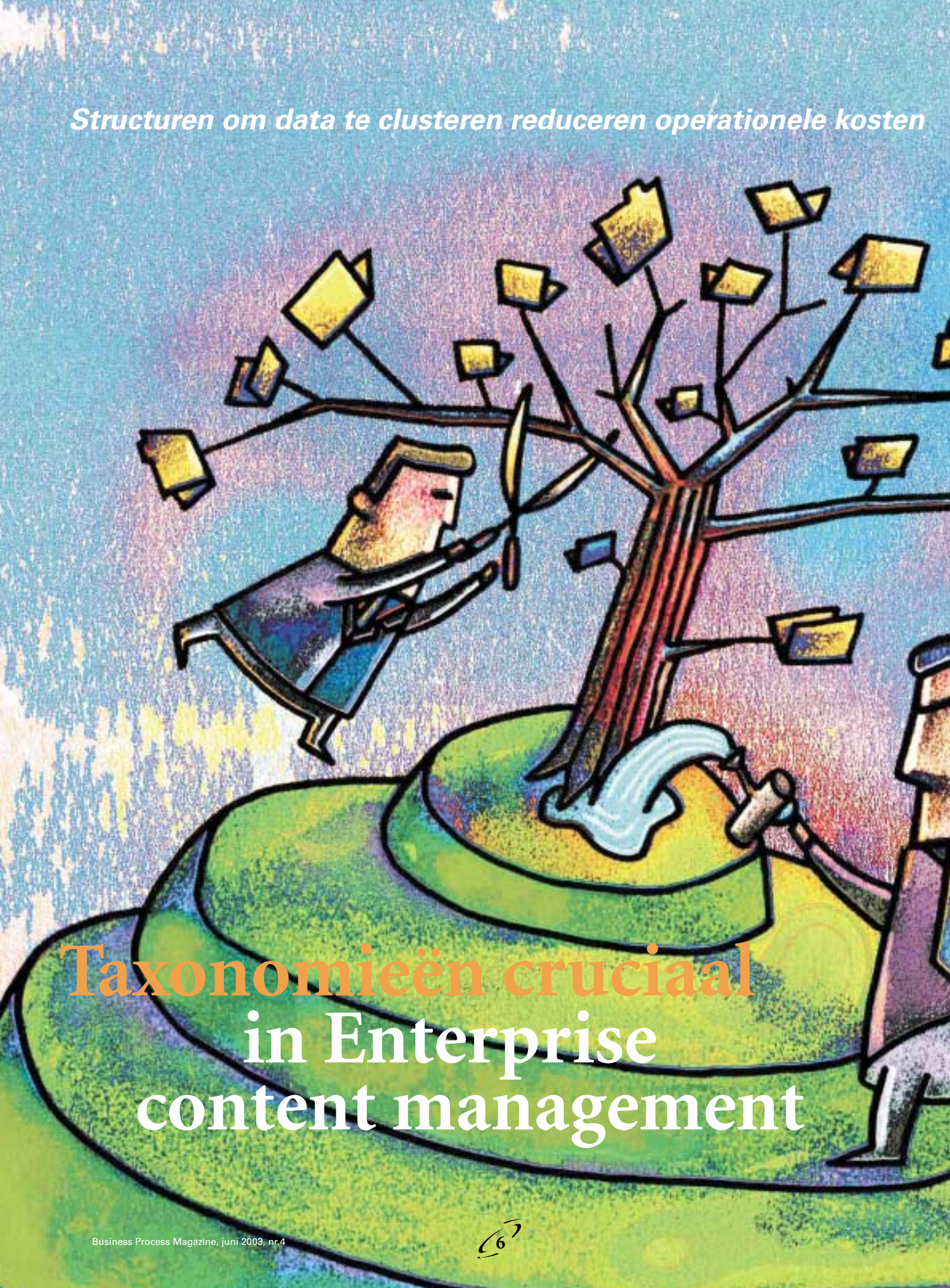
Structuren om data te clusteren reduceren operationele kosten



Taxonomieën cruciaal in Enterprise content management

Illustratie: Leon van Leeuwen

Enterprise content management-systemen zijn uiterst complex. Er is altijd sprake van meerdere technologieën en meerdere applicaties en er is een veelheid aan types content in het geding. Een goede taxonomie is een essentieel onderdeel van het ECM-concept en is het middel om content te ontsluiten en deze voor gebruikers toegankelijk en hervindbaar te maken.

