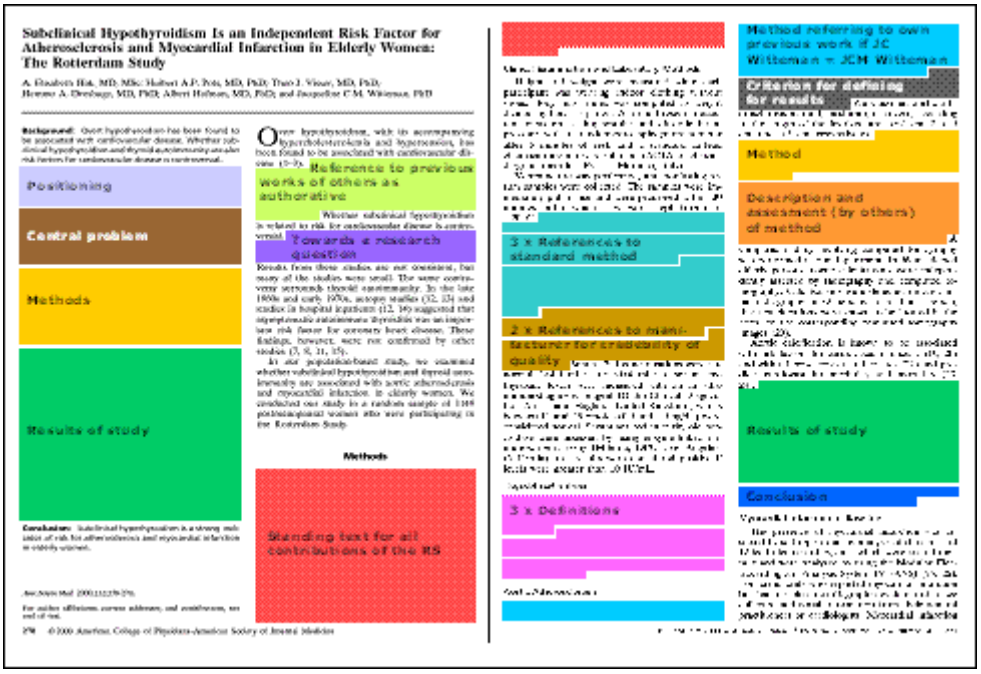
Van thesaurusverrijking naar ontologieën: bouwstenen voor semantisch publiceren

Joost Kircz, Gert Goris, Gusta Drenthe

Definitieve versie: Informatie Professional, februari 2005, jrg.9, no.2. pg

Boodschap en vorm

Wetenschappelijke publicaties zijn erop gericht een gefundeerde mededeling over te dragen naar een geïnteresseerde lezer. Hoewel het niveau van abstractie van de mededeling kan verschillen, kenmerkt de wetenschappelijke rapportage zich meestal toch door een aantal gemeenschappelijke elementen. Het artikel begint vrijwel altijd met een inleiding waar de context van het gerapporteerde in wordt uitgelegd. In dit deel staan verwijzingen naar de literatuur die ten grondslag ligt aan het onderwerp, alsmede naar collega's die gerelateerd werk doen. Dan volgt er een uiteenzetting van wat de auteurs zelf van plan waren en gedaan hebben. Afhankelijk van het onderwerp volgt dan een bespreking van de theorie, de methode, gebruikte apparatuur, de datacollectie en verwerking en afsluitend een vergelijking van de resultaten met de theorie en andere onderzoeksgroepen. Afrondend wordt er een conclusie geformuleerd, waarin de auteur de redelijkheid en het nut van de hele onderneming benadrukt en wegen tot verder werk aangeeft.

