Zonder uitgever is de zoekmachine nergens

07-11-2012 08:26

In mijn artikel “Metadata zijn geld waard” geef ik een voorbeeld hoe je omzet kunt genereren met metadata. Toch zijn er sceptici over de noodzaak van metadateren. Zij menen dat zoekmachines slimmer worden. Die zouden dan door goed te indexeren content snel kunnen vinden. Deze optimistische gedachte is naar mijn mening gedeeltelijk waar. Ja, zoekmachines worden slimmer en bijvoorbeeld Google houdt al rekening met taal, synoniemen, taalfouten en vervoegingen. Door een algoritme weet een zoekmachine dat als de zoekwoorden dichter bij elkaar staan in een zin, in een paragraaf of op een pagina, dat de content waarschijnlijk gaat over het onderwerp of trefwoord waarop gezocht is. De zoekmachine weet dan nog niet welk het meest relevant is en geeft de gevonden resultaten als een reeks van records terug. Het is dan aan jou om uit te kiezen welke relevant is.

Deze werkwijze - die de meeste mensen nu nog als prettig ervaren - werkt prima op een ongecontroleerde verzameling content zoals het internet. Het wordt complexer als je informatie zoekt waarvan het trefwoord - of synoniemen ervan - niet voorkomt in de content. Dit komt vaker voor dan gedacht. Een voorbeeld: In het Hollende kleurling-arrest Hof Amsterdam 03-06-1977, NJ 1978, 601 komt de term “Hollende kleurling” niet voor. Bijzonder detail, zoek je in Rechtspraak.nl op “Hollende kleurling-arrest” dan wordt wel een uitspraak gevonden waarin verwezen wordt naar het bedoelde arrest. Het arrest zelf wordt niet gevonden, omdat deze niet is verrijkt met de juiste termen. Een ander voorbeeld dat ik tegenkwam bij het schrijven van dit artikel, is het zoeken van een bestaande taxonomie voor de sector bouw. Alle zoekmachineresultaten gingen over het bouwen van een taxonomie …

Met andere woorden, alles wat je toevoegt aan kennis over een document of ander elektronisch object, verbetert de vindbaarheid van het document. Dit toevoegen kan dus ook het beste gebeuren door iemand die kennis heeft van de inhoud. Hij/zij is in staat aan te geven waar de content over gaat, zonder afhankelijk te zijn van de gebruikte termen in het document. Een standaardzoekmachine zou dit artikel kunnen vinden als iemand zoekt op “gevonden voorwerpen”, terwijl het daar niet over gaat. Het aanbrengen van deze toegevoegde waarde is wat uitgevers goed kunnen faciliteren.

Om het verrijken te structureren en binnen de metadata inhoudelijk te kunnen navigeren zijn er modellen beschikbaar zoals een taxonomie, thesaurus of ontologie. Door een dergelijke metadatastructuur kun je makkelijk navigeren. Zoekmachines kunnen hiervan gebruikmaken om relevante documenten te vinden zonder dat de documenten met alle mogelijke termen verrijkt zijn. Dus een groot bijkomend voordeel is dat het verrijken van de documenten eenvoudiger wordt, omdat er minder termen toegevoegd hoeven te worden. Bedenk wel eerst goed welk model het beste past bij de branche, het product en het contenttype. Probeer te voorkomen dat een reeds gekozen systeem af gaat dwingen dat er bijvoorbeeld een thesaurus gevuld moet gaan worden zonder direct aanwijsbare redenen. Een taxonomie zou kunnen volstaan.

Maar hoe begin je aan een taxonomie of thesaurus? Ik adviseer altijd om eerst uit te zoeken of er al bestaande taxonomieën of thesauri zijn binnen een branche. Mocht nu een taxonomie of thesaurus niet voorhanden zijn in een specifieke branche, dan is er misschien wel een markt voor om de “kennisboom” commercieel beschikbaar te stellen. Met de opmars van steeds meer versnipperde content zou dit wel eens een interessant uitgeefproduct kunnen worden.

Op mijn website en in mijn e-book heb ik enkele voorbeelden opgenomen hoe je zou kunnen starten met een taxonomie of thesaurus.

Edwin de Kuiper