Leo Meerman

Een taxonomie kunnen we omschrijven als een structuur die het mogelijk maakt om content over personen, organisaties, gebeurtenissen en dingen te clusteren in (hiërarchische) groepen, zodat ze gemakkelijk te identificeren, te bestuderen en terug te vinden zijn. Het ECM-concept zoals hier gepresenteerd, is tot haar essentie teruggebracht tot de deelconcepten *content management, content structuur en content access* als drie peilers. Deze drie deelconcepten worden kort besproken in het eerste deel van dit artikel.

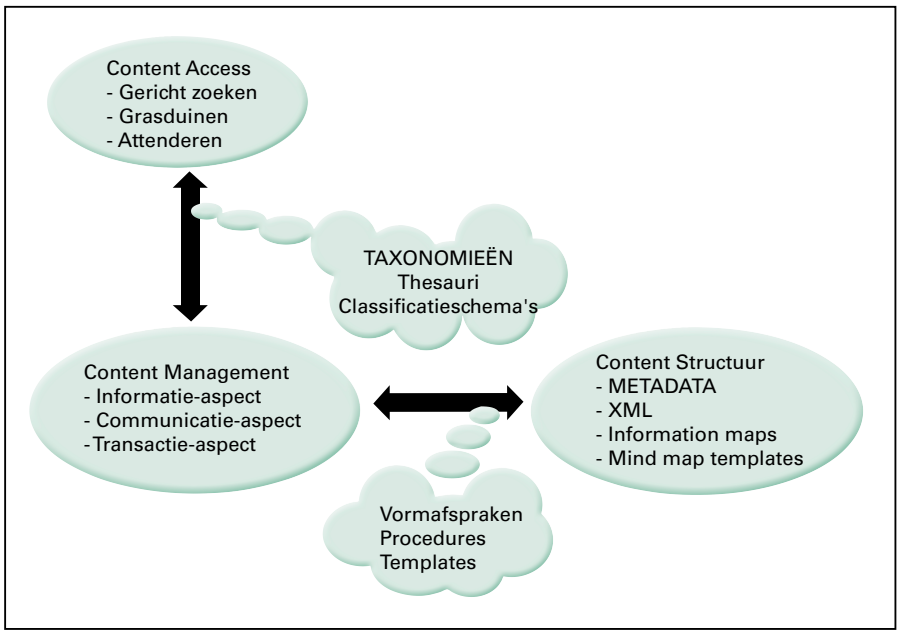
Het tweede deel is geheel gewijd aan de taxonomie. Aan de orde komen de essentiële onderdelen van de taxonomie, de stappen die genomen moeten worden in het ontwerp en een korte aanzet voor het uitvoeren van een business case.

Volledig in de sfeer van dit themanummer zal voor alle vormen van data en documenten in al hun verschijningsvormen, van 'ASCII' tot en met audio en video, het woord content gebruikt worden.

Veel schematische voorstellingen van ECM-systemen blinken uit door een veelheid aan onderdelen met grote aantallen relaties. In het kader van dit artikel volstaat het om een heel globaal, logisch model te hanteren, om de plaats van de taxonomie in een ECM-systeem aan te geven. De genoemde 'meerdere technologieën en meerdere applicaties' blijven buiten beschouwing. Afbeelding 1 geeft het ECM-concept weer met daarin de positie van de taxonomie.

Content Management

Bij content management (CM) is een drietal aspecten van belang: het informatie-aspect, het communicatie-aspect en het transactie-aspect. Het informatie-aspect heeft alles te maken met het verwerven en publiceren van de gegenereerde content. De content zal gescreend worden op de kwaliteitseisen die opgesteld zijn. Ook checks op afspraken over vertrouwelijkheid en verantwoordelijkheden zijn hier van belang. Het communicatie-aspect heeft betrekking op het faciliteren van kennis delen en samenwerken, zowel interpersoonlijk als met behulp van e-faciliteiten. Het transactie-aspect tenslotte, heeft betrekking op het goed functioneren van workflows en de daarbij betrokken content, waardoor bedrijfsprocessen op een goede manier kunnen verlopen.



Afbeelding 1. Het ECM totaalconcept.

Met betrekking tot CM zijn twee overwegingen van groot belang. Ten eerste dat de business de eigenaar van de content is en ICT is eigenaar van de infrastructuur, want business mensen zijn de enigen die de kennis bezitten om de content te onderhouden. Als tweede overweging; het implementeren van content access moet onafhankelijk gebeuren van content management, omdat een gebruiker veelal behoefte zal hebben aan content uit meerdere bronnen.

