Stop alles in het Datawarehouse en haal het er met Google weer uit

# Meaningful tags

Andries Bottema

**Op een nieuwjaarsborrel werd de vraag gesteld waarom Google niet even in een datawarehouse kan zoeken naar de contracten waar het meeste winst op wordt gemaakt. Antwoorden varieerden enorm: van "daar heb je nu boekhouders voor", en "jij snapt niets van de techniek van Datawarehouses", tot "daar heb je weer zo'n salongeleerde in BI die nog nooit een statement heeft geprogrammeerd".**

Dit is allemaal waar misschien, maar de vraag bleef wel hangen. Natuurlijk hadden we het klassiek op kunnen lossen. Definities vaststellen, constateren dat we de contracten met code 1 - 10 (bij wijze van voorbeeld) in de database moeten stoppen en de query van winst loslaten op het grootboek. Maar vaak zitten er achter relatief onschuldige (domme) vragen enkele echt grote problemen verborgen. Het gaat hier allereerst om de eenvoud waarmee we kunnen zoeken naar informatie (en eenvoudig leren zoeken hebben we te danken aan Google, alhoewel dat niet synoniem is met vinden). En het gaat hier ook om informatie die niet volgens de traditie van gestructureerd coderen met cijfers is vastgelegd.

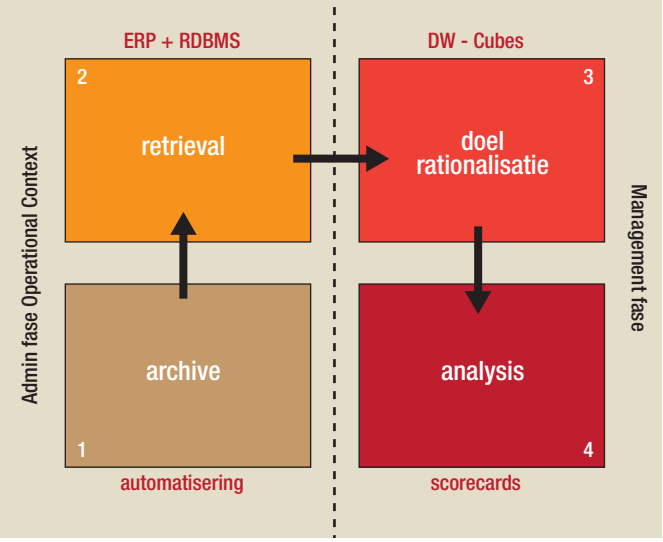
In dit artikel wordt een aantal ingrediënten besproken die er voor kunnen zorgen dat we straks Business Intelligence los kunnen laten op veel meer dan cijfers en tabelletjes. Dat wat we nu ongestructureerde data noemen (dit artikel, uw e-mail, een patiëntendossier), kan de bron worden voor interessante analyses. Wellicht dat we het op een slimme manier in een Datawarehouse (DW), of een andere manier van opslag van informatie kunnen stoppen. En we hopen dat we het er vervolgens gemakkelijk uitkrijgen. We zullen de ontstaansgeschiedenis van Business Intelligence bespreken. Enerzijds kijken we naar het structureren (lees: betekenis toekennen aan data) van cijfers vanuit de ERP-traditie. Anderzijds kijken we vanuit de wereld van documenten hoe we informatie destilleren uit op het oog ongestructureerde data. Een belangrijk begrip als UIMA passeert de revue als architectuur om dat te doen. We filosoferen vervolgens even over de opslag van al die verkregen wijsheid: Een soort Inmon visie op DW 2.0: alles zit er in? Of een Nicholls BI 2.0 visie: geen DW, maar een real-time processor voor alle informatie? Of we al dan niet een DW gebruiken bepaalt de techniek, en is dus het minst relevant aan de discussie. Tenslotte sluiten we af met de vraag hoe we, als dat DW eenmaal gevuld is met relevante data, de informatie er weer op een natuurlijke manier uithalen.

