Peter den Haan

Classificatie, taxonomieën

De opzet van een taxonomieproject

Vanaf de Griekse oudheid tot op de dag van vandaag worden taxonomieën gebruikt voor het organiseren van allerlei soorten informatie. Maar wat is het eigenlijk? Peter den Haan legt uit hoe een taxonomieproject tot een succes gemaakt kan worden.

Classificatie is iets dat we allemaal toepassen, iedere dag weer, zonder dat we het ons realiseren. Classificeren is het groeperen van soortgelijke objecten; het helpt ons om grip te krijgen op de wereld om ons heen. Wanneer er te veel informatie op ons afkomt, organiseren onze hersenen de informatie zodanig dat we snel beslissingen kunnen nemen.

Als het voor personen al noodzakelijk is om informatie te ordenen, dan geldt dat zeker ook voor organisaties waar de hoeveelheid informatie vaak exponentieel groeit. Helaas hebben veel bedrijven nog een achterstand op dit gebied. Volgens marktanalyse- en adviesbureau IDC besteedt de gemiddelde werknemer 15 procent van de werktijd aan het zoeken naar informatie. En in de helft van de gevallen is dat zelfs zonder resultaat. Dat betekent dat 7,5 procent van de werktijd verloren gaat met het zoeken naar iets dat niet wordt gevonden! Bovendien komen daar nog kosten bij: bijvoorbeeld het opnieuw produceren van dat wat niet gevonden werd, beslissingen die worden genomen op basis van verschillende versies van informatie en gemiste kansen doordat processen bij gebrek aan informatie te lang duren om nog adequaat te kunnen handelen.
Deze kosten kunnen teruggebracht worden. Een groot deel van de oplossing ligt in taxonomieën: het organiseren van de bedrijfsinformatie in een hiërarchisch systeem. Taxonomieën en classificaties zorgen voor de structuur en organisatie waarmee bedrijfsinformatie beter inzichtelijk, beheersbaar, bruikbaar en dus ook waardevoller wordt. Met andere woorden: het maakt het zoeken, maar vooral ook het vinden, van informatie eenvoudiger.

Wat is het eigenlijk?

Een taxonomie is een hiërarchische structuur in categorieën, van algemeen naar specifiek, om een bepaald domein onder te verdelen. Op zich niets nieuws, vanaf de Griekse oudheid tot op de dag van vandaag worden taxonomieën gebruikt voor het organiseren van allerlei soorten informatie.
Waarom? Wat heeft u zelf liever: een resultaatlijst met tweeduizend documenten, of een aantal categorieën die u helpen bij het verder verkennen van de gevonden informatie en het verder zoeken?

Een uitsplitsing van de zoekresultaten in een aantal categorieën maakt duidelijk dat de beschikbare informatie verschillende onderwerpen behandelt en dat het gezochte onderwerp vanuit meerdere perspectieven bekeken kan worden. Een taxonomie biedt een aanvullende ingang tot de beschikbare informatie en geeft een eerste overzicht van de aanwezige informatie. Zo kan een gebruiker door middel van navigatie een eerste selectie maken en daarbinnen een vrije zoekvraag uitvoeren.

Een taxonomie bestaat uit twee belangrijke onderdelen: de hiërarchie en de definitie van de classificatieregels ('business rules') die voor een categorie specificeren welke documenten daartoe behoren. Simpele classificatieregels kunnen worden gedefinieerd op basis van synoniemen, afkortingen en quasi-synoniemen. Complexere regels kunnen dit combineren met verdere zoektermen, attributen en andere, meer abstracte concepten (gedefinieerd in aanvullende thesauri, en dergelijke).