De content structuur (CS) zoals hier bedoeld, heeft vooral betrekking op het benoemen van metadata van de relevante contenttypes. Het vaststellen van de metadata vereist een degelijke analyse van de in gebruik zijnde content in de organisatie. Voor web-publicaties is het een goede start om uit te gaan van de Dublin Core Metadata Element Set.

Op basis van de uitkomsten van zo'n analyse kunnen dan templates en/of e-forms gebouwd worden, liefst met de mogelijkheid om *constraints* aan te brengen om bijvoorbeeld foute invoer te voorkomen of checks op gespecificeerde velden in databases te doen.

Bij CS is, in technische termen gesproken, dikwijls slechts sprake van een platte sequentiële structuur van content-items. We hoeven niet te denken in termen van bijvoorbeeld (relationele) databasestructuren. We kunnen volstaan met eenvoudige entiteiten-relatiediagrammen.

De content structuur levert de benodigdheden om een goed content management te realiseren. Belangrijk voor een zo eenduidig mogelijk beheer, zijn procedures voor het publiceren van de content. Het gebruik van bovengenoemde templates en e-forms is hierbij onontbeerlijk.

Het ontwerpen en beheren van de content structuur zal gebaat zijn bij een zo groot mogelijke mate van standaardisatie. Uiterst populair is XML als structureringsgereedschap. Maar voor de ongestructureerde data

waarmee men bij content geconfronteerd wordt, zijn op een minder gestructureerd niveau *Mind Mapping* en *Information Mapping* een goede basis om in de documentenstructuur vast te leggen. Dit kan op eenvoudige wijze, via templates die aan de gebruikers ter beschikking worden gesteld.

Het wiel uitvinden

Content access (CA) moet antwoord geven op de vraag hoe medewerkers content kunnen vinden. Organisaties zijn ervan overtuigd dat ze operationele kosten kunnen reduceren, als de medewerkers meer effectief kunnen zoeken naar en toegang hebben tot aanwezige content. Efficiënte toegang en retrievalmogelijkheden voorkomen ook het uitvinden van het wiel.
De ECM-software dient minimaal een aantal functies te bezitten. Het moet mogelijk zijn om op meerdere manieren gericht te zoeken naar content over een specifiek onderwerp, te grasduinen en de mogelijkheid te hebben te kunnen worden geattendeerd op nieuw verschenen content over opgegeven onderwerpen.
Om deze functies te kunnen uitvoeren zijn hulpmiddelen nodig die dienst doen bij het zoeken naar de gewenste content. Deze hulpmiddelen zijn trefwoordenlijsten, classificatieschema's, thesauri en taxonomieën.

Buzz-word

De noodzaak tot een steeds betere en snellere toegang tot content, leidt tot een groeiende aandacht voor methoden voor de ontsluiting van content. Het buzz-word is momenteel taxonomieën.
Taxonomieën vinden hun oorsprong in de behoefte van de mens om zaken te categoriseren, niet in het minst om overzicht te kunnen houden. Dit categoriseren gebeurde tot voor kort met behulp van classificatieschema's. Deze waren echter veel te star van structuur om toe te passen in organisaties waar de content snel groeit of verandert. Vanwege het groter aantal vrijheidsgraden bij taxonomieën, kan men flexibeler ontsluiten.
De beste manier om taxonomieën te ontwerpen is volgens de facetbenadering. In de bovenstaande omschrijving van een taxonomie is al een facetindeling opgenomen, namelijk personen, organisaties, gebeurtenissen en dingen. Deze vier begrippen

sluiten elkaar volledig uit! Dit is ook tevens het belangrijkste criterium voor het kiezen van een indeling. Bij een verdere indeling per facet moet geprobeerd worden zoveel mogelijk dit uitsluitingsprincipe te blijven hanteren. Dat is de belangrijkste garantie voor een goede taxonomie.
Een goede toets voor het testen van de juistheid van een taxonomie is dat het toevoegen van een facet zonder problemen mogelijk moet zijn. De toets is toepasbaar voor zowel de product- en materialenclassificaties, als ook voor de linguïstische indelingen.