## BI markeert de gebruikswaarde van informatie

In een zogenaamd 'expert-panel' van BI-kenners is de kortste definitie ontstaan van Business Intelligence: het inzetten van informatie als productiefactor. Oftewel die informatie kunnen presenteren waar mensen ook iets mee kunnen bereiken en verbeteren: een andere productmix voeren; van andere toe-

## Collexis en Autonomy: begrijpen wat er staat

Hoe vind je een concept in een document? Er zijn meerdere manieren om de tekst te classificeren. Een daarvan is door maar veel documenten met pure rekenkracht met elkaar te vergelijken en te correleren en zo min of meer een concept te achterhalen. Denk bijvoorbeeld aan het steeds terugkeren van bepaalde termen in bepaalde volgorde met een bepaalde waarschijnlijkheid. Vervolgens kun je zo'n concept vastleggen en uitspraken doen over andere documenten. Ook kan men een thesaurus inzetten: SOA komt voor in een tweetal betekenissen in teksten. Met het begrijpen van een concept heeft men de betekenis van een tekst te pakken. Autonomy gebruikt heel veel technieken om de teksten (of andere informatiedragers) te analyseren op de content en er vervolgens zo'n concept voor te maken. Computer based. Collexis daarentegen richt zich op veel wetenschappelijke teksten en laat zo'n framework eerst door experts samenstellen (bijvoorbeeld microbiology) en haalt er vervolgens teksten doorheen, om ze van een 'fingerprint' te voorzien: geclassificeerde informatie waar je wat mee kan.




**Afbeelding 1:** Van cijfer tot scorecard.

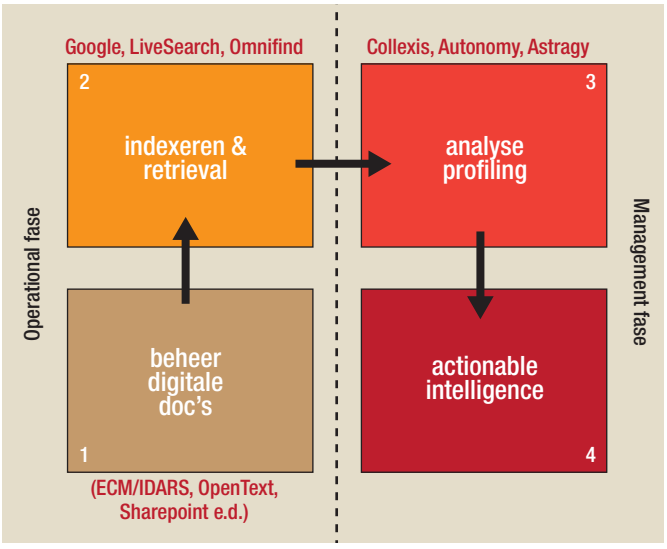
leveranciers spullen betrekken; voorspellen wanneer er meer ijs wordt verkocht.
In de dagelijkse praktijk levert een gemiddelde database met bedrijfsgegevens die informatie nauwelijks. Er moet iets mee gebeuren: transformeren naar zinvolle context (omzet/jaar, leveranciersscores etcetera) en vervolgens zo opslaan dat er effectief mee geanalyseerd kan worden. En dan moeten we nog actief aan de slag met analyseren om uiteindelijk richting te kunnen geven aan handelen (toeleverancier eruit, meer magazijnen in plaats van vrachtwagens).
Eigenlijk zijn we de afgelopen twintig jaar erg druk geweest met het coderen en structureren van allerlei gebeurtenissen die als getallen opgeslagen worden. En dat is allemaal begonnen op veel verschillende afdelingen. De sigarendoos werd te klein (de achterkant) en in spreadsheets en kleine applicaties werden de leveranciers en artikelen van een nummer voorzien (Archive-fase, zie afbeelding 1). Met deze acties hadden we in ieder geval een soort archief: ergens lagen de gegevens opgeslagen, alhoewel niet eenduidig. Vervolgens werd met ERP deze eilandautomatisering meer gerationaliseerd en geüniformeerd. Dikke coderingsboeken zijn onderdeel van een ERP-implementatie. Het doel was om uniform, relationeel en genormaliseerd data vast te leggen. Overigens niet omdat we dan beter begrepen wat er in de database stond, maar omdat de techniek dat zo belangrijk vond ("als uw receptioniste ziek is mag maar één iemand dat weten en haar ziekte heet dan 0").
En het gaat meestal om twee redenen ook nog fout: vaak een te rigide classificering ("we kunnen u geen korting geven want het systeem laat dat niet toe, maar ik schrijf het wel even op een apart papiertje") en het probleem dat we de informatie er niet eenvoudig (*technisch en logisch*) uit kunnen halen. Hoe vaak werd er door een technisch onbenul geen systeem plat gelegd, of was de datadictionary-informatie onleesbaar voor een manager? Vervolgens was het de beurt aan een Datawarehouse die de belofte van Management Informatie moest waarmaken. Met transformatie en Single Version of the Truth (winst = Earnings Before Income Tax, Depreciation and Amortization) worden de data uit de ERP (of meerdere bronsystemen) gehaald, en gepresenteerd. Eigenlijk weer een coderingsslag van gegevens naar een nog hogere graad van classificatie. Of ook wel een correctie van de ERP-mislukking in sommige gevallen.