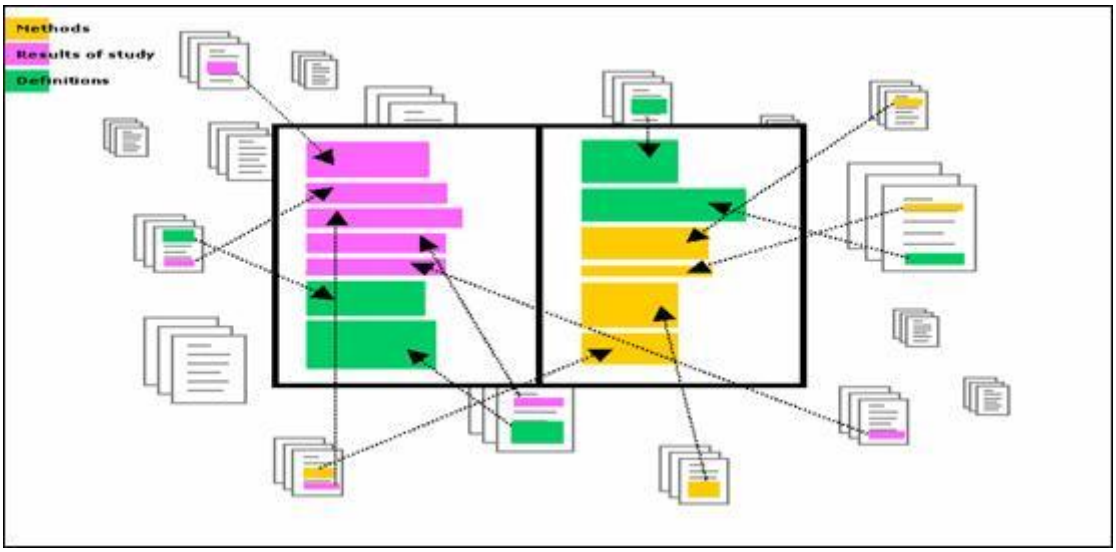


Figuur 1: voorbeeld van de structuur van een artikel

Cognitieve structuur

Dit sjabloon is de laatste halve eeuw standaard geworden, wat zich uitdrukt in de vele boeken over hoe je een wetenschappelijk artikel schrijft. Het interessante van deze vorm is dat het een cognitieve structuur legt over een lineair document. De structurele eenheden staan netjes in een rij en een snelle lezer kan delen gemakkelijk overslaan. In een elektronische omgeving, waar veel meer sprake is van losse eenheden informatie die

naar willekeur en believen gerangschikt kunnen worden, is snel doorbladeren echter niet goed mogelijk, zonder alles eerst op het scherm te projecteren. Het zou beter zijn als we de diverse onderdelen direct zouden kunnen aanspreken.



Figuur 2: Als we onderdelen van publicaties direct kunnen aanspreken is het ook mogelijk ze op een hoger niveau te aggregeren en de gewenste informatie bij elkaar te projecteren of printen

De diverse structurele eenheden moeten dus een naam krijgen die begripsmatig uitstijgt boven triviale duidingen als hoofdstuk 1 sectie 1. We hebben dus te maken met de *semantiek* van de structuur. Als we een inleiding hebben, is het evident dat allerlei gerelateerde literatuur genoemd wordt. Hebben we echter de sectie 'gegevensverwerking' dan kunnen we aannemen dat de daar gemelde literatuur direct te maken heeft met de door de auteur gebruikte methode, zij het in overeenkomstige zin of juist als voorbeeld van een andere methode, of zich hier misschien tegen afzettend. De literatuurverwijzing in kwestie heeft dan een geheel andere waarde voor de lezer dan dezelfde literatuurverwijzing in de inleiding.