Classificatieregels kunnen vaak in andere applicaties worden hergebruikt. Ze kunnen worden ingezet voor het automatisch classificeren van documenten (trefwoorden worden toegevoegd aan een nieuw document zonder dat het document in een zoekomgeving wordt ingebracht). Classificatieregels kunnen tevens als abstracte concepten (thesaurusdefinities) beschikbaar gemaakt worden voor

Foto: Egon Viebre

eindgebruikers in het zoekproces, zodat er een automatische expansie van de zoektermen gebruikt kan worden. Daarnaast kunnen classificatieregels gebruikt worden binnen een 'abonnement'-model zodat eindgebruikers automatisch geïnformeerd worden over nieuwe informatie die over een onderwerp (een categorie) beschikbaar is. Dergelijke attenderingsfuncties kunnen bijvoorbeeld ook gebruikt worden om het e-mailverkeer tussen medewerkers en klanten te controleren op correctheid (traceren van onofficiële toezeggingen, verhinderen dat interne bedrijfsinformatie naar buiten gestuurd wordt, en dergelijke).

Het gebruik

Een veel voorkomende toepassing van taxonomieën is de integratie met het navigatie- en zoeksysteem op intranet en website. Eindgebruikers kunnen een zoekopdracht invoeren, maar ook door de verschillende categorieën heenlopen. Zo kan het navigeren en zoeken in categorieën resultaten opleveren waar men nog niet aan gedacht had en wordt het zoeken effectiever wanneer het beperkt is tot één categorie. Maar het is ook mogelijk, om in twee of meer taxonomieën tegelijk te navigeren en specifiek te zoeken op combinaties van categorieën uit meerdere taxonomieën. De taxonomieën worden dan naast elkaar getoond en de gebruiker kan zonder restricties navigeren door de categorieën en niveaus van beide taxonomieën, waarbij het navigatiescherm direct aangeeft hoeveel documenten per subcategorie beschikbaar zijn.[1]
Een voorbeeld: het zoeken van een leuke tweedehands auto in een landelijke database gaat efficiënter als je vrij kunt navigeren in een geografische hiërarchie (waar wordt de auto aangeboden?), een hiërarchie van autoleveranciers en automerken, een vrije zoekopdracht en verdere selectiemogelijkheden uit andere attributen (prijsklasse of kleur). Als na iedere selectiestap direct het aantal subhits per subcategorie wordt getoond, zal het meteen duidelijk worden dat bepaalde combinaties veel, weinig of zelfs geen aanbod hebben. De gebruiker kan vervolgens overwegen om de selectiecriteria te verruimen of te verscherpen.
Afgebeeld is op p. 20 een combinatieselectie waarin de gebruiker heeft geselecteerd op provincie 'USA – New York', automerk 'Europese auto', een specifieke prijsklasse (10-15.000) en een vrije-tekst vraag heeft gebruikt om 'leren bekleding' te kunnen vinden. Het scherm toont vervolgens wat de verdeling van de hits is over de beschikbare subcategorieën.

Hoe te beginnen?

Het begin van een taxonomieproject kan moeizaam zijn, maar er zijn drie belangrijke stappen voor een succesvol project:
1. vaststellen van het taxonomieteam;
2. analyseren van de wenselijke en bestaande situatie;
3. invoeren van de taxonomie.

Vaststellen van het taxonomieteam

Het belangrijkste is om de mensen in te schakelen die veel kennis hebben van de content van de organisatie en/of

expert zijn in één of meer onderwerpen. Ook is het belangrijk om de IT-afdeling vroegtijdig over het project te informeren omdat technologie een belangrijke rol speelt. Verder adviseer ik om de uiteindelijke gebruikers van de taxonomie mee te laten werken, zodat de eisen van de verschillende afdelingen binnen de organisatie aan de orde komen.

Analyseren van de wenselijke en bestaande situatie

Voordat begonnen wordt aan het daadwerkelijk bouwen van een taxonomie, is het goed om een inventarisatie te maken. Wie maken straks gebruik van de taxonomie? En hoe gebruiken zij hem? Welke informatie moet geclassificeerd worden en welke juist niet? Hebben de documenten al metadata en wat is de kwaliteit daarvan?