Daarnaast wordt het ook nog duidelijk dat een analyse-tool niet zelf de waarheid boven tafel haalt, maar dat je met enige kennis van de business (met welk doel zoek je informatie) moet gaan zoeken tot je iets vindt waarmee je iets kunt. Dit is de BI-fase waarin we ons afvragen of we met al die vergaarde informatie tot analyse kunnen komen. Deze fase is de ultieme droom van een automatiseerder: Actionable Intelligence, bijvoorbeeld het weergeven van scorecards van leveranciers om snel te kunnen sturen als dat nodig is.

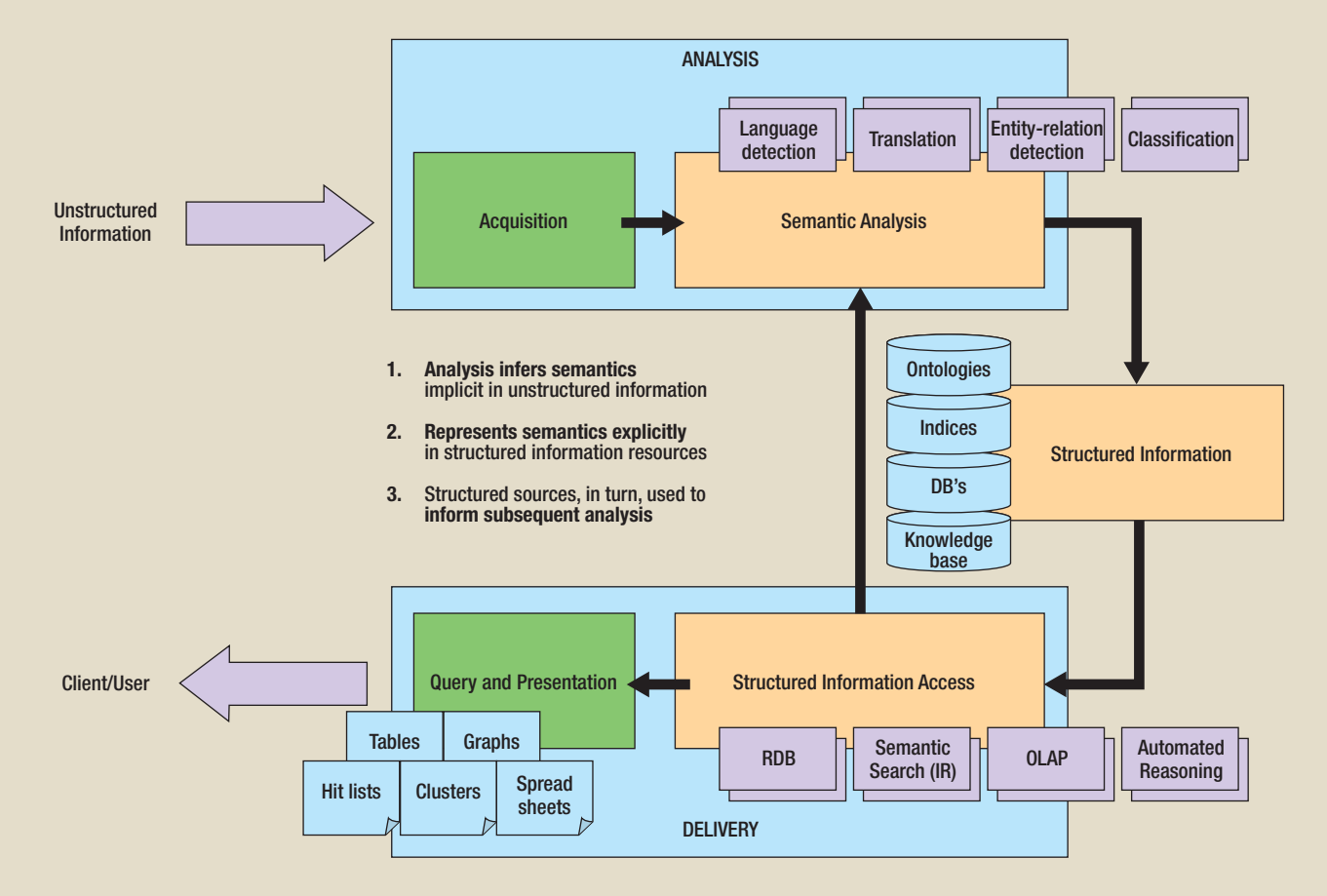
Het is duidelijk dat alle fases staan en vallen bij die codering van de data: de codering die er voor zorgt dat een structuur ontstaat waar men iets mee kan. Daar is minstens tien jaar ervaring voor nodig overigens. Maar die codering is ondanks alle metadata, business rules en dimensionele modellen overigens niet meer echt te achterhalen voor een gebruiker. Het vragen naar winst op de pyjama's (om maar een dwarsstraat te noemen), waarvan alle coderingen en definities ooit wel eens zijn vastgelegd, kun je wel vergeten. Daarbij ligt in het klassieke BI-paradigma alleen maar de informatie van de cijfertjes opgeslagen, maar is er nog een grote categorie met informatiedragers die nog niet gestructureerd zijn.