Woordsystemen

Dit voorbeeld krijgt nog een extra nadruk als we het probleem uitbreiden naar alle relevante metadata, inclusief trefwoorden. Voor de lezer is de waarde van een verwijzing of een trefwoord vooral in *context* belangrijk.
Keren we terug naar het losse traditionele artikel dan zien we dat de auteur context toevoegt via de literatuurlijst terwijl de redacteur of indexeerder dat doet door toevoeging van trefwoorden uit een gecontroleerde trefwoordenlijst of thesaurus. In de 'ouderwetse' bibliotheekomgeving met onderwerpscatalogi of online bibliografische zoekmachines probeerden we met slimme combinaties van Booleaanse operatoren relevante literatuurverwijzingen te vinden. Als dan het artikel uiteindelijk gevonden was, bleek vaak dat de gebruikte metadata wel het hele document beschreven, maar niet sloegen op de specifieke informatie die gezocht werd. Met 'fulltext' zoekmogelijkheden wordt het er niet beter op. De zoektermen zijn nu willekeurig en ontberen zowel de context van een thesaurus als de context van de structuur van het artikel.
In de wereld van de zoekmachines zien we nu gelukkig ook een tendens om woorden vooral in context te zien en niet meer als losse informatie-eenheden.

De enorme elektronische depots dwingen ons duidelijker vragen te stellen willen we nog een kleine kans hebben op enige 'precision' van de 'recall'. Het thema van contextgebonden thesauri is nu dan ook terug op de agenda, met dien verstande dat we

niet alleen maar met trefwoorden een document willen duiden, maar vooral ook nog de relevante plaats *binnen* het document.
Immers als we zowel de structuur van een document als de zoektermen kunnen benoemen, dan wordt het door segregatie van de enorme hoeveelheid 'hits' mogelijk de zoekende lezer werkelijke relevantie aan te bieden.
Met zoeken in context wordt het mogelijk een verschil te maken in relevantie tussen de zoekterm in bijvoorbeeld de sectie 'reductie meetresultaten' en de sectie 'theoretische bespiegelingen'. Voordat dit echter mogelijk wordt moeten we overgaan de trefwoorden ook daadwerkelijk aan tekstonderdelen in plaats van aan documenten als geheel toe te voegen. We moeten dus een verband leggen tussen de waarde van een trefwoord en de plaats waar dat trefwoord in de tekst van toepassing is.
Dit nu wordt mogelijk als we systemen gaan uitwerken waarbij we trefwoorden zoeken in de context van de tekst zelf en dan meteen deze trefwoorden, in een code, in het tekstdeel zelf vast leggen.

Om dat te kunnen doen, moeten we natuurlijk eerst even flink onze tanden zetten in computerondersteunde methoden voor het gestructureerd toevoegen van trefwoorden aan teksten. Daarmee wordt het belang van goed gestructureerde woordsystemen extra duidelijk. Het heeft geen zin om zo maar trefwoorden toe te voegen.

Thesaurusbouw en thesaurusverrijking

In het kader van ons Rotterdamse onderzoeksproject *R*esearch in *S*emantic *S*cholarly *P*ublishing [i] werken we met bestaande thesauri die we geautomatiseerd verrijken met nieuwe concepten, ontleend aan recente literatuurcorpora. Voor dit onderdeel maken we gebruik van het programma Collexis, zoeksoftware bestaande uit een aantal componenten om in grote hoeveelheden ongestructureerde informatie te zoeken. Indexering van tekst vindt plaats door er concepten uit te filteren, waar op basis van frequentie een gewicht aan toe wordt gekend. De verzameling van concepten vormt de zogenaamde 'fingerprint' van een publicatie. De kwaliteit van de fingerprints verbetert als de conceptuele representatie wordt geformaliseerd via één of meerdere achterliggende thesauri.

Thesauri zijn immers gevalideerde trefwoordlijsten, waarvan het gebruik meer coherente resultaten oplevert dan bij het gebruik van losse woorden.