Facet-classificatie

Foskett schreef in 2000: "Thirty years of teaching did not reveal any subject which did not lend itself to facet analysis. From its success, we must assume that the future of intellectual content retrieval lies in the first place in adequate facet analysis. Facet analysis does appear to be the most important tool to our disposal for the analysis of subjects and will be increasingly important in the future."
De originele uitgangspunten van de facet-classificatie hadden als nadeel dat het onmogelijk was om relaties te definiëren tussen bovenliggende en onderliggende begrippen, binnen hetzelfde facet (denk aan de broader en narrower terms van de thesaurus). Ook konden laterale relaties niet vastgelegd worden, terwijl deze voor het genereren van ideeën toch van wezenlijk belang zijn (lees daarvoor De Bono). Voor het oplossen van dit type problemen, wordt gebruik gemaakt van de kennis en ervaringen die met de ontologie-benadering opgedaan wordt, zie daarvoor het kader Ontologieën.
Er kleven natuurlijk wel enkele nadelen aan het gebruik van taxonomieën, zoals het arbeidsintensieve ontwerpen, bouwen en onderhouden en verder is geen enkele taxonomie perfect.

Klassieke onderdelen

De klassieke onderdelen van classificatiesystemen blijven onverkort noodzakelijk voor taxonomieën. Deze onderdelen zijn de *schema's*, de *notaties*, de *index(en)* en de *organisatie*.
De schema's zijn de direct zichtbare logische structuur van de taxonomie. Deze logische structuur kan gepresenteerd worden als een boom met vertakkingen of als een aantal kolommen, waarin iedere kolom een facet vertegenwoordigt.
De notatie in de vorm van indrukwekkende cijfer-lettercombinaties zien we tegenwoordig niet meer terug. Zij worden ook algemeen als te gebruiksonvriendelijk beschouwd. Slechts af en toe zien we nog een letternotatie met bijvoorbeeld een range van A tot en met F, voor een zestal rubrieken of iets dergelijks. Maar ook in dit soort gevallen wordt meestal de voorkeur gegeven aan betekenisvolle begrippen, zoals 'kopjes' voor de betreffende rubrieken. Notaties worden nog wel gebruikt om fysieke verzamelingen te organiseren, maar dan slechts als

Ontologieën

Ontologie is de studie van categorieën binnen een domein en vormt een logische basis voor een wetenschappelijke benadering van de weergave van kennis. Met onder andere Venn-diagrammen en boomstructuren worden relaties aangegeven. De ontologie kan bestaan uit samenhangende verzamelingen van:

- Entiteiten: zoals beweringen, normen, regels, enzovoort;
- Relaties: causaliteit (oorzaak-gevolg), erkenning, gelden als;
- Handelingen: regels toepassen, interpreteren, if ... then ...;
- Feiten: institutionele feiten, naakte feiten, erkende feiten.

Een ontologie in de betekenis van ontsluitingsmechanisme, is in het ideale geval een combinatie van een thesaurus en een taxonomie. Ze bevat zowel de relationele mogelijkheden van thesauri als de hiërarchische mogelijkheden van een taxonomie (en meer!). Thesauri en taxonomieën zijn dan ook uitstekende beginpunten voor het bouwen van ontologieën.
In het semantische web, dat in ontwikkeling is, zullen ontologieën een belangrijke positie innemen.
Voor meer informatie:
http://www.pscw.uva.nl/sociosite/websoc/semantisch.html en
http://www.searchtools.com/info/classifiers.html

plaatsingssystemen. Het meest concrete voorbeeld is het plaatsen van boeken in een bibliotheek, waarbij elk boek een signatuur krijgt, die meestal op de rug van het boek terug te vinden is. Deze signatuur geeft meestal een aanduiding van het hoofdonderwerp van het boek. Gelijksoortig zijn de aanduidingen op schappen in magazijnen. De indexen hebben de rol van ingangselementen voor het zoekproces. Afhankelijk van de manier van ontsluiten worden deze indexen volledig automatisch gegenereerd of (deels) via tussenkomst van indexeerders. Met name de inzet van deze laatste is van groot belang, als er nieuwe begrippen (concepten) in een vakgebied ontstaan, als er nieuwe synoniemen opduiken en als de inhoud van begrippen verandert of begrippen overbodig worden. Allemaal redenen om niet volledig te vertrouwen op *computer-based indexing*.
En daarmee komen we op het vierde onderdeel: de organisatie. De praktijk wijst uit dat de continuïteit van de taxonomie en daarmee de toegang tot de content, gevaar loopt als de organisatie in de vorm van onderhoud en beheer niet geregeld is. Het aanstellen van een taxonomist is wezenlijk!