## Wat staat er in een tekst?

Een tekst (of een plaatje, geluid, filmpje) kan net zoveel informatie als alle cijfertjes in een Datawarehouse bevatten. Hoe goed de voorspelling van cashflow in een onderneming ook uit het DW komt, een brief van de bank spreekt vaak boekdelen.




**Afbeelding 2:** Van tekst tot informatie.




**Afbeelding 3:** Unstructured Information Management Architecture (IBM/Apache).

En toch hebben we ons nog niet echt ingespannen om iets te verzinnen dat de computer ook iets met zo'n brief kan. Maar dat lijkt te veranderen: de ontwikkeling van het omgaan met ongestructureerde informatie lijkt dezelfde coderingsfases te ondergaan als die we hiervoor schetsten rondom de cijfertjes BI.

> Zo zijn we ongestructureerde informatie eerst maar eens per eiland gaan opslaan

Zo zijn we de zogenaamd ongestructureerde informatie eerst maar eens per eiland gaan opslaan (zie afbeelding 2: eerste kwadrant). Dan weten we waar het staat. Maar zoals een bekend vliegtuigonderdelenbouwer onlangs opmerkte "het terugvinden blijft een uitdaging". Maar daar waar er regels waren voor het opslaan van de tabelletjes in het RDBMS-tijdperk proberen we het nu voor ongestructureerde data (bijvoorbeeld een document) met steekwoorden. En als dat mislukt doen we een totale indexering van alle woorden. Search Technology is een hot topic. Het geeft ons de illusie van terugvinden. Google heeft in ieder geval die illusie van terugvinden gepopulariseerd. Maar feitelijk kun je het niveau van indexeren en retrieval van teksten in dit tweede kwadrant nog nauwelijks vergelijken met de opslag van ERP-gegevens in een RDBMS met een data-dictionary. Full text indexing die in 'shards' worden gezet en waarmee genoeg hardware snel een match naar zoekwoorden gaat maken levert inderdaad gemiddeld 190.000 hits op. Voor het betere classificeren moeten we naar het derde kwadrant: analyseren en profiling – ik wil informatie zo gaan structureren (modelleren naar dimensies wellicht) dat ik er mee kan analyseren. Eigenlijk komen we daar in de Business Intelligence-fase. We 'ETL'en' van het tweede kwadrant naar het derde kwadrant: Wat staat er in de tekst? Waar gaat het over? We komen tussen het tweede en derde kwadrant eigenlijk op een basisvraag terug: hoe structureren we informatie die ongestructureerd is? Oftewel wat staat er in een tekst? Pas dan kunnen we naar Actionable Intelligence.