U kunt zich voorstellen dat het omgekeerde ook mogelijk is: zoeken in literatuurcorpora naar concepten die *niet* met de gebruikte thesauri kunnen worden gematched. Deze concepten, die niet voorkomen in de gebruikte thesauri, zijn kandidaat om als nieuwe termen of concepten in de te verrijken thesaurus te worden opgenomen. Voor dit doel hebben we de Eurovoc thesaurus uit het domein van de economie geselecteerd - we zijn per slot werkzaam bij de Erasmus Universiteit Rotterdam - een vakgebied waarin algemene, relatief platte en qua terminologie 'arme' thesauri beschikbaar zijn of thesauri die slechts een deel van het vakgebied bestrijken en een grotere gedetailleerdheid kennen.

Het is niet moeilijk om snel nieuwe concepten te vinden. We maken daarvan platte lijsten zonder relaties met andere concepten. De nieuwe concepten doen wel mee in de volgende filtering van de literatuur op nieuwe concepten. Op die manier neemt de 'vangst' bij elke nieuwe filtering af.

Om deze platte lijsten te structureren is de volgende stap – en daarmee zijn we op dit moment volop bezig – een instrument te maken dat de omgeving van een kandidaat-term of concept conceptueel weergeeft. Deze omgevingsrepresentatie wordt vergeleken

met de onderliggende thesauri. Op deze wijze worden in de verschillende boomstructuren van de thesauri de plaatsen aangegeven waar het nieuwe concept zou kunnen passen en worden gerelateerde termen uit de thesaurus gepresenteerd. Dit geheel - een nieuw concept, de conceptuele omgevingspresentatie, de mogelijke plaatsen in de hiërarchische structuur van de gebruikte thesauri en de presentatie van gerelateerde termen - wordt gestuurd naar één of meer experts om opname van de nieuwe termen en concepten te valideren. Deze experts valideren dan dus op dezelfde manier waarop traditioneel experts binnen een vakgebied trefwoorden en nomenclatuur valideren.

Na validatie worden de nieuwe concepten gemarkeerd opgenomen en voorzien van een annotatie met datum en reden voor opname. Doordat we aangeven wanneer en waarom een term is toegevoegd kunnen zij altijd – bijvoorbeeld omdat voortschrijdend inzicht andere termen/concepten verlangt of een andere semantische context – worden aangepast of gewijzigd.

Validatieproces

Het validatieproces is essentieel. Thesaurusbouw kan onmogelijk volledig worden geautomatiseerd. We kennen zulke experimenten die hebben geleid tot dramatisch slechte systemen. De menselijke factor blijft cruciaal. Het gaat hierbij immers om het bereiken van consensus tussen vakgenoten. Vandaar dat vanaf het begin van het project samenwerking met wetenschappers in het betreffende vakgebied is gezocht, vooralsnog binnen de Erasmus Universiteit.

Voor het kunnen betrekken van (internationale) experts is het noodzakelijk te weten wie waarmee bezig is. Ook van wetenschappers is het mogelijk een 'fingerprint' maken, via hun literatuurproductie. Op deze wijze ontstaat een kenniskaart van wetenschappers. Het validatieproces zal vorm krijgen via een netwerk van wetenschappers / experts die, vergelijkbaar met het peer review proces bij tijdschriften, meewerken aan de verrijking van thesauri in hun eigen vakgebied.

We stellen ons dat als volgt voor. De experts worden persoonlijk benaderd. Hen wordt het proces van thesaurusbouw uitgelegd waarbij de nadruk ligt op het belang voor de ontwikkeling van het eigen vakgebied. Zij worden uitgenodigd om samen met collega's deel te nemen aan een redactioneel symposium. Daar wordt aan de hand van de praktijk uitgelegd hoe het proces verloopt, hoe er gecommuniceerd wordt en welke de redactionele regels zijn. Tijdens deze sessies wordt door gebruik te maken van de Collexis software naar nieuwe concepten gezocht en vindt discussie plaats over de definiëring en plaatsing van nieuwe concepten. Bestaande artikelen worden gepresenteerd met de al bekende trefwoorden en de door de computer gesuggereerde. Aan het eind van het symposium gaan de experts verrijkt naar huis en wordt het validatieproces virtueel voortgezet. Binnen de Erasmus Universiteit zal een eindredactie worden geïnstalleerd.