Het is goed om zich te realiseren dat taxonomieprojecten zelden vanuit het niets beginnen. Er is vaak al ervaring opgedaan met het organiseren van bedrijfsinformatie (vanuit bibliotheek, organisatorische structuur en productafdelingen). Deze ervaring kan gebruikt worden als basis voor het ontwerpen van de taxonomieën en het definiëren van de relevante classificatieregels.

Invoeren van een taxonomie

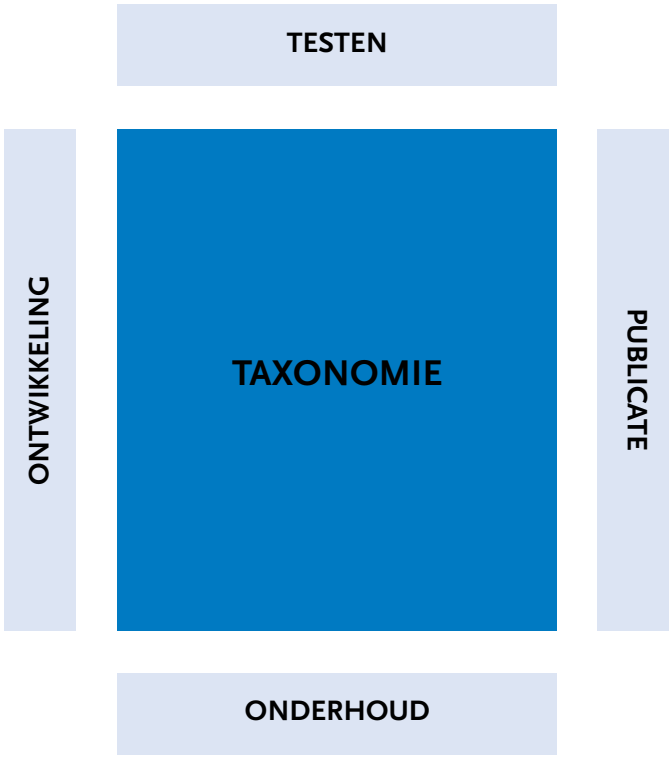
U heeft een team gevormd, de gewenste en bestaande situatie geanalyseerd en een plan van aanpak opgesteld. Nu is het tijd om daadwerkelijk de taxonomie te bouwen en de regels vast te stellen die de categorieën definiëren. Dit proces kan verdeeld worden in vier stappen:
1. taxonomieontwikkeling;
2. taxonomie testen;
3. taxonomiepublicatie;
4. taxonomieonderhoud.

Taxonomieontwikkeling

De twee belangrijkste taken bij het ontwikkelen van een taxonomie zijn:
- het bepalen van de categorieën en hun hiërarchie;
- het definiëren van de classificatieregels voor iedere categorie die bepalen welke documenten hieraan worden toegewezen.

De categorieën en hiërarchie bepalen hoe men uiteindelijk door de informatie kan navigeren en er in kan zoeken. Dit is de structuur die de eindgebruiker uiteindelijk ziet en waaronder de documenten worden gepresenteerd; het moet mogelijk zijn documenten in meerdere categorieën onder te brengen.
De naam van de categorie wordt afgeleid van de voorkeursbenaming die de organisatie hanteert voor het onderwerp. Het op deze manier zorgvuldig benoemen van iedere categorie bevordert het gebruik van de officiële terminologie van de organisatie. En dat betekent dat er minder kans is op interne en externe miscommunicatie.