Een taxonomie ontwerpen

Bij het ontwerpen en bouwen van een taxonomie zijn dezelfde fasen van belang als bij het ontwerpen en bouwen van elk ander informatiesysteem: de strategie wordt bepaald, het ontwerp wordt gemaakt, de bouw vindt plaats en de gebouwde taxonomie wordt getest en geïmplementeerd.
Het ontwerpproces bestaat uit de volgende activiteiten:

1 Verzamel kandidaat-termen en hun varianten;
2 Bepaal criteria voor de keuze van de voorkeurstermen;
3 Selecteer voorkeurstermen;
4 Ontwikkel de facet-hiërarchie(ën);
5 Schrijf het ontwerp, inclusief functionele specificaties;
6 Voer een pakketselectie uit (of bouw de applicatie);
7 Implementeer de taxonomie.

Organisaties reduceren operationele kosten als medewerkers effectief kunnen zoeken naar content

Voor het verzamelen van kandidaattermen en de voorkomende varianten, is de geijkte aanpak het opsporen van reeds bestaande (vak-) woordenlijsten, thesauri, taxonomieën, classificatieschema's en het raadplegen van gebruikers en experts. Vooral dit laatste is een kritieke succesfactor. Voor de stappen 3 en 4 geldt dat ze in een aantal iteratieve cycli worden uitgevoerd.
De volgende vragen komen altijd in herhaald overleg met gebruikers en experts meerdere keren aan de orde. Welk synoniem verdient de voorkeur? Hoe gaan we de voorkeurstermen clusteren? Moet een ruimer of juist een enger begrip gekozen worden zoals voorkeursterm, of toch beide? Welke (semantische) relaties bestaan er tussen termen en moeten deze vastgelegd worden?
Bij het vastleggen van de termen moet ook aan een aantal taalkundige aspecten aandacht besteed worden om een consistente structuur te krijgen. Enkele aspecten zijn:

1 De grammaticale vorm (gebruiken we de werkwoordsvorm of het voltooid deelwoord of corresponderend zelfstandig naamwoord?);
2 De spelling;
3 Keuze maken voor óf enkelvoud- óf meervoudsvorm;
4 Gebruik van afkortingen en acroniemen.

Als er consensus is over de facet-hiërarchie kunnen de stappen 5, 6 en 7 uitgevoerd worden, zie voor verdere informatie over tools het kader Taxonomy Tools.

Conclusie

Ook voor een taxonomie is een business case te maken. Deze verschilt in principe niet van elke andere business case, met dien verstande dat er natuurlijk altijd sprake is van een deelsysteem binnen een volledige ECM-ontwikkeling. De te beantwoorden vragen zijn ook hier wat is de ROI, wat zijn kwantificeerbare voordelen, wat zijn de risico's, wat zijn de afhankelijkheden, wat zijn de andere opties?

Om de bestedingen te rechtvaardigen zijn twee aspecten van belang, de taxonomie creëert toegevoegde waarde en genereert besparingen. De toegevoegde waarde wordt bepaald door aan te geven hoe de kwaliteit van het product (de opbrengst van de zoekopdrachten) en hoe de productiviteit van de storage- and retrieval-functie verbeteren. De besparingen bestaan uit de reductie van tijd en kosten van het zoeken en terugvinden van de content en de besparingen in personeel voor het indexeren. Enkele suggesties kunnen bij het bovenstaande gegeven worden. Bijvoorbeeld, de rol van taxonomist en indexeerder kan men uitbreiden. Maak zoekacties op meerdere databases mogelijk. Maak gebruik van de verworvenheden uit de taalanalyse-technologie (automatisch genereren van samenvattingen, automatisch categoriseren van teksten). Zorg voor een snelle verwerking van grote volumes content.

Taxonomieën vinden hun oorsprong in de behoefte van de mens om zaken te categoriseren

We kunnen stellen dat ondernemingsbrede taxonomieën een globaal karakter moeten hebben, dat taxonomieën gebruikersgericht moeten zijn en gebruikersterminologieën bevatten. Verdere geleerde lessen; taxonomieën werken het best op goed gedefinieerde domeinen en er komt langzaam een groei naar het gebruik van ontologieën.

Het toekennen en beheren van metadata vereist strenge controle en discipline. Men moet beseffen dat gebruikers op verschillende manieren zoeken en er dus meerdere manieren van zoekmogelijkheden ingebouwd moeten worden en dat er meerdere zoekfuncties noodzakelijk zijn.

Leo Meerman

Leo Meerman is directeur van CELT Consultancy, een onafhankelijk adviesbureau op de terreinen kennismanagement, enterprise content- en documentmanagement (lmeerman@celt.nl).