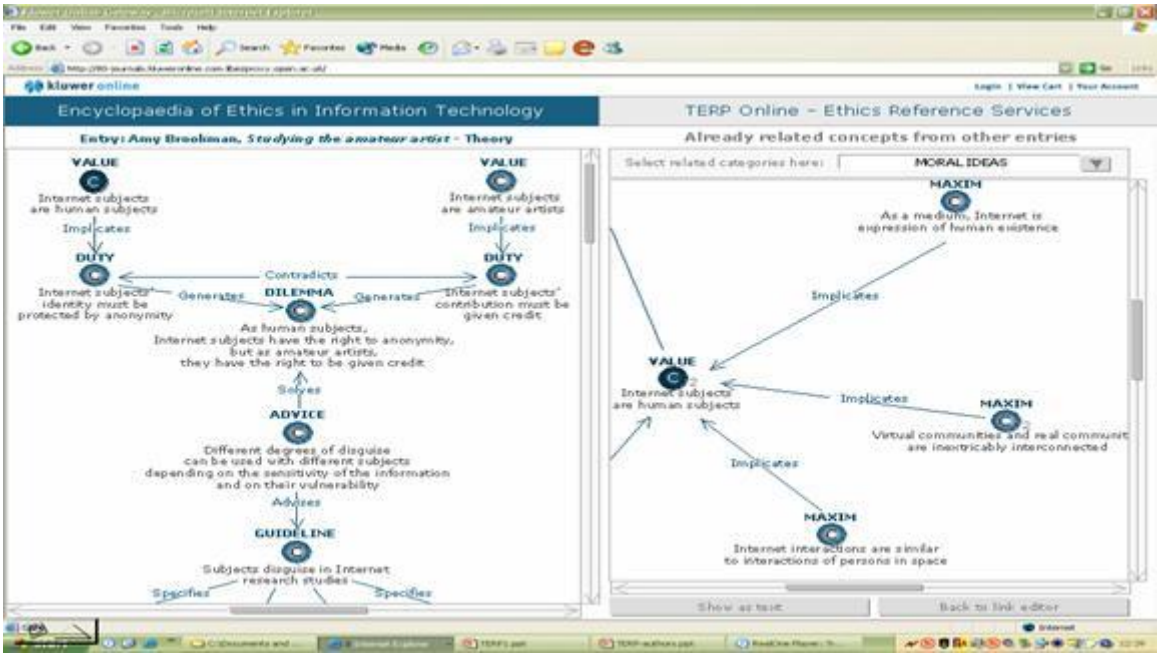
We need more!

Wat we daarmee bereiken is een rijker vocabularium met hiërarchische relaties en gerelateerde termen. Dit is de eerste stap in een verbetering van de zoekresultaten ten opzichte van het zoeken met willekeurige woorden.

De relaties tussen de verschillende concepten zijn echter nog niet benoemd. We hebben ontologieën nodig!

Thesauri bevatten dubbelzinnigheden als gevolg van een beperkte semantische structuur. De relatieset tussen concepten beperkt zich in de regel door 'broader' en 'narrower', 'use' en 'used for' en 'related' die niet worden gespecificeerd door exacte semantiek. Een

thesaurus ontbeert een expliciete en formele betekenisrepresentatie die door machines kan worden 'begrepen'. In tegenstelling tot thesauri wordt in een ontologie kennis conceptueel expliciet gemaakt door het gebruik van een formele taal met duidelijke relaties. Daarvoor dienen we een set van semantische typen te definiëren die de verschillende relatietypen specificeren.



Figuur 3: Voorbeeld van relatietypen, zoals 'contradicts', 'generates', solves', advises', specifies', implicates' [ii]

Met deze semantische structuur ontstaat een ontologie of liever een 'Semantic Knowledge Organization System' (SKOS) waarmee we op verschillende wijzen semantiek in elektronische publicatiedomeinen kunnen aanbrengen. Als dit bereikt is hebben we een eerste stap gezet naar het zoeken van informatie in context.