## Het structureren van ongestructureerde informatie

Zoals metadata voor gecodeerde cijfers de context aangeven hoe we de data moeten interpreteren, zo lijkt tagging dat voor documenten en plaatjes te zijn. Eigenlijk is tagging de gepopulariseerde versie van het indexeren van informatie: voorzien van

een key waarop we zaken kunnen terugvinden. Maar daar houdt het bij tekst niet op. Want er is meer uit te halen. Neem bijvoorbeeld Semantiek: wat betekent een zin. Populaire zoekmachines proberen betekenis te achterhalen door te kijken naar 'proximity' ("pech en automerk X staan wel dicht naast elkaar") of 'parametric search' ("als dit in de titel staat en drie keer in de eerste paragraaf, dan moet het over koffie gaan"). Maar naarmate je meer intelligent wilt coderen (wat betekent een zin nu echt), ontkom je niet aan een heel scala aan tools om de content te bestuderen. Zo kunnen er met behulp van Neural Networks-technieken concepten getraind worden: we voeren het Network eerst met een paar honderd teksten over SOA en XML, zodat vervolgens nieuwe teksten scherp geclassificeerd kunnen worden. En classificatie is het eerste begrip van wat er staat en vormt de basis voor analyse. In een voorbeeld van tekstanalyse van medische dossiers met SAS Text Miner wordt een kwantitatieve representatie van de tekst gegenereerd, waaruit met behulp van traditionele data mining-technieken relevante variabelen kunnen worden geëxtraheerd. Dat maakt het mogelijk patronen te ontdekken in behandelingsmethode en het behandelingsresultaat (University of Louisville). En daarmee zitten we in het vierde kwadrant: het equivalent van de Scorecard die vertelt wat er goed en fout is.
Maar naast het door de computer laten 'berekenen' van wat er in een tekst staat, kunnen er nog meer metadata worden toegevoegd die de latere analyse vergemakkelijken. Kortgeleden (januari 2008) sprak de CEO van REUTERS Devin Wenig over de noodzaak dat ongestructureerde informatie van meer informatie moet worden voorzien, "Add hooks", zodat men er meer mee kan (bijvoorbeeld in Business Intelligence). Hij deed die uitspraak bij het vrijgeven van de Calais API: een van de kroonjuwelen van semantische analyse van REUTERS. Maar naast semantische techniek om te structureren stelde Wenig dat we ook zeker niet moeten vergeten dat ongestructureerde informatie ook door mensen van betekenis wordt voorzien. Social networking om betekenis te verlenen aan documenten staat nog maar aan het begin! Ook gelooft Wenig dat uitgevers meer moeten doen om de informatie die ze publiceren van zowel 'meaningful tags' te voorzien, alswel van kenmerken van de beoogde doelgroep. En daarmee is personalisatie als element om ongestructureerde informatie te classificeren een megafactor van betekenis: wanneer we teksten structureren met in het achterhoofd de specifieke doelgroep, dan krijgt informatie een actualiteitswaarde. Met die parameters is Knipsel Info Services (Almere) al enige tijd op een aardige manier aan het stoeien: alle informatie die in dag- en weekbladen staat in Nederland wordt 'ge-meta-tagt' door tientallen professionele lezers die de doelgroep kennen en betekenis kunnen verlenen aan het artikel op basis van die doelgroep(en).

## Het UIMA initiatief

IBM is jaren bezig geweest om al het voorgaande te vertalen naar een architectuur. De basisvraag was hoe we iets zinnigs kunnen doen met al die ongestructureerde informatie. Sterker nog, als we meer inzicht zouden hebben in de informatie op internet, zouden we feitelijk voorspellingen kunnen doen over hoe lang men (nog) leeft, of men volgend jaar qua bedrijf nog groeit etcetera: "All I really need to know I learned from Google" (Oren Etzioni, Turing Center).

