van alleen gegevens over data (meetwaarden), het ook belangrijk is om gegevens over bijvoorbeeld personen en modellen te hebben bij wie de data is gebruikt. Om bij het voorbeeld van de zaag en meubels te blijven, weten waarvoor en door wie de zaag gebruikt gaat worden.

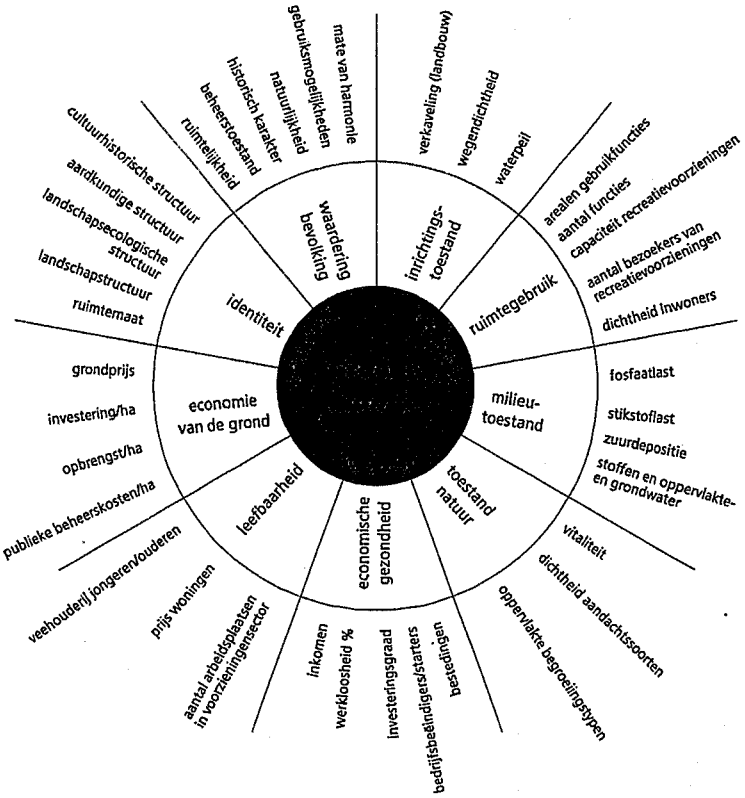
Wat is data eigenlijk en wat is kennis?

Maar wat is data nu eigenlijk en wat wordt er binnen de context van het project onder verstaan? Data op zich is niets anders dan een verzameling (alfa)numerieke gegevens. Beschrijving van data is informatie, geen meta-data of meta-informatie. Het is Informatie die toegevoegde waarde krijgt, samen met een gebruiker of in combinatie met een model, als het aangewend wordt voor beantwoording van bijvoorbeeld een onderzoeksvraag. Deze toegevoegde waarde wordt gedefinieerd als kennis. Van data naar informatie naar kennis. Dus naast geïnteresseerd te zijn in informatie is het belangrijk te weten voor welke context de informatie aangewend kan worden. Met andere woorden, we zijn geïnteresseerd in de kennis in plaats van de informatie. Want informatie alleen in plaats van kennis veronderstelt kennis van de gebruiker over de informatie.

Om een vraag zoals: 'Ik ben op zoek naar kennis', te beantwoorden is een andere benadering nodig dan de benadering van de methodiek van een informatiesysteem. Wanneer namelijk een informatiesysteem gebruikt wordt om kennis te genereren, wordt de gebruiker van het informatiesysteem verondersteld inhoudelijk kennis te hebben om de informatie te kunnen gaan gebruiken. Hergebruik van data zal op deze manier alleen binnen het domein kunnen worden hergebruikt waarvoor het gemaakt is. Want alleen de mensen binnen het domein weten waarvoor de data verzameld is en kunnen het hergebruiken. Zelfs als het een niet-domeingebruiker duidelijk is waar bepaalde data te vinden is, zal allerlei kwalitatieve en organisatorische informatie over de data bekend moeten zijn. Zoals wie het gebruiken mag, hoe de nauwkeurigheid is en wat de kwaliteit is.