Hoe u de taxonomiestructuur met de categorieën definieert, hangt onder andere af van wat u beschikbaar heeft aan materiaal en menskracht.
Een *bottom-up* benadering kan heel goed door bijvoorbeeld alle beschikbare informatie door een softwareprogramma



zoals *Verity Thematic Mapping* te laten analyseren op de hoofdonderwerpen en thema's. Thematic mapping is een methode waarbij de software zelf op basis van de aanwezige documenten de belangrijkste onderwerpen destilleert en de thema's hiërarchisch rangschikt en logische namen toekent aan de verschillende niveaus in de hiërarchie. Dit is vooral makkelijk wanneer vanaf nul moet worden begonnen. Deze informatie kan dan als basis dienen voor verdere, handmatige verfijning.
Andersom kunt u ook een *top-down* benadering toepassen. Uw bedrijf heeft al bestaande, vaak globale taxonomiedefinities die als een eerste basis kunnen worden gebruikt en door een specialist kunnen worden verfijnd. Het is daarnaast vaak mogelijk taxonomieën extern te verkrijgen, of gebruik te maken van ervaring bij andere, gelijksoortige organisaties. Een dergelijke basistaxonomie kan verder worden gecombineerd met al bestaande interne bedrijfstaxonomieën en kan worden aangepast aan de verdere wensen van de organisatie.
Het is van belang de eigen specialisten in de organisatie hierbij te betrekken. De expertise over hoe informatie gebruikt wordt binnen de organisatie, is belangrijk bij het verfijnen van de taxonomiestructuur en het definiëren van de classificatieregels.

In de praktijk zullen de verschillende benaderingen worden gecombineerd. Zo komt u tot een eerste opzet, die kan worden uitgewerkt na evaluatie en goedkeuring door de diverse betrokken partijen.

Het vaststellen van de classificatieregels

Iedere categorie in de taxonomie moet worden gekoppeld aan een classificatieregel.[2] Deze specificeert welke documenten voor deze categorie getoond moeten worden. De

twee belangrijkste methoden om de classificatieregels te bepalen zijn het handmatig definiëren en het gebruik van trainingsdocumenten (ook automatische classificatie genoemd).
De handmatige methode houdt in dat specialisten op een bepaald vakgebied vastleggen welke documenten in welke categorie horen op basis van een aantal criteria.[3] De uiteindelijke classificatieregel kan gebruik maken van bepaalde woorden en uitdrukkingen, eventueel aangepaste relevantie indicatoren per element, de plaats in het document, eventuele attributen, enzovoort. Om op een effectieve manier handmatig nauwkeurige classificatieregels te bepalen, moet de specialist een goede beheer- en (interactieve) testomgeving tot zijn beschikking hebben.
Naast directe input van de specialist kunnen ook andere bestaande bronnen nuttige suggesties voor classificatieterminologie bieden; denk aan search logs, begrippenlijsten, site maps, synoniemenlijsten, productlijsten en overzichten met afkortingen. Het zorgvuldig samenstellen van de classificatieregels moet ervoor zorgen, dat alle relevante informatie gevonden kan worden, ook wanneer documenten niet de officiële terminologie gebruiken.
Bij training worden per categorie representatieve documenten gezocht en ingevoerd in de trainingsoftware. De technologie leert zo waar deze documenten over gaan en creëert automatisch de regel voor de categorie. Ten behoeve van de nauwkeurigheid kan het nuttig zijn om een expert de automatisch gegenereerde classificatieregels vervolgens handmatig te laten bewerken. Daarnaast kan ook de automatische Thematic-mapping methode suggesties voor classificatieregels aanleveren.
Een combinatie van de bovengenoemde methodes levert vaak nog betere resultaten op.

Nadat de classificatieregels voor de categorieën zijn bepaald, kan automatisch een navigatie-index worden gecreëerd op basis van de documenten. Dit gebeurt veelal in combinatie met een standaard index ten behoeve van een 'gewone' zoekomgeving, zodat eindgebruikers inderdaad beide methoden kunnen gebruiken.