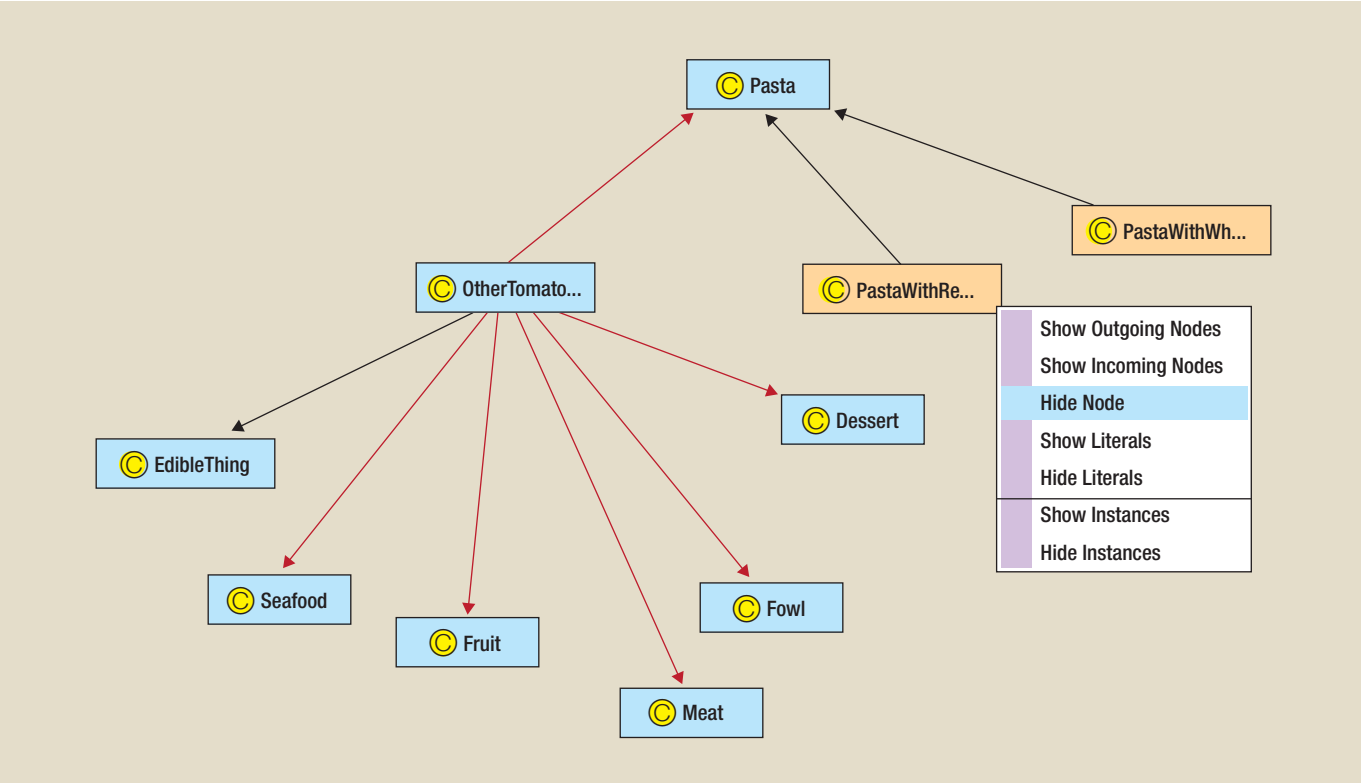
IBM heeft een architectuur neergezet die er voor zorgt dat ongestructureerde informatie gestructureerd en (dus) bruikbaar wordt. Deze 'Unstructured Information Management Architecture' (UIMA) omschrijft op welke manier verschillende technologieën samenwerken om input in een omvattend systeem van analyse te verwerken en waarbij de output in *back-end information processing machines* kan worden gebruikt.

> Classificatie is het eerste begrip van wat er staat en vormt de basis voor analyse

We zien in afbeelding 3 terug dat met name het structureren van data alle mogelijke technieken toelaat. De 'loop' van Analysis naar Structured Information naar Structured Information Access is zeker geen rocket-science (het is een soort ETL-tool), maar het aardige is dat IBM het framework aan Apache als open-source Incubator initiatief heeft gegeven, met als doel om talloze tools toe te voegen die op het middenstuk van UIMA tekst, audio en video kunnen analyseren tot gestructureerde content.
Het Extract-deel (zoals in ETL) is waar de informatie klaargezet wordt voor behandeling. Daarvoor is er semantische analyse nodig om de eerste structuur toe te kennen. Een soort data-dictionary wordt losgelaten op de (bijvoorbeeld) tekst, waarna taaldetectie, vertaling, relaties en het toekennen van types gebeurt (alle lidwoorden eruit, zoeken naar onderwerp etcetera).