Onze aandacht richt zich momenteel op twee verschillende toepassingen:

1. associatief zoeken in bestaande lineaire publicaties om achter verbanden te komen die niet expliciet in de literatuur zijn beschreven [iii] .

2. het ontwikkelen van een modulair semantisch publiceermodel en bijbehorende software voor toepassing in wetenschappelijk publiceren en communiceren.

In dit laatste onderzoek willen we nieuwe modellen ontwikkelen die de integratie van tekst, beeld, geluid, simulaties, ruwe data, software en het leggen van betekenisvolle verbanden daartussen toestaat.

Vorm en cognitieve structuur

Een publicatie kan worden opgesplitst in kleinere, cognitief zelfstandige informatie-eenheden, modules genaamd. Deze modules kunnen zowel talig als niet-talig zijn en kunnen in een digitale omgeving op een veelzijdige en betekenisvolle manier met elkaar worden verbonden. Op die manier ontstaat een publicitaire eenheid of document met een eigen identiteit: authentiek, vindbaar en citeerbaar, maar dynamisch in zijn samenstelling. Een dergelijk model maakt het wetenschappers mogelijk 'semantische links' aan te leggen tussen fragmenten of onderdelen van elke publicatie. Hiertoe wordt

een publiceerprogramma ontwikkeld dat auteurs in staat stelt betekenis te hechten aan een verwijzing naar gerelateerde informatie. Het programma ondersteunt de auteur door duidelijke aanwijzingen te geven van de keuzemogelijkheden en linkbetekenissen.

Daarvoor moeten ontologieën ontwikkeld worden via een classificatie van relatietypen die tussen deze modules kunnen voorkomen. De classificatie van modules en hun onderlinge relaties resulteert in een logisch systeem van metadata. Zo'n hyperlinktypologie maakt de interne linkstructuur van een publicatie op zichzelf informatiedragend en verruimt de informatie- en argumentatieruimte, waarbinnen de wetenschapper verschillende sporen of routes kan volgen. Op deze manier ontstaat een publicatieomgeving, waarin cognitieve structuren aan de oppervlakte komen door middel van verbanden tussen wetenschappelijke concepten, zoals ideeën, hypothesen, bewijsvoeringen, weerleggingen, interpretaties, commentaren.

Per discipline zijn er grote verschillen in communicatiecultuur, taal en semantiek [iv] . De discipline-ontologieën geven de formele en semantische structuur aan het publiceermodel. Het is naar onze mening aan te bevelen om thesauri / ontologieën in verschillende deeldisciplines naast elkaar te laten bestaan en een metathesaurus te ontwikkelen die de dwarsverbanden benoemt.

De link tussen verrijkte thesauri, ontologieën en semantisch publiceren wordt nu steeds duidelijker.

==========

Drs. *Gert Goris* is adjunct-bibliothecaris en hoofd documentaire informatie bij de Universiteitsbibliotheek van de Erasmus Universiteit Rotterdam (EUR). Hij is project-manager van het project Research in Semantic Scholarly Publishing (RSSP) van de EUR
Dr. *Joost Kircz* heeft na een opleiding in de natuurwetenschappen jarenlang als wetenschappelijk uitgever gewerkt. Tegelijkertijd heeft hij academische onderzoeksprojecten op het gebied van elektronische kennisoverdracht uitgevoerd. Op dit moment is hij actief als zelfstandig onderzoeker en parttime onderzoeker bij het RSSP-project.
Drs. *Gusta Drenthe* is vakreferent sociale wetenschappen bij de Universiteitsbibliotheek van de EUR en betrokken bij het sociaal-wetenschappelijke deel van het RSSP-project.