Taxonomie testen

Tussen het echte bouwen van de taxonomie en het publiceren daarvan, ligt een van de belangrijkste stappen, namelijk het testen van de taxonomie. Dit betekent dat categorieën en subcategorieën doorlopen moeten worden om te bekijken of het allemaal klopt. Waar categorieën 'veel' documenten bevatten, kan er eventueel verder onderverdeeld worden. Waar ze 'weinig' documenten bevatten, kan er wellicht samengevoegd worden.
Daarnaast is het belangrijk de hiërarchie van de taxonomie te analyseren om er zeker van te zijn dat gebruikers niet al te diep hoeven te graven om iets te vinden. Documenteer hierbij uw bevindingen, de aanpassingen die u gedaan heeft en waarom u die gedaan heeft.

Taxonomiepublicatie

Daarna is het systeem toe aan de eindgebruikers. Het is een goed idee om een basistraining via intranet of een website beschikbaar te stellen. Ook al is het de bedoeling dat een taxonomie voor zichzelf spreekt, toch is het vaak zo nieuw dat het nuttig is om een aantal tips te geven voor de verschillende manieren van zoeken die nu kunnen worden toegepast. Is de taxonomie bedoeld voor een intranetapplicatie? Organiseer dan webinars voor de verschillende gebruiksgroepen. Naast een goede introductie is natuurlijk ook een goed feedback-systeem essentieel.

Taxonomieonderhoud

Het werk is nog niet voorbij. Nu de taxonomie in gebruik is genomen, wordt onderhoud essentieel. Waarom? Omdat werknemers nieuwe documenten produceren over nieuwe technologieën, producten, klanten, markten enzovoort. En ook de manier waarop werknemers de bestaande content gebruiken zal blijven veranderen. En terminologie verandert. Een taxonomie moet met de organisatie en de informatie kunnen meegroeien en veranderen.
De hoeveelheid tijd die nodig is voor het onderhoud, hangt onder andere af van de omloopsnelheid van informatie binnen de organisatie en de grootte en complexiteit van de taxonomie.
Maar het is belangrijk dat er procedures zijn voor het regelmatig bijhouden en herzien van de taxonomie. Denk bijvoorbeeld tijdig aan veranderingen, zoals nieuwe productlijnen, of een nieuw bedrijfsonderdeel, waardoor nieuwe categorieën nodig zijn.

Een goede evaluatiemethode is het naast elkaar leggen van de taxonomie en de rapportages van de zoekmachine en het taxonomiegebruik. Dit kan bijvoorbeeld leiden tot het verfijnen van categorieën of business rules. Ook de feedback van medewerkers is onmisbaar, net als het regelmatig laten testen van het systeem door ervaren gebruikers. Het geeft de gelegenheid om het systeem te verfijnen, zodat de zoekresultaten steeds relevanter worden.

Dit klinkt allemaal zeer complex en uitgebreid. Maar het is onze ervaring dat iedereen uiteindelijk het belang inziet van het onderhouden van een taxonomie en het als bijna vanzelfsprekend meeneemt in de dagelijkse werkzaamheden. Het onderhoud van specifieke subcategorieën kan ook worden neergelegd bij de verschillende, in de organisatie aanwezige specialisten. Dit bevordert meteen de samenwerking tussen de inrichters van de taxonomie en de eindgebruikers.

Noten

1. Deze functionaliteit lijkt sterk op wat binnen de DBMS-wereld ook omschreven wordt als 'datawarehousing' technieken. Door middel van flexibele navigatie- en presentatiefuncties kan een eindgebruiker makkelijker informatie vinden, en indien er veel of weinig informatie beschikbaar is voor de actuele selectie, verder in- of uitzoomen.
2. Categorieën op een tussenniveau in de hiërarchie hebben niet altijd een eigen classificatieregel, maar presenteren bijvoorbeeld automatisch de documenten beschikbaar in alle subcategorieën.
3. Voor een brede informatieomgeving kan het nodig zijn dat meer specialisten moeten samenwerken om de gehele taxonomiestructuur te kunnen ontwikkelen.

Peter den Haan is Technical Director bij Verity Europe.