## Calais API voor het Semantische Web

Het structureren van de ongestructureerde informatie is mogelijk door een nieuwe technology tool die REUTERS heeft vrijgegeven: de Calais API. Deze technologie maakt het mogelijk om een aantal duizenden documenten door de wringer te halen en de uitkomst te presenteren in RDF-Grafity: een grafische weergave van de concepten, samenhang en logica (in het voorbeeld: voedsel-Pasta), zoals te zien in afbeelding 4. De grafische structuur zou de definitie van 'winst' op kunnen leveren, als men er een aantal jaarrapporten door heen walst.




**Afbeelding 4:** De Calais API voor het Semantische Web: begrijpen wat er in al die webpagina's staat.

Het hart, IBM noemt dit CAS (Common Analysis Structure), stelt verschillende technologieën in staat om content door te geven en verder te analyseren (T. Götz, 2004). Feitelijk een ingewikkelde Transform en Load ineen. Hier vindt bijvoorbeeld de koppeling van een artikel over rechtspraak naar de thesaurus van juridische zaken plaats, en voor artsen naar een medische ontologie. Voor illustraties van dit fenomeen: Clearforest Corporation (een UIMA compliant analyse tool) heeft daar een aantal leuke voorbeelden van die laten zien hoe zij teksten analyseert en classificeert naar een concept. En biedt dit aan als een software service.

Maar het venijn zit in de staart. UIMA is nog nauwelijks in staat geweest om voor te schrijven hoe we al die gestructureerde informatie gaan opslaan. In een relationele database zouden we kunnen veronderstellen, of in een Datawarehouse 2.0. Het 'delivery' deel staat binnen UIMA slechts toe alleen de resultaten van de classificatie op de een of andere manier te tonen. En dat betekent dat we ergens halverwege de metadata laten liggen. Als we met veel moeite een paar duizend e-mails van klagende klanten hebben gecodeerd (semantisch geanalyseerd, concept gevonden ("neem bijvoorbeeld klacht = klant + het noemen van minimaal 2 x niet betalen + 1 x stemmingswoord (boos, verdrietig etcetera) + 2 x merknaam"), dan is als uitkomst wel aardig om te weten welke klaagmails het ergste zijn, hoeveel en wanneer we deze het meest tegenkomen, maar dan willen we ook die metadata (het concept 'klacht') blijvend kunnen gebruiken als een definitie, om later eens aan het DW te vragen in hoeverre klachten invloed hebben op betaalgedrag. UIMA is een aardige aanzet, maar moet nog een theorie ontwikkelen rondom de opslag. In traditionele BI-bewoordingen: of we relationeel of dimensioneel willen modelleren en hoe we de metadata gaan opslaan.

## BI op veel soorten gestructureerde informatie

Er zijn veel software-tools die in staat zijn om tekst, plaatjes, spraak te kunnen analyseren. Bespraken we hiervoor met name text analyse, er zijn ook tools die andere media kunnen analyseren. Neem Utopy.com: het analyseren van call-center gesprekken en daar een correlatie in structuur (toonhoogte, monoloog versus dialoog en dergelijke) en succes (verkoop van bijvoorbeeld een hypotheek) van het gesprek in kunnen ontdekken. De echte BI'er wil straks al deze verschillende structuren aan elkaar koppelen op zoek naar nieuwe inzichten. Er moeten dus kruisverbanden worden gelegd: relaties naar andere data uit andere bronnen; informatie koppelen; opnieuw dimensioneren. We zijn hier al ver voorbij wat UIMA beoogt: zij richt zich slechts op het structureren van ongestructureerde data. Stel nu ook eens dat we die achterliggende structuur gaan opslaan en we ook nog de metadata kunnen gaan gebruiken om de data er weer vraagsgewijs uit te halen, dan lijkt de Google query op de informatie ineens niet meer zo ver weg. Zodat we kunnen vragen: "Waarom maken we het meeste winst op contracten die minder woorden besteden aan de juridische voorwaarden".

**Andries Bottema** (a.bottema@pl.hanze.nl) is Lector Business Intelligence aan de Hanzehogeschool Groningen.