Verdere oriëntatie

Voorbeelden van verrijkte thesauri:

Dagobert Soergel, Boris Lauser, Anita Liang, Frehiwot Fisseha, Johannes Keizer and Stephen Katz; "Reengineering thesauri for new applications: the AGROVOC example". In: Journal of digital information, Vol 4 issue 4; article No. 257, 2004-03-17. Zie: http://jodi.ecs.soton.acuk/Articles/v04/i04/Soergel/
Mathew Weaver, Timothy Tolle, Marianne Lykke Nielsen and Lois Delcambre Knowledge Structures for a Domain-Specific Digital Library for Natural Resource Managers.

Voorbeeld van een uitwerking van 'semantic scholarly publishing':
Victoria Uren, Simon Buckingham Shum, Gangmin Li, John Domingue, Enrico Motta; Scholarly Publishing and Argument in Hyperspace. Zie: http://kmi.open.ac.uk/publications/papers/kmi-tr-127.pdf(april 2003

Voorbeelden van toepassingen van woordsystemen voor het semantisch web:
http://www.delos.info/eventlist/LUB1.html
KOS: Simpel Knowledge Organization Systems for the Semantic Web. Zie bijvoorbeeld:

http://www.delos.info/eventlist/LUB1/Alistair_Miles/Lund_presentation.ppt/
http://www2.db.dk/nkos-workshop/

Noten

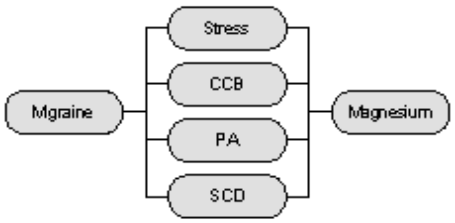
[i] Zie http://rssp.org. Dit onderzoeksproject vindt plaats aan de universiteitsbibliotheek van de Erasmus Universiteit Rotterdam. Het onderdeel thesaurusverrijking wordt in nauwe samenwerking uitgevoerd met: de capaciteitsgroep informatica van de Faculteit Economische Wetenschappen van de EUR: dr. Jan van den Berg en de doctoraalstudent Jaron Boheemen, in het kader van het VICORE-project (zie ook www.vicore.nl) en de afdeling medische informatica van het Erasmus MC: dr. Barend Mons en dr. Erik van Mulligen (bouwers Collexis-software).

[ii]Dit voorbeeld is gehaald uit een ontwerp dat in het kader van RSSP in samenwerking met het Knowledge Media Institute (UK) is gemaakt voor een Online Encyclopedia of Applied Ethics.

[iii] Een inmiddels klassiek voorbeeld van het doel van tekst data mining is het onderzoek naar het verband tussen migraine en magnesium. In dit onderzoek werden de titels van medische wetenschappelijke artikelen onderzocht, hetgeen het volgende opleverde:

· CCB (calcium channel blockers) beperken migraine
· SCD (spreading cortical depression) is geassocieerd met migraine
· Een hoog magnesiumgehalte bestrijdt SCD
·Migraine patiënten vertonen een hoge PA (platelet aggregability)
· Magnesium onderdrukt de PA
·Stress leidt tot lager magnesium gehalte
· Migraine wordt geassocieerd met stress (spanning)

Hoewel in geen van deze teksten afzonderlijk de concepten migraine en magnesium gerelateerd worden, kan op basis van tekst data mining de volgende relatiestructuur worden afgeleid.



Figuur 4: Grafische weergave van de relatie tussen de concepten migraine en magnesium met de gevonden tussenliggende concepten. Een gebruiker kan met dit systeem zonder dat hij/zij zelf alle artikelen hoeft te lezen gebruik maken van de automatisch gevonden relaties.

[iv] Algemene achtergrond over de invloed van informatietechnologie op het wetenschapsbedrijf:Michael Nentwich. Cyberscience: research in the age of the Internet. Vienna Austrian of Science Press, 2003. 490 + 79 p. ISBN: 3-7001-3188-7