Metawarehouse

Hoe kan deze kennis die nodig is voor het gebruik van deze data aangewend worden in een generiek systeem, zonder dat de data



Figuur 1 – Kwaliteitsindicatoren Groene Ruimte

Data Groene Ruimte

Omdat voor dit project met name kennis ten aanzien van de groene ruimte onderwerp van onderzoek was, wordt er een case uitgewerkt. Het deelprogramma Monitoring Kwaliteit Groene Ruimte (MKGR) gebruikt in haar communicatie naar en met beleidsmedewerkers een metafoor voor de kwaliteitsindicatoren in de vorm van een zonnetje, zie figuur 1. Als case voor Datamining Groene Ruimte is dit zonnetje beschreven met de thesaurustermen, met als basis de terminologie van het beleid. Door nu deze, geüniformeerde, thesaurustermen te confronteren met de thesaurustermen uit de datasets, kan bepaald worden welke gegevens ontbreken om te komen tot een compleet monitoringsysteem, zonder dat beleidsmedewerkers domeinspecifieke kennis moeten hebben.

Tevens kan er door de aanwezige synoniemrelaties in de thesaurus tussen verschillende termen toch een koppeling worden gevonden. Zo heeft 'groene ruimte' bijvoorbeeld een synoniemrelatie met 'landelijk gebied'. Op deze manier kunnen vraag en aanbod samenkomen, niet gehinderd door abstractie- of detailniveau of door alleen domeinspecifieke kennis.

daarbij fysiek aanwezig is?

Hiervoor is een systeem ontwikkeld dat als werktitel *metawarehouse* heeft meegekregen. Een op internet werkend kennissysteem waarmee zowel geregistreerd als gezocht kan worden. In het systeem is meta-informatie over datasets beschreven aan de hand van een uitgebreide thesaurus. Datasets in het metawarehouse zijn naast gegevens van meetwaarden, ook modellen en personen.

De reden dat gebruikgemaakt wordt van een thesaurus is tweeledig. Aan de ene kant worden de gegevens die in het systeem zijn opgenomen, gekoppeld aan woorden waarvan een definitie bekend is. Zodat iedereen kan begrijpen wat er bedoeld wordt door de definitie van het woord op te zoeken. Aan de andere kant kunnen synoniemrelaties en bredere relaties die aanwezig zijn in een thesaurus gebruikt worden tijdens het zoeken.

Op twee manieren kunnen gegevens worden opgenomen in het systeem. De eerste manier is de registratie van beschreven data, de informatie. Dit kan een beschrijving zijn van een digitaal bodemgebruikbestand, maar het kan ook het curriculum vitae van een wetenschappelijk onderzoeker zijn. De tweede manier van registratie is aan de vraagkant van het systeem. Een vraag van een gebruiker wordt ook als dataset geregistreerd. En hier komt het marketingaspect van het systeem naar voren. Namelijk als een gebruiker op zoek is naar bepaalde kennis over een onderwerp, maar dit niet kan vinden, dan kan dat gebruikt worden als informatie voor een nieuwe dataverzameling. Om in termen van een kennismatrix te spreken, je kunt erachter komen wat je niet hebt, maar waar wel behoefte aan is.

Tot slot

In dit artikel ben ik niet diep ingegaan op de conceptuele kant van de ontwikkelde methodologie en de technische oplossingen die daarvoor gevonden zijn. Wanneer de gelegenheid zich voordoet in een ander themanummer van AI zal ik deze daarvoor benutten. Ideeën, vragen, opmerkingen maar vooral discussie over dit onderwerp is altijd welkom (j.h.vanrijswijk@lei.dlo.nl) en, denk ik, erg belangrijk. @